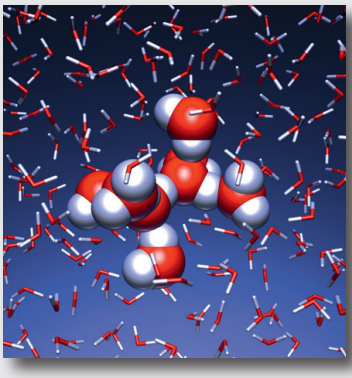
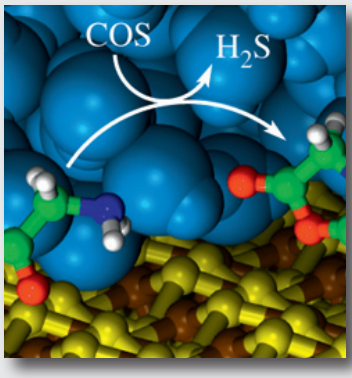
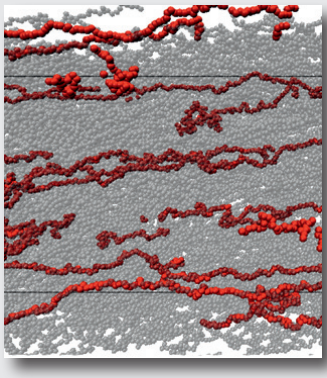
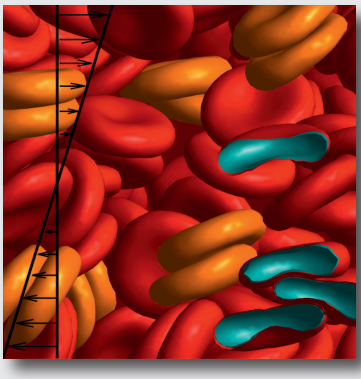
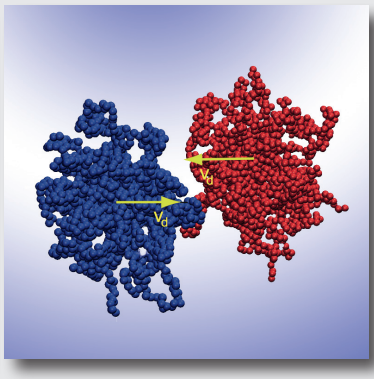
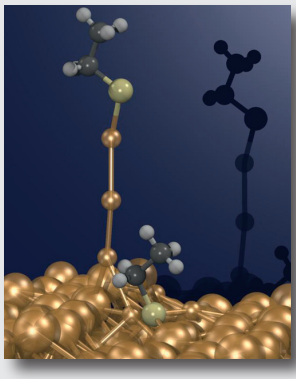












Hierarchical Methods for Dynamics in Complex Molecular Systems Lecture Notes

edited by Johannes Grotendorst, Godehard Sutmann, Gerhard Gompper, Dominik Marx

JÜLICH
FORSCHUNGSZENTRUM

Mitglied der Helmholtz-Gemeinschaft

Schriften des Forschungszentrums Jülich
IAS Series Volume 10

Forschungszentrum Jülich GmbH
Institute for Advanced Simulation (IAS)
Jülich Supercomputing Centre (JSC)

Hierarchical Methods for Dynamics in Complex Molecular Systems

edited by Johannes Grotendorst, Godehard Sutmann,
Gerhard Gompper, Dominik Marx

IAS Winter School, 5 – 9 March 2012
Forschungszentrum Jülich GmbH
Lecture Notes

Schriften des Forschungszentrums Jülich
IAS Series Volume 10

ISSN 1868-8489 ISBN 978-3-89336-768-9

Bibliographic information published by the Deutsche Nationalbibliothek.
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.d-nb.de.

Publisher and Distributor:	Forschungszentrum Jülich GmbH Zentralbibliothek 52425 Jülich Phone +49 (0) 24 61 61-53 68 · Fax +49 (0) 24 61 61-61 03 e-mail: zb-publikation@fz-juelich.de Internet: http://www.fz-juelich.de/zb
Cover Design:	Jülich Supercomputing Centre, Forschungszentrum Jülich GmbH
Printer:	Grafische Medien, Forschungszentrum Jülich GmbH
Copyright:	Forschungszentrum Jülich 2012

Schriften des Forschungszentrums Jülich
IAS Series Volume 10

ISSN 1868-8489
ISBN 978-3-89336-768-9

The complete volume is freely available on the Internet on the Jülicher Open Access Server (JUWEL) at http://www.fz-juelich.de/zb/juwel

Persistent Identifier: urn:nbn:de:0001-2012020208
Resolving URL: http://www.persistent-identifier.de/?link=610

Preface

Generating and analyzing the dynamics of molecular systems is a true challenge to molecular simulation. It includes processes that happen on the femtosecond scale, such as photoinduced nonadiabatic (bio)chemical reactions, and touches the range of seconds, being e.g. relevant in biophysics to cellular processes or in material sciences to crack propagation. Thus, many orders of magnitude in time need to be covered either concurrently or hierarchically. In the latest edition of this series of Winter Schools in 2009 we addressed the topic of *Multiscale Simulation Methods in Molecular Sciences* with a strong focus on methods which cover diversities of length scales. The key issue of the present school is to dwell on hierarchical methods for dynamics having primarily in mind systems described in terms of many atoms or molecules. One extreme end of relevant time scales is found in the sub-femtosecond range but which influence dynamical events which are orders of magnitude slower. Examples for such phenomena might be photo-induced switching of individual molecules, which results in large-amplitude relaxation in liquids or photodriven phase transitions of liquid crystals, phenomena for which nonadiabatic quantum dynamics methods were developed. The other end of relevant time scales is found in a broad range of microseconds, seconds or beyond and which governs e.g. non-equilibrium dynamics in polymer flows or blood cells in complex geometries like microvessels. Special mesoscopic techniques are applied for these time- and length-scales to couple the atomistic nature of particles to the hydrodynamics of flows.

This Winter School has a daily stratification pattern starting with dynamics within the realm of Materials Science with a focus on slow processes which nevertheless require most detailed input at the level of electronic structure and interatomic potentials. In Biomolecular Science one challenge is the concurrent handling of an electronic structure based description of a "hot spot" within an enzyme with a computationally efficient treatment of the protein environment in terms of parameterized interactions. Accelerated sampling is a key issue whenever both slow and fast motion is relevant and applies methods in the fields of metadynamics, force probe molecular dynamics or nonequilibrium dynamics using fluctuation theorems. Finally, getting rid of atoms and molecules but still keeping a particle perspective is achieved by coarse-graining procedures. In Soft Matter and Life Science, the dynamics is often governed by the hydrodynamics of the solvent. A particular challenge is here to bridge the large length- and time-scale gap between the small solvent molecules and the embedded macromolecules or macromolecular assemblies (polymers, colloids, vesicles, cells). Therefore, several mesoscale simulation approaches have been developed recently, including Lattice Boltzmann, Dissipative Particle Dynamics and Multi-Particle Collision Dynamics, which rely on a strong simplification of the microscopic dynamics with a simultaneous implementation of conservation laws on mass, momentum and energy.

Last but not least most efficient implementation on current-day hardware is a strong requirement to overcome computational barriers and to tackle large systems in multiscale environments. Examples will be provided covering basic methods or well-established optimal numerical methods like multigrid. In addition to lectures and poster sessions

this Winter School offers an introductory course to parallel computing techniques with practical sessions.

The target group of this IAS Winter School, organized and supported by the Jülich CECAM Node, are young scientists, especially PhD students and early postdocs.

Many individuals have significantly contributed to the success of the School. First of all we are very grateful to the lecturers for preparing extended lecture notes in due time, in spite of the heavy work load they all have to carry. Without their effort such an excellent reference book on *Hierarchical Methods for Dynamics in Complex Molecular Systems* would not have been possible.

We are greatly indebted to the School's secretaries Elke Bielitza and Britta Hoßfeld who were indispensable for this School by taking care of logistical details, transports, registration and catering. Also, we would like to express our gratitude both to Monika Marx for realizing the book of poster abstracts and to Oliver Bücker for technical and administrational support. Particular thanks go to Martina Kamps, who compiled the contributions and created the present high quality book.

Jülich and Bochum
February 2012

Johannes Grotendorst
Godehard Sutmann
Gerhard Gompper
Dominik Marx

Contents

Hard and Soft Materials

Biomolecular Systems

Advanced Methods

Flow Simulation and Transport

Parallel Computing and Numerical Methods

Simulating Light-Induced Phenomena in Soft Matter

Nikos L. Doltsinis

Institut für Festkörpertheorie
Universität Münster, 48149 Münster, Germany
E-mail: nikos.doltsinis@wwu.de

The absorption of light by soft (bio-)materials initially causes photophysical or photochemical molecular processes which are highly local in space and time. These fast quantum-mechanical events often trigger much slower, macroscopically observable phenomena. Thus to simulate light-induced macroscopic functionality of nanomaterials one needs to bridge many orders of magnitude in space and time. In this lecture a suitable multiscale simulation strategy is outlined which connects the quantum to the mesoscopic level by bringing together nonadiabatic *ab initio* molecular dynamics (QM), classical (force field) molecular dynamics (MM), and coarse grained (CG) simulation techniques. This methodology is applied to model light-induced phase transitions in a liquid crystal containing the azobenzene photoswitch – a prototypical example for a wide range of light-triggered phenomena in soft matter.

1 Introduction

Although quantum-mechanical processes are typically local in time and space, they can trigger a cascade of events on different length and time scales leading to macroscopically observable phenomena. From the standpoint of computer simulation, the use of different hierarchical levels of theory is required to describe such mesoscopic or even macroscopic phenomena. However, traditionally, theoretical methods have been tailored to deal with specific, limited, time and length scales without connecting to the levels above or below.

At the most detailed level, quantum-mechanical (QM) dynamical calculations of soft matter can be performed using the *ab initio* molecular dynamics (AIMD) method[1–3]. This technique is able to resolve the system's electronic structure "on the fly" in an approximate fashion, typically within the framework of density functional theory. Due to the high computational cost involved, however, applications are limited to systems containing a few hundred atoms and to trajectories of a few tens of picoseconds.

Abandoning electronic structure while retaining atomistic resolution one arrives at the molecular mechanics (MM) approach, i.e. classical molecular dynamics using (empirical) force field potentials. While the MM method extends the simulation range to a few thousand atoms on the nanosecond time scale, it typically does not allow for chemical reactions, i.e. bond breaking and formation, or other inherently quantum-mechanical events such as photoinduced processes and charge transfer. Such issues can be addressed using hybrid QM/MM techniques, which treat the chemically active centre quantum-mechanically and the chemically inert environment classically[4–9].

To go beyond the MM time and length scales, atomistic resolution needs to be abandoned. By mapping groups of atomic centres onto an effective particle, the number of interactions in a system can be significantly reduced[10,11]. Coarse grained (CG) interaction potentials between the effective particles are usually derived from the atomistic model by fitting the parameters to reproduce thermodynamic and/or structural properties[12–20]. CG models have been applied successfully to a variety of systems including polymers[21–29],

lipid systems[30,14,31,32], peptides[33–37], and proteins[38–40]. More on recent advances in this field can be found in Ref. 41.

The challenge is now to connect the different simulations levels – QM, MM, and CG – in a suitable way. While in the bottom–up direction, mapping each representation onto the next one above is, in principle, straightforward, the situation is less obvious in the opposite direction. However, suitable back-mapping recipes have been developed to switch back from the CG to MM description whenever a higher resolution is required[22,42,24,43,35,44,36,37,45].

Multiscale simulation methods generally attempt to connect more than two hierarchical levels. In this lecture we describe our recent efforts to interlink the three simulation levels QM, MM, and CG for molecular systems[46] with particular emphasis on light-induced processes in biosystems or photoswitchable materials. This introduces an additional level of complexity as we need to go beyond the standard QM description, i.e. a ground state Born-Oppenheimer treatment, and use a nonadiabatic approach (na-QM) coupling multiple electronic states. Here we describe our na-QM/MM/CG multiscale simulation approach, which is applicable to a wide range of mesoscopic photoinduced phenomena. As an application prototypical of the vast area of light-controllable materials based on the azobenzene (AB) photoswitch[47–58], we discuss the photoinduced phase transitions in the 8AB8 liquid crystal[59].

The development of a na-QM/MM/CG multiscale method required us to couple a previously developed na-QM approach[60,61] to the MM level using a suitable QM/MM interface[62]. The new na-QM/MM simulation tool is generally applicable to a wide range of photoexcited systems; it has already been employed to study the photoisomerisation of AB in the bulk liquid phase[63,64] and the light-triggered unfolding of a polypeptide[65]. To be able to carry out multiscale simulations photoswitchable materials based on AB-containing organic chains a suitable atomistic force field first needed to be developed[46] by mapping the results from QM (AIMD) simulations. In a subsequent step, a CG model of the 8AB8 liquid crystal was derived based on the atomistic azo force field[44].

With the necessary tools in place, multiscale simulations of photoactive azo-materials combining the na-QM, MM, and CG hierarchical levels can now be performed. We can distinguish between two different ways. A sequential approach involves separate simulations at the different levels and cross-linking by passing on the information obtained from the simulation at the next higher/lower level in the form of initial conditions (see Fig. 1a). The overall timespan covered by the simulation this way is long, as it is essentially determined by the CG level. The drawback is that the effects of a larger CG environment on the QM subsystem is not taken into account. The alternative approach is a *simultaneous* na-QM/MM/CG simulation (Fig. 1b), which would allow the modelling of larger systems compared to the sequential approach at the expense of a much shorter timescale determined by the fast QM process.

In simultaneous multiscale simulations the issue of partitioning the total system into the QM, MM, and CG subsystems arises. Standard QM/MM simulations use fixed partitions, i.e. atoms are assigned to the QM or MM regions prior to the calculation and remain in their respective partitions throughout the run. Thus, no exchange of particles between the QM and MM subsystems can take place, caused either by a moving QM/MM boundary or by particle diffusion through the QM/MM boundary. This precludes the application of the QM/MM method to the study of such important physical problems as, for instance, crack

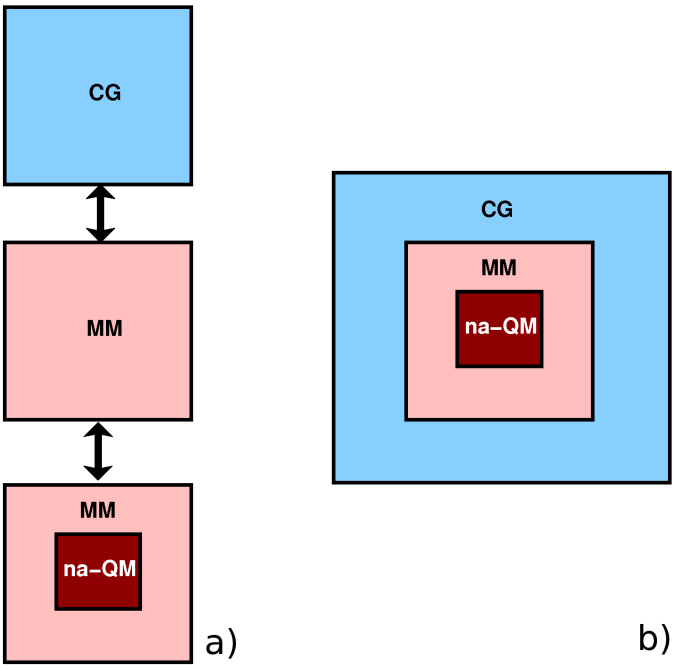


Figure 1. Schematic representations of the two different multiscale simulation approaches: sequential (a) and simultaneous (b).

propagation in materials or solvated active sites in biomolecules.

Adaptive QM/MM resolution schemes designed to overcome the fixed partitioning problem have been proposed[66–72], but they generally suffer from either an excessive computational overhead or a lack conceptual consistency (see Sec. 2.8), hence the number of applications is small.

Other extensions would involve the inclusion of two or more QM centres at a given time. In photoactive materials, this situation occurs when two chromophores in close proximity are excited within a short time span. In the case of azomaterials, for example, one may ask the question as to what extent the final state (*cis* or *trans*) of one AB chromophore depends on the photoisomerisation dynamics of its neighbours? Currently, there is no QM/MM scheme that allows for two separate QM regions.

In addition to spatial partitioning, there is also the issue of temporal switches between different representations. For instance, different AB chromophores in an azo-material absorb photons undergo photoisomerisation at different times. This scenario requires "switching off" one QM subsystem (i.e. turning it into an MM system) and simultaneously "switching on" another QM subsystem using a suitable switching function.

In analogy to the QM/MM case, certain physical problems at the mesoscopic level call for a hybrid MM/CG approach where the same partitioning problems persist. Therefore, in the area of materials science, MM/CG dual resolution methods have been developed recently, including a novel adaptive resolution molecular dynamics technique[73–76] in which the representation of a particular molecule (i.e. MM or CG) is dynamically adjusted ac-

cording to its position relative to the MM/CG boundary.

In Sec. 2 some general background theory and the different hierarchical molecular dynamics simulation methods, which represent the building blocks for our multiscale model, are briefly reviewed. Sec. 3 then presents illustrative example applications of these techniques to various aspects contributing to our ultimate target system, namely the light-induced phase transition in the 8AB8 liquid crystal.

2 Theoretical Background

2.1 Born-Oppenheimer Approximation

A complete, non-relativistic, description of a system of N atoms having the positions $\mathbf{R} = (\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_K, \ldots, \mathbf{R}_N)$ with n electrons located at $\mathbf{r} = (\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_K, \ldots, \mathbf{r}_n)$ is provided by the time-dependent Schrödinger equation

$$\mathcal{H}\Xi(\mathbf{r}, \mathbf{R}; t) = i\hbar \frac{\partial}{\partial t} \Xi(\mathbf{r}, \mathbf{R}; t) \quad , \tag{1}$$

with the total Hamiltonian

$$\mathcal{H}(\mathbf{r}, \mathbf{R}) = \mathcal{T}(\mathbf{R}) + \mathcal{T}(\mathbf{r}) + \mathcal{V}(\mathbf{R}) + \mathcal{V}(\mathbf{r}, \mathbf{R}) + \mathcal{V}(\mathbf{r}) \quad , \tag{2}$$

being the sum of kinetic energy of the atomic nuclei,

$$\mathcal{T}(\mathbf{R}) = -\frac{\hbar^2}{2} \sum_{K=1}^{N} \frac{\boldsymbol{\nabla}_K^2}{M_K} \quad , \tag{3}$$

kinetic energy of the electrons,

$$\mathcal{T}(\mathbf{r}) = -\frac{\hbar^2}{2m_e} \sum_{k=1}^{n} \boldsymbol{\nabla}_k^2 \quad , \tag{4}$$

internuclear repulsion,

$$\mathcal{V}(\mathbf{R}) = \frac{e^2}{4\pi\epsilon_0} \sum_{K=1}^{N-1} \sum_{L>K}^{N} \frac{Z_K Z_L}{|\mathbf{R}_K - \mathbf{R}_L|} \quad , \tag{5}$$

electronic – nuclear attraction,

$$\mathcal{V}(\mathbf{r}, \mathbf{R}) = -\frac{e^2}{4\pi\epsilon_0} \sum_{K=1}^{N} \sum_{k=1}^{n} \frac{Z_K}{|\mathbf{r}_k - \mathbf{R}_K|} \quad , \tag{6}$$

and interelectronic repulsion,

$$\mathcal{V}(\mathbf{r}) = \frac{e^2}{4\pi\epsilon_0} \sum_{k=1}^{n-1} \sum_{l>k}^{n} \frac{1}{|\mathbf{r}_k - \mathbf{r}_l|} \quad . \tag{7}$$

Here, M_K and Z_K denote the mass and atomic number of nucleus K; m_e and e are the electronic mass and elementary charge, and ϵ_0 is the permittivity of vacuum. The nabla operators $\boldsymbol{\nabla}_K$ and $\boldsymbol{\nabla}_k$ act on the coordinates of nucleus K and electron k, respectively.

Defining the electronic Hamiltonian (fixed-nuclei approximation of $\mathcal{H}$) as

$$\mathcal{H}_{\mathrm{el}}(\mathbf{r},\mathbf{R}) = \mathcal{T}(\mathbf{r}) + \mathcal{V}(\mathbf{R}) + \mathcal{V}(\mathbf{r},\mathbf{R}) + \mathcal{V}(\mathbf{r}) \quad , \tag{8}$$

we can rewrite the total Hamiltonian as

$$\mathcal{H}(\mathbf{r},\mathbf{R}) = \mathcal{T}(\mathbf{R}) + \mathcal{H}_{\mathrm{el}}(\mathbf{r},\mathbf{R}) \quad . \tag{9}$$

Let us suppose the solutions of the time-independent (electronic) Schrödinger equation,

$$\mathcal{H}_{\mathrm{el}}(\mathbf{r},\mathbf{R})\phi_i(\mathbf{r},\mathbf{R}) = E_i(\mathbf{R})\phi_i(\mathbf{r},\mathbf{R}) \quad , \tag{10}$$

are known. Furthermore, the spectrum of $\mathcal{H}_{\mathrm{el}}(\mathbf{r},\mathbf{R})$ is assumed to be discrete and the eigenfunctions orthonormalised:

$$\int_{-\infty}^{\infty} \phi_i^*(\mathbf{r},\mathbf{R})\phi_j(\mathbf{r},\mathbf{R})\mathrm{d}\mathbf{r} \equiv \langle\phi_i|\phi_j\rangle = \delta_{ij} \quad . \tag{11}$$

The total wavefunction Ξ can be expanded in terms of the eigenfunctions of $\mathcal{H}_{\mathrm{el}}$ since these form a complete set:

$$\Xi(\mathbf{r},\mathbf{R};t) = \sum_j \phi_j(\mathbf{r},\mathbf{R})\chi_j(\mathbf{R},t) \quad . \tag{12}$$

Insertion of this ansatz into the time-dependent Schrödinger equation (1) followed by multiplication from the left by $\phi_i^*(\mathbf{r},\mathbf{R})$ and integration over the electronic coordinates leads to a set of coupled differential equations:

$$\left[\mathcal{T}(\mathbf{R}) + E_i(\mathbf{R})\right]\chi_i + \sum_j \mathcal{C}_{ij}\chi_j = i\hbar\frac{\partial}{\partial t}\chi_i \quad , \tag{13}$$

where the coupling operator $\mathcal{C}_{ij}$ is defined as

$$\mathcal{C}_{ij} \equiv \langle\phi_i|\mathcal{T}(\mathbf{R})|\phi_j\rangle - \sum_K \frac{\hbar^2}{M_K}\langle\phi_i|\boldsymbol{\nabla}_K|\phi_j\rangle\boldsymbol{\nabla}_K \quad . \tag{14}$$

The diagonal term $\mathcal{C}_{ii}$ represents a correction to the (adiabatic) eigenvalue E_i of the electronic Schrödinger equation (10). In the case that all coupling operators $\mathcal{C}_{ij}$ are negligible, the set of differential Eqs. 13 becomes uncoupled:

$$\left[\mathcal{T}(\mathbf{R}) + E_i(\mathbf{R})\right]\chi_i = i\hbar\frac{\partial}{\partial t}\chi_i \quad . \tag{15}$$

This means that the nuclear motion proceeds without changes of the quantum state of the electron cloud and, correspondingly, the wavefunction (12) is reduced to a single term (adiabatic approximation):

$$\Xi(\mathbf{r},\mathbf{R};t) \approx \phi_i(\mathbf{r},\mathbf{R})\chi_i(\mathbf{R},t) \quad . \tag{16}$$

For a great number of physical situations the Born-Oppenheimer approximation can be safely applied. On the other hand, there are many important chemical phenomena like, for instance, charge transfer and photoisomerisation reactions, whose very existence is due to the inseparability of electronic and nuclear motion. Inclusion of nonadiabatic effects will be the subject of the following sections.

2.2 Mixed Quantum–Classical Approach

Further simplification of the problem can be achieved by describing nuclear motion by classical mechanics and only the electrons quantum mechanically. In this so-called mixed quantum–classical (sometimes referred to as semiclassical) approach[77,78], the atomic nuclei follow some trajectory $\mathbf{R}(t)$ while the electronic motion is captured by some time-dependent total wavefunction $\Phi(\mathbf{r};t)$ satisfying the time-dependent electronic Schrödinger equation,

$$\mathcal{H}_{\mathrm{el}}(\mathbf{r},\mathbf{R}(t))\Phi(\mathbf{r};t) = i\hbar\frac{\partial}{\partial t}\Phi(\mathbf{r};t) \quad . \tag{17}$$

Again, the total wavefunction is written as a linear combination of adiabatic eigenfunctions $\phi_i(\mathbf{r},\mathbf{R})$ (solutions of the time-independent Schrödinger equation (10)):

$$\Phi(\mathbf{r};t) = \sum_j a_j(t)\phi_j(\mathbf{r},\mathbf{R})e^{-\frac{i}{\hbar}\int E_j(\mathbf{R})\mathrm{d}t} \quad . \tag{18}$$

Insertion of this ansatz into the time-dependent electronic Schrödinger equation (17) followed by multiplication from the left by $\phi_i^*(\mathbf{r},\mathbf{R})$ and integration over the electronic coordinates leads to a set of coupled differential equations:

$$\dot{a}_i = -\sum_j a_j C_{ij} e^{-\frac{i}{\hbar}\int (E_j - E_i)\mathrm{d}t} \quad , \tag{19}$$

where

$$C_{ij} \equiv \langle\phi_i|\frac{\partial}{\partial t}|\phi_j\rangle \tag{20}$$

are the nonadiabatic coupling elements. Integration of Eqs. 19 yields the expansion coefficients $a_i(t)$ whose square modulus, $|a_i(t)|^2$, can be interpreted as the probability of finding the system in the adiabatic state i at time t.

We now want to develop a condition for the validity of the Born-Oppenheimer approximation based on qualitative arguments. For this purpose, we shall consider a two-state system. To illustrate the problem, Fig. 2 shows the avoided crossing between the covalent and ionic potential energy curves of NaCl[79,80]. As we can see, the adiabatic wavefunctions ϕ_1 and ϕ_2 change their character as the bond length is varied. The characteristic length, l, over which ϕ_1 and ϕ_2 change significantly clearly depends on the nuclear configuration $\mathbf{R}$; in the vicinity of the NaCl avoided crossing, for instance, the character of the wavefunctions varies rapidly, whereas at large separations it remains more or less constant.

Division of the characteristic length l by the velocity of the nuclei, $\dot{R} = |\dot{\mathbf{R}}|$, at a particular configuration $\mathbf{R}$ defines the time the system needs to travel the distance l around $\mathbf{R}$:

$$\text{passage time } \tau_p = \frac{l}{\dot{R}} \quad . \tag{21}$$

In order for the Born-Oppenheimer approximation to be valid, the electron cloud has to adjust instantly to the nuclear changes. The time scale characteristic of electronic motion can be obtained from the relation

$$\Delta E = |E_1 - E_2| = \hbar\omega \tag{22}$$

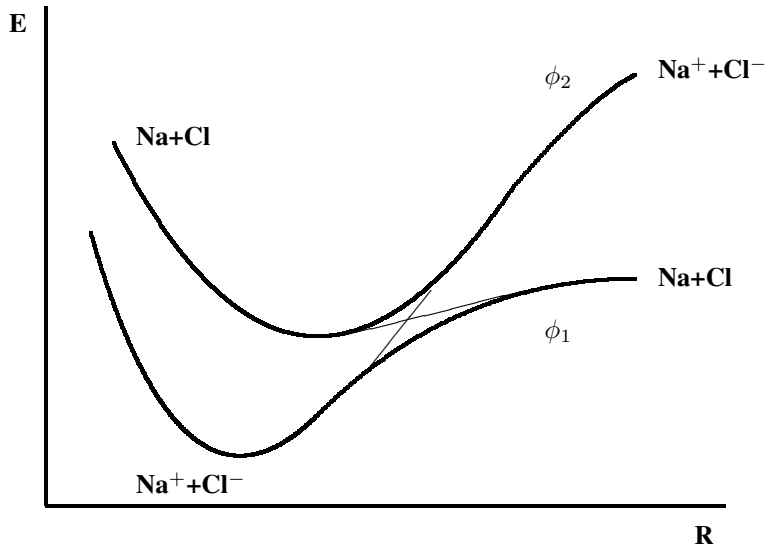


Figure 2. Avoided crossing between the covalent and ionic adiabatic potential curves of NaCl (thin lines: crossing of diabatic states).

by taking the inverse transition frequency:

$$\tau_e = \frac{1}{\omega} = \frac{\hbar}{\Delta E} \quad . \tag{23}$$

The ratio

$$\xi = \frac{\tau_p}{\tau_e} = \frac{\Delta E\, l}{\hbar \dot{R}} \tag{24}$$

is the so-called Massay parameter. For values $\xi \gg 1$, i.e. large energy gaps ΔE and small velocities $\dot{R}$, nonadiabatic effects are negligible. In this case, if the system is prepared in some pure adiabatic state i ($|a_i|^2 = 1$) at time $t = 0$, the rhs of Eq. 19 will be zero at all times and the wavefunction expansion (Eq. 18) can be replaced by a single term:

$$\Phi(\mathbf{r}; t) = \phi_i(\mathbf{r}, \mathbf{R}) e^{-\frac{i}{\hbar} \int E_i(\mathbf{R}) \mathrm{d}t} \quad . \tag{25}$$

The atomic nuclei are then propagated by solving Newton's equations

$$M_K \ddot{\mathbf{R}}_K = \mathbf{F}_K(\mathbf{R}) \quad , \tag{26}$$

where

$$\mathbf{F}_K(\mathbf{R}) = -\boldsymbol{\nabla}_K E_i(\mathbf{R}) \tag{27}$$

is the force on atom K.

2.3 Approaches to Nonadiabatic Dynamics

2.3.1 Mean-Field (Ehrenfest) Method

As we have discussed in the previous section, nonadiabaticity involves changes in the adiabatic state populations $|a_i|^2$ with changing nuclear configuration. Clearly, such a distortion of the electron cloud will, in turn, influence the nuclear trajectory. Although there are situations in which the impact of electronic nonadiabaticity on nuclear motion is negligible (e.g. for high energy collisions or small energy separations between adiabatic states), for many chemical systems it is of prime importance to properly incorporate electronic–nuclear feedback[81,82].

The simplest way of doing this is to replace the adiabatic potential energy surface E_i in Eq. 27 by the energy expectation value

$$E^{\text{eff}} = \langle\Phi|\mathcal{H}_{\text{el}}|\Phi\rangle = \sum_i |a_i|^2 E_i \quad , \tag{28}$$

where we have used Eq. 18. Thus, the atoms evolve on an effective potential representing an average over the adiabatic states weighted by their state populations $|a_i|^2$ (as illustrated in Fig. 3). The method is therefore referred to as mean-field (also known as Ehrenfest) approach.

It is instructive to derive an expression for the nuclear forces either from the gradient of Eq. 28 or using the Hellmann-Feynman theorem

$$\mathbf{F}_K = -\langle\Phi|\boldsymbol{\nabla}_K\mathcal{H}_{\text{el}}|\Phi\rangle \quad . \tag{29}$$

Opting for the latter, we start by writing down the relation

$$\begin{aligned}
\boldsymbol{\nabla}_K\langle\phi_i|\mathcal{H}_{\text{el}}|\phi_j\rangle &= \boldsymbol{\nabla}_K E_i\delta_{ij} && (30)\\
&= \langle\boldsymbol{\nabla}_K\phi_i|\mathcal{H}_{\text{el}}|\phi_j\rangle + \langle\phi_i|\boldsymbol{\nabla}_K\mathcal{H}_{\text{el}}|\phi_j\rangle + \langle\phi_i|\mathcal{H}_{\text{el}}|\boldsymbol{\nabla}_K\phi_j\rangle && (31)\\
&= \langle\phi_i|\boldsymbol{\nabla}_K\mathcal{H}_{\text{el}}|\phi_j\rangle + (E_j - E_i)\mathbf{d}_{ji} \quad , && (32)
\end{aligned}$$

where we have defined the nonadiabatic coupling vectors, $\mathbf{d}_{ji}$, as

$$\mathbf{d}_{ji} = \langle\phi_j|\boldsymbol{\nabla}_K|\phi_i\rangle \quad , \tag{33}$$

and used Eq. 10 together with the hermiticity of $\mathcal{H}_{\text{el}}$:

$$\langle\phi_i|\mathcal{H}_{\text{el}}|\boldsymbol{\nabla}_K\phi_j\rangle = \langle\boldsymbol{\nabla}_K\phi_j|\mathcal{H}_{\text{el}}|\phi_i\rangle^* = \langle\boldsymbol{\nabla}_K\phi_j|E_j\phi_i\rangle^* = E_i\mathbf{d}_{ij}^* = -E_i\mathbf{d}_{ji} \quad . \tag{34}$$

Note that

$$\mathbf{d}_{ji}^* = -\mathbf{d}_{ij} \quad , \tag{35}$$

because

$$\begin{aligned}
\boldsymbol{\nabla}_K\langle\phi_i|\phi_j\rangle &= \boldsymbol{\nabla}_K\delta_{ij} = 0 && (36)\\
&= \langle\boldsymbol{\nabla}_K\phi_i|\phi_j\rangle + \langle\phi_i|\boldsymbol{\nabla}_K\phi_j\rangle = \mathbf{d}_{ji}^* + \mathbf{d}_{ij} \quad . && (37)
\end{aligned}$$

Equating the rhss of Eqs. 30 and 32 one obtains after rearranging,

$$\langle\phi_i|\boldsymbol{\nabla}_K\mathcal{H}_{\text{el}}|\phi_j\rangle = \boldsymbol{\nabla}_K E_i\delta_{ij} - (E_j - E_i)\mathbf{d}_{ji} \quad . \tag{38}$$

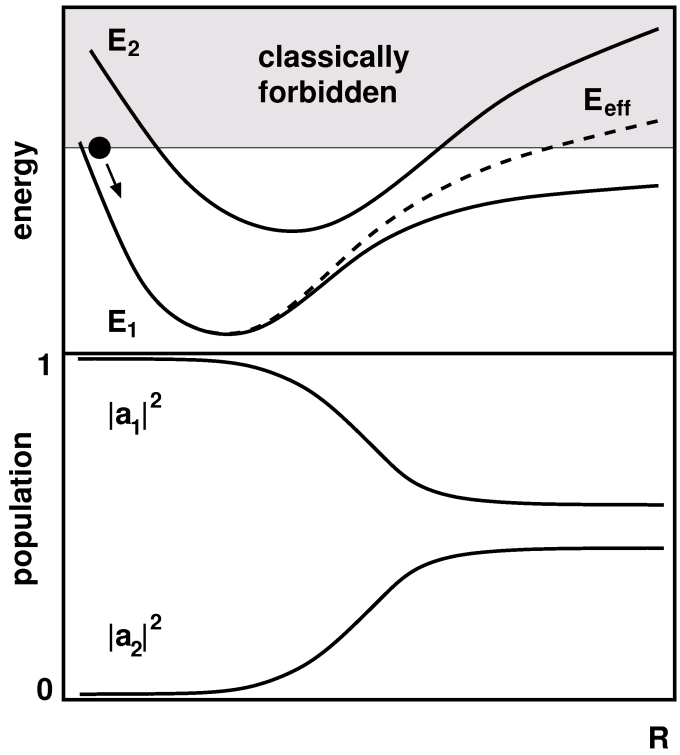


Figure 3. Top: avoided crossing between two adiabatic PES, E_1 and E_2, and effective potential, $E_{\rm eff}$, on which the nuclei are propagated in the Ehrenfest method. In the asymptotic region (right) $E_{\rm eff}$ contains contributions from classically forbidden regions of E_2. Bottom: corresponding adiabatic state populations $|a_1|^2$ and $|a_1|^2$. The system is prepared in state 1 initially with zero kinetic energy. Upon entering the coupling region state 2 is increasingly populated.

The nuclear forces (29) are thus given by

$$\mathbf{F}_K = -\sum_i |a_i|^2 \boldsymbol{\nabla}_K E_i + \sum_{i,j} a_i^* a_j (E_j - E_i) \mathbf{d}_{ji} \quad . \tag{39}$$

Eq. 39 illustrates the two contributions to the nuclear forces; the first term is simply the population-weighted average force over the adiabatic states, while the second term takes into account nonadiabatic changes of the adiabatic state occupations. We would like to point out here that the nonadiabatic contributions to the nuclear forces are in the direction of the nonadiabatic coupling vectors $\mathbf{d}_{ji}$.

The Ehrenfest method has been applied with great success to a number of chemical problems including energy transfer at metal surfaces[83]. However, due to its mean-field character the method has some serious limitations. A system that was initially prepared in a pure adiabatic state will be in a mixed state when leaving the region of strong nonadiabatic coupling. In general, the pure adiabatic character of the wavefunction cannot be recovered even in the asymptotic regions of configuration space. In cases where the differences in the adiabatic potential energy landscapes are pronounced, it is clear that an

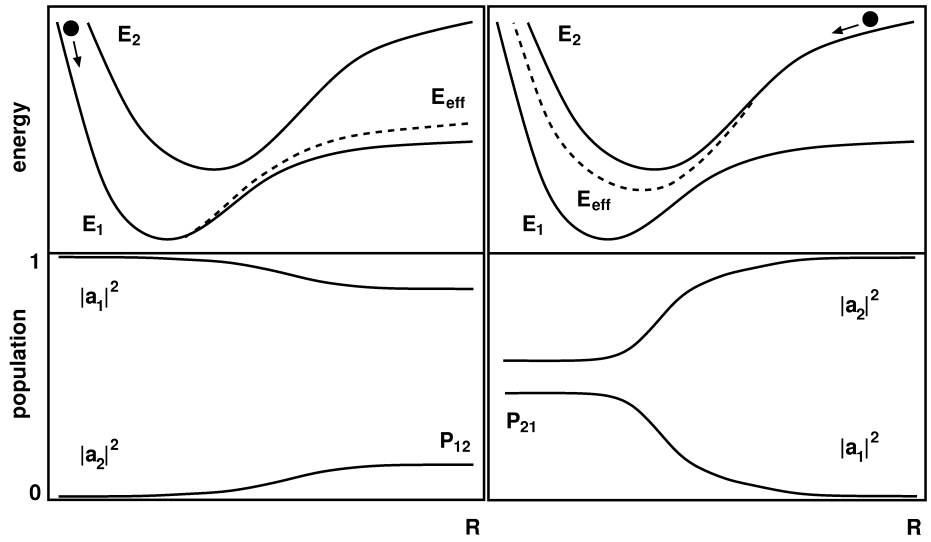


Figure 4. Top left: forward path effective potential, $E_{\rm eff}$, for two weakly coupled adiabatic PES, E_1 and E_2. Bottom left: state occupations for a system initially prepared in state 1. The final value of $|a_2|^2$ is equal to the transition probability P_{12}. Top right: backward path effective potential, $E_{\rm eff}$, for two weakly coupled adiabatic PES, E_1 and E_2. Bottom left: state occupations for a system initially prepared in state 2. The final value of $|a_1|^2$ is equal to the transition probability P_{21}.

average potential will be unable to describe all reaction channels adequately. In particular, if one is interested in a reaction branch whose occupation number is very small, the average path is likely to diverge from the true trajectory. Furthermore, the total wavefunction may contain significant contributions from adiabatic states that are energetically inaccessible (see Fig. 3).

Fig. 4 illustrates another severe drawback of the mean-field approach. The principle of microscopic reversibility demands that the forward path probability, $P_{12}^{\rm for} = |a_2^{\rm final}|^2$ for a system that was initially prepared in state 1 to end up in state 2 must be equal to the backward path probability, $P_{21}^{\rm back} = |a_1^{\rm final}|^2$ for a system that was initially prepared in state 2 to end up in state 1. One can easily think of situations, like the one depicted in Fig. 4, for which the effective potentials for the forward and backward paths are very different, resulting also in different populations, $|a_i|^2$. The Ehrenfest method, therefore, violates microscopic reversibility.

It should be noted that the expansion of the total wavefunction in terms of (adiabatic) basis functions (Eq. 18) is not a necessary requirement for the Ehrenfest method; the wavepacket Φ can be propagated numerically using Eq. 17. However, projection of Φ onto the adiabatic states facilitates interpretation. Knowledge of the expansion coefficients, a_i, is also the key to refinements of the method such as the surface hopping technique.

2.3.2 Surface Hopping

We have argued above that after exiting a well localised nonadiabatic coupling region it is unphysical for nuclear motion to be governed by a mixture of adiabatic states. Rather it

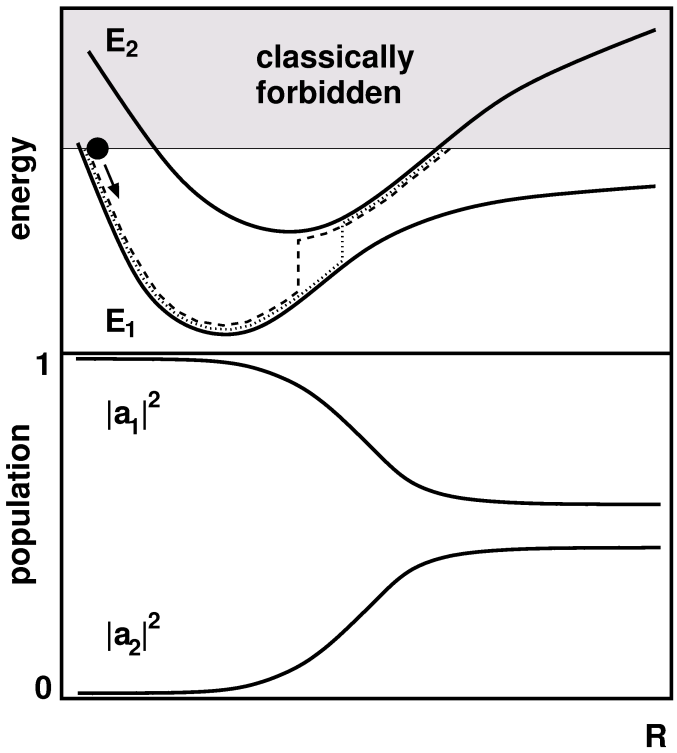


Figure 5. Top: avoided crossing between two adiabatic PES, E_1 and E_2, and two typical forward surface hopping trajectories. Nonadiabatic transitions are most likely to occur in the coupling region. Bottom: corresponding adiabatic state populations $|a_1|^2$ and $|a_1|^2$. The system is prepared to be in state 1 initially with zero kinetic energy. Upon entering the coupling region state 2 is increasingly populated.

would be desirable that in asymptotic regions the system evolves on a pure adiabatic PES. This idea is fundamental to the surface hopping approach. Instead of calculating the "best" (i.e., state-averaged) path like in the Ehrenfest method, the surface hopping technique involves an ensemble of trajectories. At any moment in time, the system is propagated on some pure adiabatic state i, which is selected according to its state population $|a_i|^2$. Changing adiabatic state occupations can thus result in nonadiabatic transitions between different adiabatic PESs (see Fig. 5). The ensemble averaged number of trajectories evolving on adiabatic state i at any time is equal to its occupation number $|a_i|^2$.

In the original formulation of the surface hopping method by Tully and Preston[84], switches between adiabatic states were allowed only at certain locations defined prior to the simulation. Tully[85] later generalized the method in such a way that nonadiabatic transitions can occur at any point in configuration space. At the same time, an algorithm – the so-called fewest switches criterion – was proposed which minimises the number of surface hops per trajectory whilst guaranteeing the correct ensemble averaged state populations at all times. The latter is important because excessive surface switching effectively results in weighted averaging over the adiabatic states much like in the case of the Ehrenfest method.

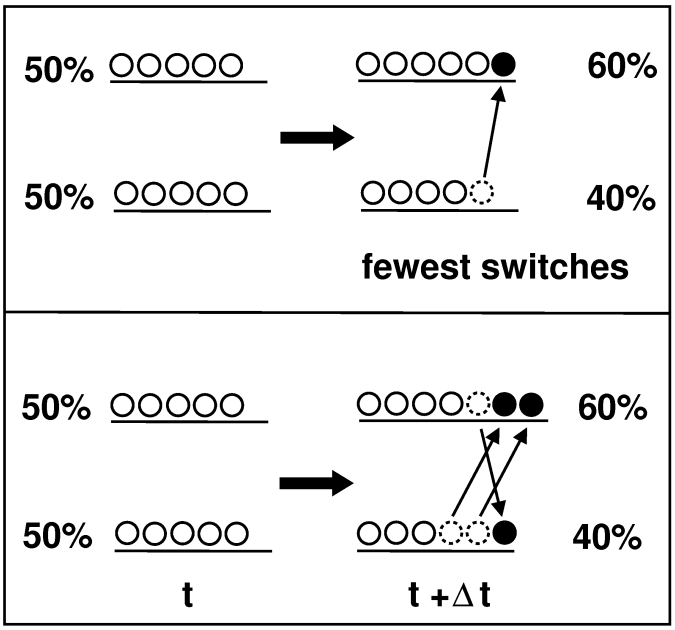


Figure 6. A two-state system with each state being equally (50%) populated at time t. At time $t + \Delta t$ the lower and the upper state are populated by 40 % and 60 % of ensemble members, respectively. The top panel shows how this distribution can be achieved with the minimum number of transitions, whereas the bottom panel shows *one* alternative route involving a larger number of transitions.

We shall now derive the fewest switches criterion. Out of a total of N trajectories, N_i will be in state i at time t,

$$N_i(t) = \rho_{ii}(t)N \quad . \tag{40}$$

Here we have introduced the density matrix notation

$$\rho_{ij}(t) = a_i^*(t)a_j(t) \quad . \tag{41}$$

At a later time $t' = t + \delta t$ the new occupation numbers are

$$N_i(t') = \rho_{ii}(t')N \tag{42}$$

Let us suppose that $N_i(t') < N_i(t)$ or $\delta N = N_i(t) - N_i(t') > 0$. Then the minimum number of transitions required to go from $N_i(t)$ to $N_i(t')$ is δN hops from state i to any other state and zero hops from any other state to state i (see Fig. 6). The probability $P_i(t, \delta t)$ for a transition out of state i to any other state during the time interval $[t, t + \delta t]$ is then given by

$$P_i(t, \delta t) = \frac{\delta N}{N_i} = \frac{\rho_{ii}(t) - \rho_{ii}(t')}{\rho_{ii}} \approx -\frac{\dot{\rho_{ii}}\delta t}{\rho_{ii}} \quad , \tag{43}$$

where we have used

$$\dot{\rho_{ii}} \approx \frac{\rho_{ii}(t') - \rho_{ii}(t)}{\delta t} \quad . \tag{44}$$

The lhs of Eq. 44 can be written as

$$\dot{\rho_{ii}} = \frac{\mathrm{d}}{\mathrm{d}t}(a_i^* a_i) = \dot{a_i^*} a_i + a_i^* \dot{a_i} = (a_i^* \dot{a_i})^* + a_i^* \dot{a_i} = 2\Re(a_i^* \dot{a_i}) \quad . \tag{45}$$

Inserting Eq. 19 into Eq. 45 we obtain

$$\dot{\rho_{ii}} = -2\Re\left(\sum_j \rho_{ij} C_{ij} e^{-\frac{i}{\hbar}\int (E_j - E_i)\mathrm{d}t}\right) \quad . \tag{46}$$

Substituting expression (46) into Eq. 43 the probability P_i can be rewritten as follows

$$P_i(t, \delta t) = \frac{2\Re\left(\sum_j \rho_{ij} C_{ij} e^{-\frac{i}{\hbar}\int (E_j - E_i)\mathrm{d}t}\right) \delta t}{\rho_{ii}} \quad . \tag{47}$$

Since the probability, P_i, for a switch from state i to any other state must be the sum over all states of the probabilities, P_{ij}, for a transition from state i to a specific state j,

$$P_i(t, \delta t) = \sum_j P_{ij}(t, \delta t) \quad , \tag{48}$$

it follows from Eq. 47 that

$$P_{ij}(t, \delta t) = \frac{2\Re\left(\rho_{ij} C_{ij} e^{-\frac{i}{\hbar}\int (E_j - E_i)\mathrm{d}t}\right) \delta t}{\rho_{ii}} \quad . \tag{49}$$

A transition from state i to state k is now invoked if

$$P_i^{(k)} < \zeta < P_i^{(k+1)} \quad , \tag{50}$$

where ζ ($0 \leq \zeta \leq 1$) is a uniform random number and $P_i^{(k)}$ is the sum of the transition probabilities for the first k states,

$$P_i^{(k)} = \sum_j^k P_{ij} \quad . \tag{51}$$

In order to conserve total energy after a surface hop has been carried out, the atomic velocities have to be rescaled. The usual procedure is to adjust only the velocity components in the direction of the nonadiabatic coupling vector $d_{ik}(\mathbf{R})$ (Eq. 33)[85]. We can qualitatively justify this practice by our earlier observation that the nonadiabatic contribution to the Ehrenfest forces also are in the direction of the nonadiabatic coupling vector $d_{ik}(\mathbf{R})$ (see Eq. 39). Certainly, such discontinuities in nuclear velocities must be regarded as a flaw of the surface hopping approach. In most physical scenarios, however, nonadiabatic surface switches take place only at relatively small potential energy separations so that the necessary adjustment to the nuclear velocities is reasonably small. Nevertheless, a severe limitation of the method is presented by its inability to properly deal with situations in which the amount of kinetic energy is insufficient to compensate for the difference in potential energy (so-called classically forbidden transitions). Tully's original suggestion not to carry out a surface hop while retaining the nuclear velocities in such cases has been demonstrated[86] to be more accurate than later proposals to reverse the velocity components

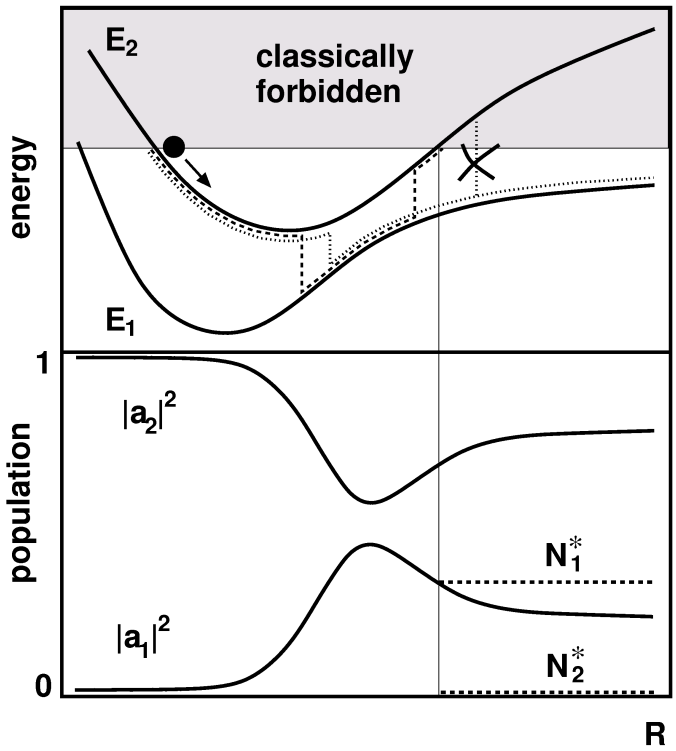


Figure 7. Top: avoided crossing between two adiabatic PES, E_1 and E_2, and two typical forward surface hopping trajectories. Nonadiabatic transitions are most likely to occur in the coupling region. The cross indicates a classically forbidden transition; no switch is carried out in this case. Bottom: corresponding adiabatic state populations $|a_1|^2$ and $|a_1|^2$. The system is prepared in state 2 initially with zero kinetic energy. Upon entering the coupling region state 1 is increasingly populated. Upon exiting the coupling region, state population 1 decreases. For configurations $\mathbf{R}$ for which E_2 is in the classsically forbidden region, the percentages of trajectories in state i, N_i^*, are unequal to $|a_i|^2$; N_2^* is zero whereas N_1^* remains constant.

in the direction of the nonadiabatic coupling vector $d_{ik}(\mathbf{R})$[87,88]. The example presented in Figure 7 illuminates how classically forbidden transitions cause divergence between the target occupation numbers, $|a_i|^2$, and the actual percentages of trajectories evolving in state i, N_i^*.

It should be noted that surface hopping exhibits a large degree of electronic coherence through continuous integration of Eqs. 19 along the entire trajectory. On the one hand, this enables the method to reproduce quantum interference effects[85] such as Stueckelberg oscillations[77]. On the other hand, due to treating nuclei classically, dephasing of the electronic degrees of freedom may be too slow, a shortcoming shared by the surface hopping and the Ehrenfest method alike. A number of semiclassical approaches incorporating decoherence have, therefore, been proposed[89–95]. Some of these alternative methods attempt to combine the advantages of surface hopping (mainly, pure adiabatic states in asymptotic regions) with those of the mean-field method (no discontinuities in potential energy, no disallowed transitions) by employing an effective potential whilst enforcing gradual demixing of the total wavefunction away from the coupling regions[93–95].

2.4 Nonadiabatic *ab initio* Molecular Dynamics

In the last decade, a number of surface hopping implementations based on "on the fly" *ab initio* molecular dynamics using different electronic structure methods have been published[60,151,152] (for a recent review see Ref. 153). The first general nonadiabatic AIMD surface hopping method was formulated within the framework of density functional theory[60,98,61,99]. It used the Restricted Open-Shell Kohn-Sham (ROKS) method for the excited state and standard Kohn-Sham theory for the ground state. In the two-state formulation, the total electronic wavefunction, Φ, is represented as a linear combination of the S_0 and S_1 adiabatic state functions, ϕ_0 and ϕ_1,

$$\Phi(\mathbf{r},t) = a_0'(t)\phi_0(\mathbf{r},\mathbf{R}) + a_1'(t)\phi_1(\mathbf{r},\mathbf{R}) \tag{52}$$

where the time-dependent expansion coefficients $a_0'(t)$ and $a_1'(t)$ are to be determined such that Φ is a solution to the time-dependent electronic Schrödinger equation (17). The prime indicates that the coefficients now include the exponential factor, i.e. $a_k'(t) = a_k(t)e^{-\frac{i}{\hbar}\int E_k \mathrm{d}t}$, for the sake of compactness.

In the present case, our adiabatic basis functions are the S_0 closed-shell Kohn-Sham ground state determinant,

$$\phi_0 = |\varphi_1^{(0)}\bar{\varphi}_1^{(0)}\cdots\varphi_n^{(0)}\bar{\varphi}_n^{(0)}\rangle \tag{53}$$

and the orthonormalized S_1 wavefunction

$$\phi_1 = \frac{1}{\sqrt{1-S^2}}\left[-S\phi_0 + \phi_1'\right] \tag{54}$$

where

$$S = \langle\phi_0|\phi_1'\rangle \tag{55}$$

is the overlap between the ground state wavefunction and the ROKS excited state wavefunction[100–102]

$$\phi_1' = \frac{1}{\sqrt{2}}\left\{|\varphi_1^{(1)}\bar{\varphi}_1^{(1)}\cdots\varphi_n^{(1)}\bar{\varphi}_{n+1}^{(1)}\rangle + |\varphi_1^{(1)}\bar{\varphi}_1^{(1)}\cdots\bar{\varphi}_n^{(1)}\varphi_{n+1}^{(1)}\rangle\right\} \tag{56}$$

n being half the (even) number of electrons. Separate variational optimization of ϕ_0 and ϕ_1' generally results in nonorthogonality, the molecular orbitals $\varphi_l^{(0)}$ and $\varphi_l^{(1)}$ are different. Please note, however, that for small S, $\phi_1 \approx \phi_1'$.

Substitution of ansatz (52) into (17) and integration over the electronic coordinates following multiplication by $\phi_k^*\quad(k=0,1)$ from the left yields the coupled equations of motion for the wavefunction coefficients

$$\dot{a}_k'(t) = -\frac{i}{\hbar}a_k'(t)E_k - \sum_l a_l'(t)D_{kl}\quad(k,l=0,1) \tag{57}$$

where E_k is the energy eigenvalue associated with the wavefunction ϕ_k. For the nonadiabatic coupling matrix elements

$$D_{kl} = \langle\phi_k|\frac{\partial}{\partial t}|\phi_l\rangle \tag{58}$$

the relations $D_{kk} = 0$ and $D_{kl} = -D_{lk}$ hold, as our ϕ_k are real and orthonormal.

In the Car-Parrinello molecular dynamics (CP-MD) formalism[1,2], computation of the nonadiabatic coupling elements, D_{kl}, is straightforward and efficient, since the orbital velocities, $\dot{\varphi}_l$, are available at no additional cost due to the underlying dynamical propagation scheme. If, instead of being dynamically propagated, the wavefunctions are optimized at each point of the trajectory (so-called Born-Oppenheimer mode), the nonadiabatic coupling elements are calculated using a finite difference scheme.

Numerical integration of (57) yields the expansion coefficients a'_k, whose square moduli, $|a'_0|^2$ and $|a'_1|^2$, can be interpreted as the occupation numbers of ground and excited state, respectively.

Following Tully's *fewest switches criterion*[85] recipe, the nonadiabatic transition probability from state k to state l is

$$\Pi_{kl} = \max(0, P_{kl}) \tag{59}$$

with the transition parameter

$$P_{kl} = -\delta t \, \frac{\frac{d}{dt}|a'_k|^2}{|a'_k|^2} \tag{60}$$

where δt is the MD time step.

A hop from surface k to surface l is carried out when a uniform random number $\zeta < \Pi_{kl}$ provided that the potential energy E_l is smaller than the total energy of the system. The latter condition rules out any so-called classically forbidden transitions. After each surface jump atomic velocities are rescaled in order to conserve total energy. In the case of a classically forbidden transition, we retain the nuclear velocities, since this procedure has been demonstrated to be more accurate than alternative suggestions[86] .

The two-state surface hopping formalism presented here can be easily generalized to include multiple excited states[85]. However, calculating a large number of electronic states including nonadiabatic couplings between them from first principles is often either not straightforward or too computationally demanding in practice.

2.5 Nonadiabatic QM/MM Molecular Dynamics

The nonadiabatic AIMD approach of the previous Section has the same limitations with regard to system size as regular, adiabatic, AIMD. Therefore, in order to be able to study light-induced processes in complex environments, the method has been extended and implemented in a QM/MM framework, in which the electrons in the QM subsystem are represented by a total wavefunction $\Phi^{\mathrm{QM/MM}}$ satisfying the time-dependent Schrödinger equation (17), while the photochemically inert environment is described by an analytical force field. The QM/MM coupling is established via a Hamiltonian[62] $\mathcal{H}^{\mathrm{QM/MM}}$ which is a function of *all* the nuclear coordinates, i.e. both of the QM and the MM subsystems. Likewise, the total wavefunction, $\Phi^{\mathrm{QM/MM}}$, depends on the entire set of nuclear coordinates and is expanded

$$\Phi^{\mathrm{QM/MM}}(\mathbf{r}, \mathbf{R}, t) = \sum_i a'_i(t) \phi_i^{\mathrm{QM/MM}}(\mathbf{r}, \mathbf{R}) \tag{61}$$

in terms of known electronic state functions, $\phi_i^{\mathrm{QM/MM}}(\mathbf{r}, \mathbf{R})$. The time-dependent expansion coefficients, $a'_i(t)$, are determined by inserting this ansatz into the time-dependent

Schrödinger equation (17), resulting in a system of coupled differential equations,

$$\dot{a}'_i(t) = -\frac{i}{\hbar} a'_i(t) E_i^{\mathrm{QM/MM}} - \sum_j a'_j(t) C_{ij}^{\mathrm{QM/MM}} \tag{62}$$

where $E_i^{\mathrm{QM/MM}}$ is the energy of electronic state i, and

$$C_{ij}^{\mathrm{QM/MM}} = \langle \phi_i^{\mathrm{QM/MM}} | \frac{\partial}{\partial t} | \phi_j^{\mathrm{QM/MM}} \rangle = \dot{\mathbf{R}} \langle \phi_i^{\mathrm{QM/MM}} | \frac{\partial}{\partial \mathbf{R}} | \phi_j^{\mathrm{QM/MM}} \rangle \tag{63}$$

are the nonadiabatic couplings between states i and j.

Here we discuss a two-state implementation which couples nonadiabatically the closed-shell Kohn-Sham ground state, $\phi_0^{\mathrm{QM/MM}}$, to the re-orthonormalized ROKS representation[100,103] of the S_1 first singlet excited state, $\phi_1^{\mathrm{QM/MM}}$, following the successful single-scale na-QM technique[60,104–109]. As a two-determinant representation (for reviews see Refs. 2,61,3), the ROKS S_1 state provides an improved reference to compute nonadiabatic couplings[110], and yields reliable S_1 nonradiative lifetimes and decay mechanisms when nonadiabatically coupled to the KS ground state[60,104–109]; the level of accuracy obtained by this efficient approach for the system of interest to this study, i.e. AB photoisomerisation, will be demonstrated in detail in Sec. 3.1.2.

In addition to the nonadiabatic QM component, the QM/MM approach also requires a force field parameterization suitable for condensed phase simulations[46] to define the MM part of $\mathcal{H}^{\mathrm{QM/MM}}$. For the QM↔MM electrostatic coupling terms of $\mathcal{H}^{\mathrm{QM/MM}}$ it is important to include the correct electron density of the state the system is propagated in according to the surface hopping prescription.

2.6 Classical Atomistic Molecular Dynamics

Although, strictly speaking, the term classical molecular dynamics refers to the fact that the atomic nuclei are treated as classical particles, it is also synonymous with methods that use analytical interatomic interaction potentials (known as force fields). Generally, an empirical force field consists of terms that model the non-bonded interactions (E_{nonbond}), which include both the van der Waals and Coulombic interactions, the bond interactions (E_{bond}), the angle bending interactions (E_{angle}) and the dihedral (bond rotations) interactions (E_{dihedral}):

$$E(\mathbf{R}) = E_{\mathrm{nonbond}} + E_{\mathrm{bond}} + E_{\mathrm{angle}} + E_{\mathrm{dihedral}}. \tag{64}$$

A large number of force fields, parametrised with different applications in mind, exist. Because we have employed it in our multiscale simulations presented below, we discuss here the Gromos force field[111] in which covalent bonds are described by

$$E_{\mathrm{bond}} = \sum_i^{N_b} \frac{1}{4} k_{b,i} (R_i^2 - R_{0,i}^2)^2 \tag{65}$$

$$E_{\mathrm{angle}} = \sum_i^{N_\theta} \frac{1}{2} k_{\theta,i} (\cos\theta_i - \cos\theta_{0,i})^2 \tag{66}$$

$$E_{\mathrm{dihedral}} = E_{\mathrm{dih,improp}} + E_{\mathrm{dih,prop}} \tag{67}$$

where

$$E_{\mathrm{dih,improp}} = \sum_{i}^{N_\xi} \frac{1}{2} k_{\xi,i}(\xi_i - \xi_{0,i})^2 \tag{68}$$

and

$$E_{\mathrm{dih,prop}} = \sum_{i}^{N_\phi} k_{\phi,i}[1 + \cos(\delta_i)\cos(m_i\phi_i)] \tag{69}$$

The nonbonded interactions are given by

$$E_{\mathrm{nonbond}} = E_{\mathrm{vdW}} + E_{\mathrm{el}} \tag{70}$$

where

$$E_{\mathrm{vdW}} = \sum_{i}^{N_p} \frac{C_{12,i}}{R_i^{12}} - \frac{C_{6,i}}{R_i^6} \tag{71}$$

and

$$E_{\mathrm{el}} = \sum_{i}^{N_p} \frac{Q_i}{4\pi\epsilon_0 R_i} \tag{72}$$

Q_i being the product of the two atomic charges involved in pair i.

2.7 Coarse Grained Classical Molecular Dynamics

A number of different approaches have been proposed for systematically constructing CG models in a way suitable for multiscale simulations, where one wants to switch back and forth between different levels of resolution. Usually, coarse graining efforts underly the constraint that they must reproduce structural and thermodynamic properties of a higher resolution model.

In one class of coarse graining methods the derivation of CG interaction potentials uses thermodynamic properties such as energies or free energies as a reference[14,16]. While this approach yields a good CG description of certain thermodynamic properties it is less suited to describe the structure of the system[112]. The other class are the so-called structure-based methods, where the CG interactions are fitted such that the model reproduces certain structural properties – often described by a set of radial distribution functions obtained from all-atom molecular simulations[12,13,44,20]. While these structure-based methods are able to reproduce local structures, and are thus well suited to reinsert atomistic coordinates, they are less reliable when it comes to predicting thermodynamic properties. There is also uncertainty as to how well they reproduce higher-order (e.g. three-body) structural correlations if they are not included explicitly in the parameterization process[113]. In the so-called force matching method[114–116] the difference between the (instantaneous) CG forces and the underlying atomistic forces are minimized. The thus optimized CG interactions can be

shown to reproduce a many-body multidimensional potential of mean force of the underlying atomistic system. Therefore the force matching method is related to the structure-based CG methods, the only difference being that the latter usually rely on pair distribution functions, i.e. pair potentials of mean force[17,113,37,117]. Current research investigates the limitations of the different approaches and explores possible hydrid CG potentials that are both thermodynamically and structurally consistent with an atomistic description[15,118] and transferable to different state points or system compositions[117,18,119].

A structure-based method was used to parameterize the CG model for the liquid crystalline 8AB8 system[44] and we dedicate the remainder of this Section to discussing this approach in more detail. Typically one distinguishes between bonded/covalent and non-bonded CG potentials, which are developed separately. The total potential energy U^{CG} can then be written as

$$U^{CG} = \sum U_B^{CG} + \sum U_{NB}^{CG} \tag{73}$$

Bonded interactions are tuned in such a way that they yield the correct conformational statistics of the molecules in the CG simulation. In addition to being temperature dependent, the conformational distributions P^{CG} are usually functions of specific bond lengths R, angles θ, and torsions ϕ between any pair, triple and quadruple of CG beads respectively, i.e. $P^{CG}(R,\theta,\phi,T)$. The reference distributions are obtained from atomistic simulations (see below). Assuming that the different CG internal degrees of freedom are uncorrelated, $P^{CG}(R,\theta,\phi,T)$ factorizes into independent probability distributions of bond, angle and torsional degrees of freedom

$$P^{CG}(R,\theta,\phi,T) = P^{CG}(R,T)\ P^{CG}(\theta,T)\ P^{CG}(\phi,T)\ . \tag{74}$$

The corresponding CG potentials are then obtained by Boltzmann inversion of the individual probability distributions $P^{CG}(R,T)$, $P^{CG}(\theta,T)$, and $P^{CG}(\phi,T)$:

$$U^{CG}(R,T) = -k_BT\ \ln\left(P^{CG}(R,T)/R^2\right)\ +\ C_R \tag{75}$$

$$U^{CG}(\theta,T) = -k_BT\ \ln\left(P^{CG}(\theta,T)/\sin(\theta)\right)\ +\ C_\theta \tag{76}$$

$$U^{CG}(\phi,T) = -k_BT\ \ln\ P^{CG}(\phi,T)\ +\ C_\phi\ , \tag{77}$$

C_R, C_θ, and C_ϕ being constants used to shift the respective potential minima to zero.

This factorization can, however, lead to artifacts in the CG model. One way to improve this is by a better choice of coarse units (mapping points). Furthermore, the introduction of intramolecular potentials can alleviate this problem[29].

Structure-based approaches to derive nonbonded CG interaction functions often employ the inverse Monte Carlo or the iterative Boltzmann inversion method[12,120] to numerically generate a tabulated potential that accurately reproduces a given radial distribution function $g(R)$. These methods require an initial guess for a nonbonded potential $U_{NB,0}^{CG}$. This is often taken to be the Boltzmann inverse of the target $g(R)$, i.e. the potential of mean force,

$$U_{NB,0}^{CG} = -k_BT \ln g(R)\ , \tag{78}$$

which is then used to perform a CG simulation of the liquid. This first step will not reproduce the target structure very well, since the potential of mean force is only a good estimate

for the potential energy in the limit of high dilution, due to the neglect of multi-body interactions. The iterative Boltzmann method then refines the CG potential self-consistently until the desired structure is reproduced. It uses the following iteration scheme:

$$U^{CG}_{NB,i+1} = U^{CG}_{NB,i} + k_B T \ \ln(\frac{g_i(R)}{g(R)}) \ , \tag{79}$$

An alternative solution would involve more complex, coupled potentials. Once the mapping is defined, U^{CG} can be determined in an automated way using, for instance, the recently developed VOTCA package[113].

2.8 Adaptive Resolution Molecular Dynamics

We have seen in the previous Section that coarse-graining strategies based on atomistic reference models allow us to reduce the number of degrees of freedom of a problem in order to study properties which are typical of mesoscopic (macroscopic) scales where atomistic resolution does not play a direct role. Unfortunately, there are many situations in soft matter physics in which a complete separation of scales is not possible. Rather, there is a delicate interconnection between them, which impacts on the properties of the system. For instance, in certain regions of a system an interesting phenomenon might occur at the atomistic level, in while the remaining system a coarse-grained resolution is sufficient to describe its equilibrium thermodynamic properties. In such cases it is desirable to properly couple the atomistic description in the interesting region with the coarse-grained description of the rest of the system. For liquids and soft matter, where fluctuations are typically large, the coupling must allow free exchange of particles between the two regions. When a molecule leaves the atomistic region it slowly loses the atomistic degrees of freedom and becomes a coarse-grained molecule, and vice versa from the atomistic to the coarse-grained region. A particular challenge for adaptive resolution theories is the condition that the overall thermodynamic equilibrium of the system should not be perturbed by the fluctuations in the degrees of freedom. Several approaches along these lines have been proposed recently[121,122,70,123], which differ mainly in the way thermodynamic equilibrium is ensured.

We shall briefly describe here the AdResS method[73–76], which has two basic ingredients. Firstly, a transition region is introduced between the atomistic and coarse grained partitions, characterized by a continuous and monotonic switching function $w(x)$ that is zero in the coarse-grained region and unity in the atomistic region (see also Fig. 8). Secondly, the forces in the transition region are taken to be weighted average of the atomistic and coarse-grained forces according to the formula:

$$\mathbf{F}_{\alpha\beta} = w(X_\alpha)w(X_\beta)\mathbf{F}^{\text{atom}}_{\alpha\beta} + [1 - w(X_\alpha)w(X_\beta)]\mathbf{F}^{\text{cg}}_{\alpha\beta} \tag{80}$$

where the indices α and β label two different molecules, $\mathbf{F}^{\text{atom}}_{\alpha\beta}$ is obtained from the atomistic interactions between the atoms of molecule α and those of β, and $\mathbf{F}^{\text{cg}}_{\alpha\beta}$ is obtained from the coarse-grained pair potential between the centres of mass of α and β. The recipe in Eq. 80 leads to a smooth transition from atomistic to coarse-grained trajectories without any major perturbation of the overall evolution of the system. A crucial point concerning Eq. 80 is that, by construction, it obeys Newton's Third Law demanding that the forces of two bodies on each other are always equal and are directed in opposite directions. This

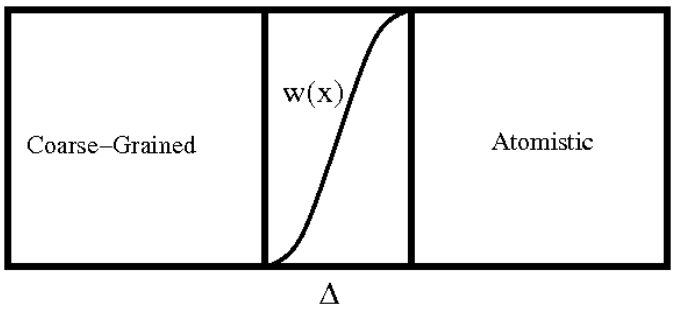


Figure 8. Schematic picture of the AdResS box. Atomistic region on the right, coarse-grained left, and transition region Δ in between. $w(x)$ is the switching function for the transition from a coarse grained to an atomistic resolution and vice versa. Below the example of tetrahedral molecule (test system) that changes resolution according to $w(x)$. This representation is taken from Ref. 73.

adaptive force is, however, insufficient to ensure thermodynamic equilibrium, since it is not conservative. This means that no potential energy expression can be given explicitly. Thus conservation of energy is not ensured and as a consequence the equilibrium of the system cannot be controlled. The problem boils down to the fact that different molecular resolutions (atomistic, coarse-grained and a continuous sequence of hybrid resolutions in the transition region) are coupled, each characterized by its own (intrinsic) chemical potential. As a consequence, if one starts from a homogeneous density (e.g. the stationary state of the full atomistic simulation), the system will evolve towards a stationary state with non-homogeneous density, because the fugacity at the start is not uniform. This is, of course, undesired, as the AdResS simulation should reproduce the uniform density of the full atomistic target system. A solution to this problem has been proposed which involves introducing a thermodynamic force and the coupling to a local thermostat. The thermostat guarantees that locally the right amount of kinetic energy is provided so that the slowly introduced degrees of freedom are at equilibrium with the surrounding as the molecule crosses the transition region. In this way the molecule can enter the atomistic region without encountering any kinetic barriers. An extension of the equipartition theorem to fractional degrees of freedom is used to define the temperature in the transition region[75,76]. The methodology introduced above is derived from basic principles of thermodynamics and statistical mechanics and has been shown to be rather robust in ensuring overall thermodynamic equilibrium in AdResS simulations even for the delicate situation of binary mixtures[124].

It is tempting to apply the interpolation formula to Hamiltonians instead of forces, but this has been shown to violate physical and mathematical principles[125]. The AdResS force interpolation scheme has been applied to a number of different systems including polymers in solution[126], a series of solvated bucky balls ($C_{60} - C_{2160}$)[131], and liquid water[127,128]. It has also been extended to the coupling with a continuum[129,130].

Transferring the above ideas quantum–classical transitions is somewhat tricky, as this would not only require a change in the number of degrees of freedom but also in the physical principle involved. While classical mechanics is governed by deterministic evolution, quantum mechanics is of a probabilistic nature. When electrons are involved, a proper

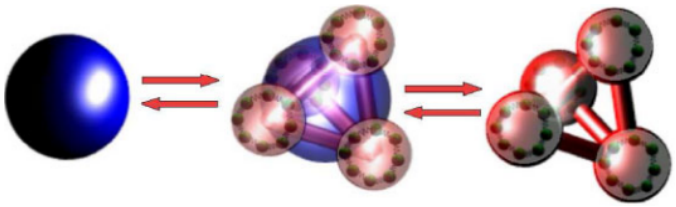


Figure 9. Adaptive resolution from path integral (ring polymer) representation to coarse-grained through a continuous sequence of intermediate resolution for the tetrahedral molecule. A liquid of such molecules was studied with this adaptive resolution; results show that the equilibrium structures in the quantum (path integral) region are the same as in a full path integral simulation. This representation is taken from Ref. 72 .

classical–quantum adaptive scheme based on the Schrödinger equation must be able to deal with the problem of a variable number of particles and thus a varying particle normalization condition as the system evolves. Existing schemes for this case are based on pragmatism rather than providing a complete and consistent theoretical framework[71,69]. If instead the quantum particles are atoms without explicity considering the electrons, then the quantum problem can be mapped on an effective classical one. Thus, the adaptive coupling is between two classical descriptions. The idea is based on the path integral description of atoms in which a quantum atom is represented as a classical polymer ring. In this concept, the beads of the polymer ring are fictitious classical particles. We can now imagine a system, which has atomistic or coarse-grained resolution in a certain region and a path integral resolution in another, where each atom is represented by a polymer ring. In such a situation it is straightforward to apply the principles of AdResS, as this is equivalent to the case of two classical regions with different numbers of degrees of freedom. It has been shown from an application to study equilibrium statistical properties of a liquid of tetrahedral molecules that this method is rather robust both conceptually and numerically[72] (see also Fig. 9).

3 Results and Discussion

3.1 Azobenzene in the Gas Phase

3.1.1 Force Field Development

A new force field suitable for AB was developed using the GROMOS 45a3 force field (see Sec. 2.6) as a starting point, adjusting only the bonded parameters and the charges of the azo group while keeping the original values for the remaining parameters. The parametrization was carried out in such a way as to achieve maximum agreement between the dynamical distributions obtained from force field (MM) and *ab initio* molecular dynamics (QM) simulations in the gas phase at 300 K concerning the relevant bond lengths, bond angles and dihedral angles. This ensures maximum compatibility between the QM and MM descriptions – so that switching adaptively between the two descriptions for a given AB unit is as smooth as possible in future adaptive multiscale applications.

Tab. 1 summarizes our results for the non–standard parameters. In all cases, the values for the force constants are adapted to reproduce the widths of the distributions of bond

entity		force constant	reference
Bonds:	NN	1.40×10^3 kJ/(mol Å^4)	1.2625 Å
	CN	0.72×10^3 kJ/(mol Å^4)	1.4325 Å
Angles:	CNN	650.0 kJ/mol	116.5°
	CCN	560.0 kJ/mol	120.0°
Dihedrals:	CNNC	70.0 kJ/mol	180.0°
	CCNN	6.0 kJ/mol	180.0°
	XCCX	40.0 kJ/mol	180.0°
Point charges:	N		-0.20 e
	$C^{(1)}$		0.20 e
	$C^{(2)}$ – $C^{(6)}$		-0.10 e
	H		0.10 e

Table 1. Non-standard force field parameters derived for azobenzene. For azobenzene structure and atomic numbering scheme see Fig. 10a (X denotes *any* atom).

lengths, angle and dihedrals, starting out from the force field's standard values for chemically similar internal coordinates. In this spirit, we also decided to use the same force constants and point charges for the *cis* and *trans* isomers; only the equilibrium reference values for the bonded potentials differ.

The point charges for the atoms $C^{(1)}$ and N were adapted from average RESP charges[132] computed along the QM reference trajectories. Since the values obtained for the aromatic ring atoms $C^{(2)}$–$C^{(6)}$ and the hydrogen atoms were close to their standard force field values, we decided to use the standard values and thus to take advantage of the resulting small charge groups. Note that in contrast to the GROMOS convention but in line with the AMBER convention[133] and in order to distribute the forces evenly over all contributing atoms we define explicitly *all* dihedral angles involving the two CN bonds, i.e. the dihedral angles $C^{(6)}C^{(1)}NN'$, $C^{(2)}C^{(1)}NN'$, $C^{(6')}C^{(1')}N'N$, and $C^{(2')}C^{(1')}N'N$, which consequently results in smaller force constants per dihedral compared to standard force field values.

It was our aim to derive a single, unified force field for *cis* and *trans* which can be easily applied to study, for instance, a mixture of *trans* and *cis* AB molecules in the condensed phase. In addition, such a force field does not need to be modified during a simulation once a photoinduced *cis*↔*trans* isomerisation has occurred in a preceding nonadiabatic QM/MM simulation. Therefore, in the applications described below, we employ the average values determined for *cis* and *trans* (see Tab. 1). This is a minor approximation, as the only differences between the two isomers concern the NN and CN bond lengths and the CNN bond angle, which are merely 0.015 Å, and 5°, respectively.

The most difficult parameter to adjust was the force constant for the dihedral angle ∠CNNC. Here we settled for the best compromise between sufficient dihedral flexibility and a high enough barrier for the thermal *trans* ↔*cis* isomerization in the ground state. Our new force field yields a barrier along the torsional reaction coordinate of $\approx$ 140 kJ/mol, in reasonable agreement with recent *ab initio* calculations[134] predicting $\approx$ 160 kJ/mol. Increasing the barrier in the force field would yield to an even narrower distribution function for the dihedral CNNC angles. On the other hand, the presently parametrized barrier

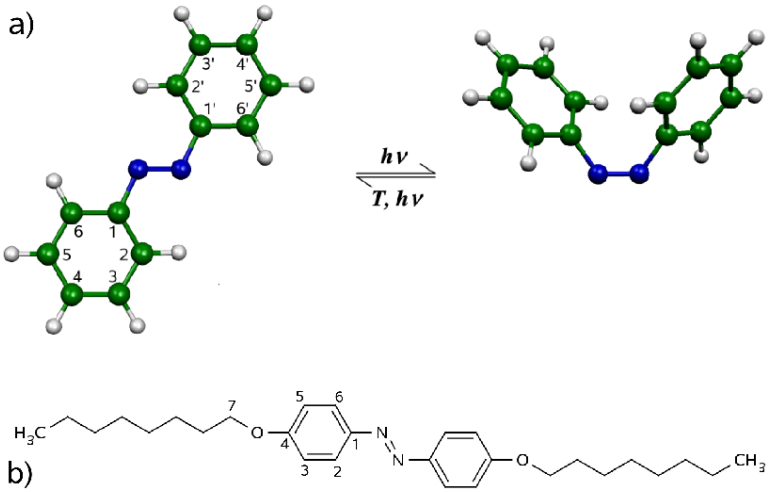


Figure 10. a) Structure and atom numbering scheme of *trans* and *cis* azobenzene in the left and right panels, respectively. b) Chemical structure of 4,4'-dioctyloxy-azobenzene (8AB8).

height is sufficiently large to prevent thermally induced *cis* ↔*trans* isomerizations in the simulations of bulk AB at 400 K (corresponding to an energy of about 3.3 kJ/mol) (see Sec. 3.2.1), which will not occur on the timescale accessible to classical simulation.

The new force field also reproduces the energy difference between the *cis* and *trans* isomers reasonably well. It predicts the *cis* structure to be higher in energy by 36 kJ/mol, which is close to the value of 50 kJ/mol obtained experimentally[135] and from CASPT2 *ab initio* calculations[134].

A detailed comparison of dynamical and optimised structural data from QM and MM calculations can be found in Ref. 46. The data underline the good quality of our parameter set for *trans* and *cis* AB, respectively.

Based on this force field for the AB chromophore we have extended the set of force field parameters to be able to study materials containing this photoswitch such as 8AB8, introduced in Fig. 10, where aliphatic side chains are attached to the phenyl rings of AB via ether bridges. We used methyl-phenyl-ether (H_3C–O–C_6H_5) as a model system to derive the force field parameters necessary to describe the $C^{(4)}$–O–$C^{(7)}$ link unit (see Fig. 10 for atomic numbering scheme) whereas the remainder of these side chains are treated using standard force field parameters.

As for AB itself, a Car–Parrinello run at 300 K was performed with H_3C–O–C_6H_5 as QM reference with the aim to parametrize those internal coordinates that involve the oxygen atom of the ether group. In order to take into account dynamical fluctuations of the $C^{(5)}C^{(4)}OC^{(7)}$ dihedral angle, RESP charges for $C^{(4)}$, O, and $C^{(7)}$ were calculated for different angles between 0° and 90° in steps of 10°. The resulting charge of the methyl group is taken as the charge of the alkyl carbon atom $C^{(7)}$ (united atom approach), and the resulting charge of the aryl carbon atom $C^{(4)}$ is adjusted so as to yield a neutral $C^{(4)}$–O–$C^{(7)}$ unit. Force field point charges were then obtained by Boltzmann averaging over the torsion–angle dependent RESP charges. The resulting parameters for the ether linkage are collected in Tab. 2.

entity		force constant	reference value
bonds:	$OC^{(7)}$	8.18×10^2 kJ/(mol Å^4)	1.430 Å
	$C^{(4)}O$	1.02×10^3 kJ/(mol Å^4)	1.360 Å
angle:	$C^{(4)}OC^{(7)}$	620.0 kJ/mol	116.0°
dihedral:	$C^{(3,5)}C^{(4)}OC^{(7)}$	6.0 kJ/mol	180.0°
point charges:	O		-0.332 e
	$C^{(7)}$		0.178 e
	$C^{(4)}$		0.154 e

Table 2. Extension of azobenzene force field to include ether linkage $C^{(4)}$–O–$C^{(7)}$ (see Fig. 10 for structure and atomic numbering scheme).

3.1.2 Photoisomerisation of Azobenzene

The photoisomerisation of AB was studied using the nonadiabatic AIMD method introduced in Sec. 2.4. Ten surface hopping trajectories were calculated for both directions, $trans \rightarrow cis$ and $cis \rightarrow trans$, sampling the initial conditions randomly from ground state AIMD runs at 300 K.

Potential Energy Landscapes It is important for the discussion below of the photoisomerisation dynamics and mechanism to determine first the shape of the ground and excited state potential landscapes, and to assess the quality of the ROKS S_1 PES used in the na-AIMD and na-QM/MM simulations. We have therefore calculated ROKS, CASSCF, and CASPT2

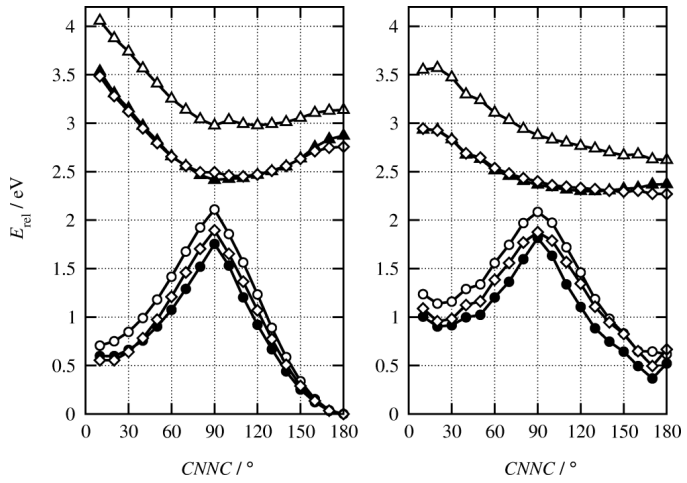


Figure 11. S_0 and S_1 energy profiles along the CNNC dihedral angle using structures optimized using DFT (PBE) for the S_0 ground state (left) and ROKS (PBE) for the S_1 first excited state (right). The ROKS energies (▲—▲) are compared to the state-averaged CAS(14,12) (△—△) and CASPT2 (◇—◇) data. S_0 ground state energies are shown for DFT (●—●), state-averaged CAS(14,12) (○—○) and CASPT2 (◇—◇). All energies are relative to the respective S_0 energies of *trans*-AB optimized with DFT in the S_0 state; 0.7 eV have been added to the ROKS energies as explained in Sec. 3.1.2.

energy profiles along the CNNC dihedral angle, which is an important internal coordinate for photoisomerisation in the S_1 state[64]. On the left hand side of Fig. 11 we present DFT, CASSCF, and CASPT2 S_0 and ROKS, CASSCF, and CASPT2 S_1 potential energy curves for the DFT ground state minimum energy path (MEP) along the CNNC dihedral angle, θ. Note that the ROKS energies have been corrected upwards by constant shift of 0.7 eV which was originally determined from the difference between ROKS and experimental vertical excitation energies[64]. All ground state curves in Fig. 11 have a maximum at 90°, the DFT curve being very close to the CASPT2 curve, while CASSCF is seen to overestimate by about 0.2 eV compared to CASPT2 over a wide range of θ. The agreement between the ROKS and CASPT2 excited state curves throughout the entire range of the isomerization coordinate θ is remarkable, whereas the CASSCF S_1 curve is very similar in shape but shifted upwards by about 0.5 eV.

The right hand side of Fig. 11 shows ground and excited state energy profiles along the ROKS optimised S_1 excited state MEP along the CNNC dihedral. It is seen that the ROKS curve has a shallow S_1 minimum around $\theta \approx 120°$ which can be reached by a barrierless path from both the *cis* and *trans* Franck-Condon (FC) points at $\theta = 10°$ and $\theta = 180°$, respectively. As for the S_0 MEP, there is again striking agreement between the ROKS and CASPT2 excited state curves, while the CASSCF energies are higher than the CASPT2 data by 0.3–0.5 eV. With regards to the differences between the $trans \rightarrow cis$ and $cis \rightarrow trans$ photoisomerisation dynamics which we shall discuss below, it is important to realise that there is a considerably larger potential energy difference between the FC point and the S_1 global minimum upon vertical excitation of the *cis* isomer as compared to *trans*-AB.

At this stage it is concluded, based on the direct comparison to CASPT2 reference data

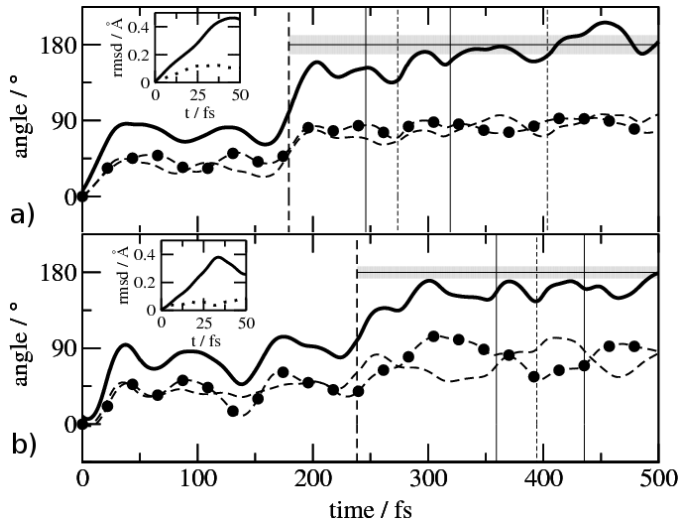


Figure 12. Time evolution of $\psi^{NN'}$ (—), ψ^{N} (- -), and -$\psi^{N'}$ (•–•) for typical $cis \rightarrow trans$ trajectories in the gas phase (a) and the liquid (b). Vertical lines indicate $S_1 \rightarrow S_0$ (- -) and $S_0 \rightarrow S_1$ (—) hops. Grey bars indicate the $\psi^{NN'}$ *trans* reference range. The insets show the rmsd's of N, N′ (solid) and R, R′ (dashed) relative to $t = 0$.

in Fig. 11, that the aforementioned constant blue-shift applied to the ROKS data indeed corrects consistently the gap not only at 90° but along the entire MEPs along the CNNC dihedral angle, which is a major component of the reaction coordinate of AB photoisomerization, both in the S_0 and in the S_1 in the full range between 0° and 180°.

Definition of internal coordinates To analyze in detail the isomerisation mechanism and possible differences between the gas and the liquid phase, we describe internal motion in terms of the plane normal vectors $\mathbf{n}^R$ and $\mathbf{n}^{R'}$, of the two aromatic rings at their geometric centres, R and R′, together with the normal vectors $\mathbf{n}^N$ and $\mathbf{n}^{N'}$ of the $C^{(1)}NN'$ and $NN'C^{(1')}$ coordination planes at N and N′. To measure torsion of the $C^{(1)}NN'$ and $NN'C^{(1')}$ coordination planes, we define an intrinsic (right-handed) coordinate system whose origin is at the geometric midpoint of the two nitrogen atoms, the x-axis is parallel to the N=N′bond and the z-axis is parallel to the arithmetic mean of $\mathbf{n}^N$ and $\mathbf{n}^{N'}$. The absolute torsion is then monitored as the change in angle of the projection of the normal vector $\mathbf{n}^\alpha$ onto the yz plane, $\mathbf{n}^\alpha_{yz}$, $\psi^\alpha(t) = \angle(\mathbf{n}^\alpha_{yz}(t), \mathbf{n}^\alpha_{yz}(0))$, while $\psi^{\alpha\beta}(t) = \angle(\mathbf{n}^\alpha(t), \mathbf{n}^\beta(t))$, with $\alpha, \beta \in \{\mathrm{N}, \mathrm{N}', \mathrm{R}, \mathrm{R}'\}$, gives relative changes, e.g. $\psi^{NN'}$ is the $C^{(1)}NN'C^{(1')}$ dihedral angle and ψ^{RN} captures rotation of the phenyl rings.

Cis→trans In this section, we analyze in detail the mechanism of *cis*-AB to *trans*-AB photoisomerisation. In Fig. 12a we present the time evolution of $\psi^{NN'}$ (i.e. the CNNC

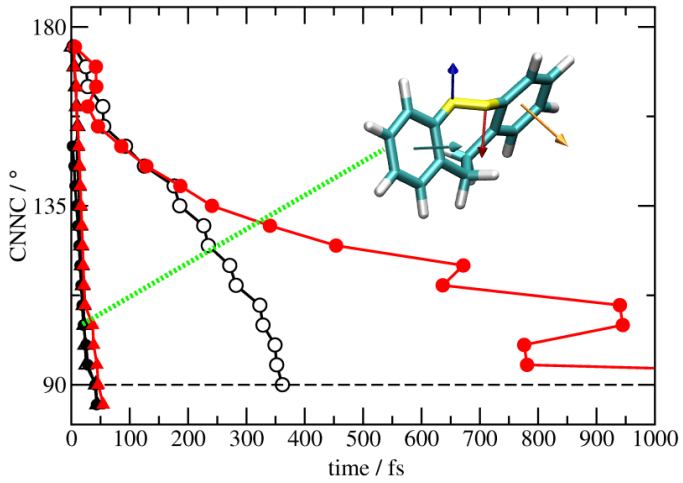


Figure 13. Ensemble average of the CNNC dihedral angle, $\psi^{NN'}(t)$, during time–evolution in the first excited state, S_1, where the specific value $\psi^{NN'}$ at time t is reached for the first time after vertical photoexcitation from S_0 at $t = 0$: gas phase *trans*-AB-C_2 (•–•), gas phase *trans*-AB (○–○), gas phase *cis*-AB (△–△) liquid phase *trans*-AB (•–•), liquid phase *cis*-AB (▲–▲). In the case of *cis*-AB, $180° - \psi^{NN'}$ is plotted. The perpendicular conformation, i.e. $\psi^{NN'} = 90°$, is marked by the horizontal dashed line. For liquid *trans*-AB, the non-monotonicity arises from the fact that ensemble averages were obtained with different numbers of trajectories. The inset shows the optimised structure of *trans*-AB-C_2; normal vectors indicate the orientation of the phenyl rings, $\mathbf{n}^\mathrm{R}$ and $\mathbf{n}^{\mathrm{R}'}$ (cyan, orange), and of the CNN coordination plane of the N atoms, $\mathbf{n}^\mathrm{N}$ and $\mathbf{n}^{\mathrm{N}'}$ (red, blue).

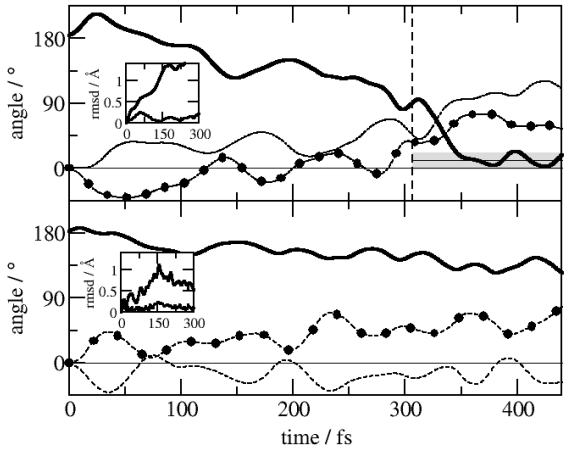


Figure 14. Time evolution of $\psi^{NN'}$(—), ψ^{N}(- -), and -$\psi^{N'}$(•–•) for typical $trans \rightarrow cis$ trajectories in the liquid (top) and in the gas phase (bottom). Grey bars indicate the $\psi^{NN'}$ *cis* reference range. See caption of Fig. 12.

dihedral) for a typical $cis \rightarrow trans$ trajectory together with the corresponding time evolution of ψ^{N} and $\psi^{N'}$. It is seen that $\psi^{NN'}$ initially changes rapidly and after about 30 fs reaches a value of $\approx 90°$. After the $S_1 \rightarrow S_0$ transition to the ground state, $\psi^{NN'}$ reaches a value of $\approx 180°$, thus indicating a successful $cis \rightarrow trans$ isomerisation. Inspection of the order parameters ψ^{N} and $\psi^{N'}$ (cf. Fig. 12a) reveals that the total change in $\psi^{NN'}$ is due to equal contributions from the two coordination planes at N and N′ in opposite directions. Photoisomerisation is dominated by a pedal motion of the CNNC group and *not* by large amplitude rotation of the phenyl rings. This is illustrated by the inset of Fig. 12a which shows the rmsd's of the N atoms and the phenyl ring centres R. We can see that during the first 30 fs it is mainly the translocation of the N atoms that is responsible for the change in $\psi^{NN'}$ by $\approx 90°$, while the phenyl rings remain largely fixed in space. Fig. 13 shows the ensemble averaged changes of the CNNC angle (i.e. $\psi^{NN'}$) as a function of time after vertical photoexcitation. The average time it takes a molecule to reach $\psi^{NN'} = 90°$ is just 42 fs.

Trans→cis The time evolution of $\psi^{NN'}$ in the $trans \rightarrow cis$ case is displayed in Fig. 14a. It shows a more or less smooth decrease from $180°$ at $t = 0$ to $\approx 90°$ at $t \approx 300$ fs, where a $S_1 \rightarrow S_0$ transition occurs. After the hop $\psi^{NN'}$ rapidly falls to a value close to $0°$, thus indicating a successful $trans \rightarrow cis$ isomerisation. As observed for the $cis \rightarrow trans$ isomerisation, the total change of $\approx 180°$ in $\psi^{NN'}$ is eventually accomplished by equal contributions ($\approx 90°$) from ψ^{N} and $\psi^{N'}$ (dashed and circled lines), in *opposite* directions. Interestingly, however, initially both nitrogen coordination planes rotate in the *same* direction, thus producing no net change of $\psi^{NN'}$ (e.g. at $t = 50$ fs). This is a crucial difference from the $cis \rightarrow trans$ photoisomerisation mechanism.

As for the $cis \rightarrow trans$ isomerisation, the change in $\psi^{NN'}$ is produced by the translocation of the N atoms, again indicating a pedal–motion–like mechanism (see inset of Fig. 14a). Interestingly, a similar translocation at fixed $\psi^{NN'}$ has been found in x-ray diffraction experiments of azobenzene crystals[136].

Being the characteristic feature also of the $trans \rightarrow cis$ photoisomerisation, the pedal motion of the two N atoms accounts for the experimental finding of fast isomerisation dynamics in rotation–restricted *trans*-AB[137] and clearly rules out an inversion-type mechanism as deduced from resonance raman intensity analysis[138]. Note, that there is no large-amplitude rotation of the phenyl rings involved as has been suggested for the interpretation of fluorescence anisotropy data[139].

The $trans \rightarrow cis$ photoisomerisation is significantly slower than $cis \rightarrow trans$. This is illustrated by the ensemble averaged changes of the CNNC angle shown in Fig. 13. The average time it takes a molecule to reach $\psi^{NN'} = 90°$ is about 360 fs, one order of magnitude longer than for $cis \rightarrow trans$. This observation is in accord with the longer S_1 lifetimes measured experimentally for *trans*-AB compared to *cis*-AB[140].

3.1.3 Photoisomerisation of Chemically Bridged Azobenzene

Recently, a greatly enhanced $trans \rightarrow cis$ quantum yield $\Phi^{\mathrm{AB-C_2}}_{trans \rightarrow cis}$ was reported[141] for a bridged azobenzene (AB-C_2 in Fig. 13) in the S_1 state as compared to the parent molecule AB. This finding seemed surprising at first as the structural changes involved in isomerisation should be expected to be hindered by the restriction due to the presence of a bridge interconnecting the phenyl rings. Subsequent na-AIMD simulations[142] then demonstrated that counterintuitively the bridge does not hinder photoisomerization. On the contrary, it suitably pre-orients the phenyl rings such that AB-C_2 can more easily undergo $trans \rightarrow cis$ isomerization thus yielding not only an enhanced quantum yield $\Phi^{\mathrm{AB-C_2}}_{trans \rightarrow cis}$ but also ultra–short S_1 lifetimes.

Upon chemical modification of AB to form AB-C_2 (i.e. addition of a –CH_2–CH_2– bridge in *ortho* position of the phenyl rings) the *trans* isomer becomes nonplanar, due to the orientation of the phenyl rings, while only minor structural changes arise for the *cis* isomer (Fig. 13). The mechanical strain introduced into *trans*-AB-C_2 makes it less stable than *cis*-AB-C_2 by 0.31 eV according to DFT (which again compares favourably to the CASPT2 value of 0.35 eV).

Performing nonadiabatic AIMD simulations after vertical $S_0 \rightarrow S_1$ photoexcitation of *trans*-AB-C_2 roughly half of them are found to result in successful $trans \rightarrow cis$ isomerization, the quantum yield being, more precisely, $\Phi^{\mathrm{AB-C_2}}_{trans \rightarrow cis} = (47 \pm 10)\,\%$ including the statistical error of the sample computed using the blocking method[143]. The computed number is consistent with the experimental finding[141] of $\Phi^{\mathrm{AB-C_2}}_{trans \rightarrow cis} = (50 \pm 10)\,\%$. It is noted in passing that the surface hopping method applied here is known to overemphasize coherence[85,144], which is a potential source of error when extracting quantitative information such as quantum yields or excited state lifetimes. However, it has been shown[145] to perform rather well when compared to other approximate methods suitable to simulate complex molecular systems and, moreover, it has been demonstrated[146] that surface hopping quantum yields are only slightly underestimated for bare AB thus validating this approach for the specific case. In stark contrast to *trans*-AB-C_2, the parent compound AB features a much lower experimental value of $\Phi^{\mathrm{AB}}_{trans \rightarrow cis} = 24\,\%$[141], which is again in harmony with

nonadiabatic AIMD where only roughly 20 % of the trajectories yielded isomerization[64]. Thus, the computational approach reproduces the observed[141] striking difference between AB-C_2 *versus* AB in terms of quantum yields. The calculations demonstrate that photoisomerisation of *trans*-AB-C_2 is astonishingly fast according to Fig. 13, which shows that *trans*-AB-C_2 reaches the decisive perpendicular conformation, i.e. $\psi^{NN'} = 90°$, in only ≈ 40 fs which is *one order of magnitude faster* than for the corresponding parent *trans*-AB and, in fact, comparable to plain *cis*-AB. Moreover, in contrast to *trans*-AB, *trans*-AB-C_2 is found to rapidly reach the region beyond the ground state barrier to $trans \rightarrow cis$ isomerisation along the CNNC coordinate characterized by $\psi^{NN'} \ll 90°$, thus entering the *cis*-AB-C_2 product potential well leading to successful photoisomerization (the optimized value for the *cis*-AB-C_2 product being $\psi_0^{NN'} = 6.6°$ in S_0). This behavior rationalizes the greatly enhanced isomerization quantum yield found for *trans*-AB-C_2 compared to *trans*-AB consistently both in experiment[141] and in these simulations.

But why is photoisomerization *promoted upon bridging*? First principles simulations of the photoiomerization of AB in the bulk[63,64] (see Sec. 3.2.2) and suspended between gold electrodes[147,148] reveal that spatial confinement and mechanical constraints, respectively, only mildly affect $cis \rightarrow trans$ isomerisation while a pronounced slowing down is seen for $trans \rightarrow cis$. The reason can be traced back to the ultrafast pedal motion of the N atoms yielding a CNNC angle of $\psi^{NN'} \approx 90°$ in the S_1, which competes with the need to achieve co-planarity of the CNN planes with their adjacent rings, $\psi^{RN} \approx 0°$. In this sense, *trans*-AB-C_2 is very similar to $cis \rightarrow trans$, since both have a very similar non-co-planar phenyl ring orientation in the S_0 equilibrium structure ($\psi^{RN} = \psi^{R'N'} \approx 58°$ for *cis*-AB and $\approx 53°$ for *trans*-AB-C_2, see Fig. 13). The energetic reward associated with achieving co-planarity turns out to be a driving force for ultrafast photoisomerisation. The same reasoning can be applied to explain the large difference between *trans*-AB-C_2 and its unbridged parent *trans*-AB. While *trans*-AB-C_2 tries to compensate the initial non-co-planarity of the phenyl rings and the CNN planes, *trans*-AB is initially planar and therefore completely lacks this incentive.

Hence, the "bridging" of AB – commonly viewed as a severe steric hindrance to photoisomerisation – counterintuitively yields a drastically improved photoswitch which isomerizes on a much shorter timescale with a significantly enhanced quantum yield. Extending the same reasoning it is expected that bridging *cis*-AB to yield *cis*-AB-C_2 will have only a minor effect on the mechanism and, therefore, the $cis \rightarrow trans$ photoisomerization of AB-C_2 should be as fast as for unbridged AB. This implies that both $cis \rightarrow trans$ and $trans \rightarrow cis$ photoswitching of AB-C_2 should be similarly effective. In fact, experiment predicts a $cis \rightarrow trans$ quantum yield of 72 ± 4 %, even larger than for $trans \rightarrow cis$[141]. A recent semiempirical nonadiabatic dynamics study of AB-C_2 has confirmed that nonadiabatic relaxation is ultrafast (< 100 fs) for both bridged isomers[149]. According to those simulations, however, the $cis \rightarrow trans$ quantum yield is only 23 %. The authors attribute this underestimation to early surface hops caused by subtle errors in the underlying semiempirical potential energy surfaces.

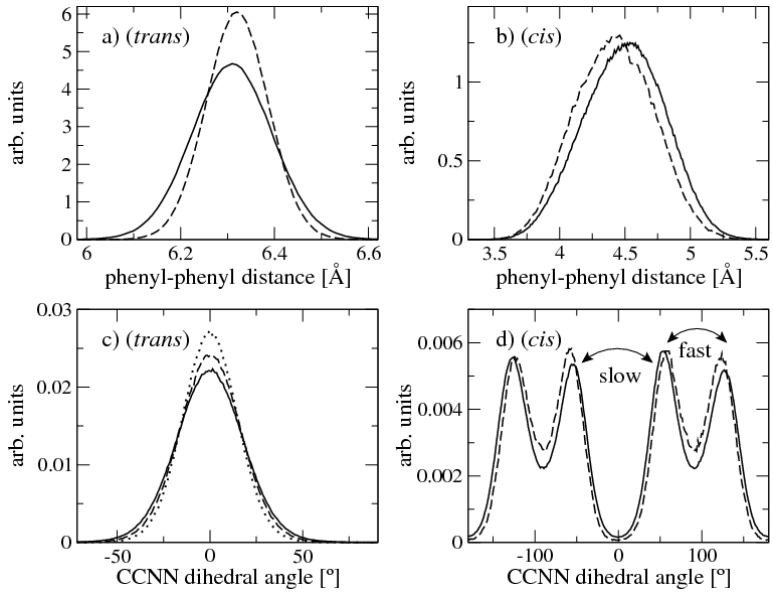


Figure 15. Distributions of structural parameters of the AB unit under various conditions at 400 K. Panels a and b: Distance of the geometrical centres of the two phenyl rings. Panels c and d: CCNN dihedral angle. Panel a: *trans* AB in liquid phase as well as in vacuum – solid line, *trans* 8AB8 in liquid phase (both isotropic and smectic) and in vacuum – dashed line; panel b: *cis* AB in liquid phase – solid line, *cis* AB in vacuum – dashed line; panel c: *trans* AB in liquid phase (equivalent to *trans* AB in vacuum and *trans* 8AB8 in vacuum) – solid line, *trans* 8AB8 in isotropic liquid – dashed line, *trans* 8AB8 in anisotropic (smectic) phase – dotted line; panel d: *cis* AB in liquid phase – solid line, *cis* AB in vacuum – dashed line.

3.2 Liquid Azobenzene

3.2.1 Classical Molecular Dynamics in the Ground State

Parametrizations of the azo group had been carried out at 300 K having in mind future applications of the force field at ambient conditions. The present applications to the study of liquid AB (with a melting point of 341 K) and liquid crystalline AB-containing compounds (with phase transition temperatures of 8AB8 between 372 and 385 K) required testing the validity of the classical force field at an increased temperature of 400 K.

We therefore applied the new force field to study liquid AB at 400 K and analyze separately the influence of the liquid environment on the structural properties of the *cis* and *trans* conformers of AB by comparison with gas phase simulations at 400 K. As a measure for the extension of the AB unit serves the distribution of the distance between the geometric centres of the two phenyl rings as shown in Fig. 15a) and b). For *trans* AB the distributions of the single molecule and the liquid phase are indistinguishable (solid line in Fig. 15a), whereas the conformations of the *cis* isomer are slightly more affected by the bulk environment (see Fig. 15b). In the liquid phase the *cis* AB unit is slightly stretched out compared to the vacuum simulations. Similar observations can be made when analyzing the out-of-plane motions of the phenyl rings by monitoring the distribution functions of the dihedral angle between the normal vectors of the two phenyl rings (data not shown) and

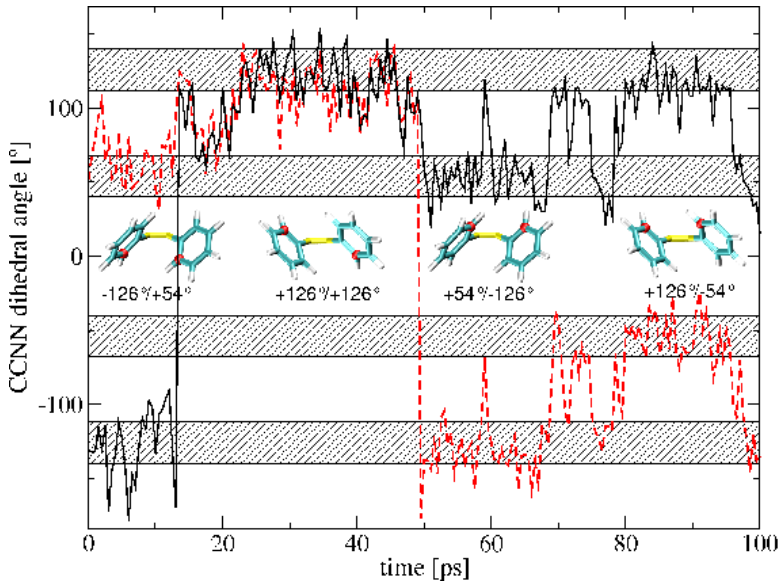


Figure 16. Time evolution of the CCNN dihedral angles of a *cis* AB molecule in liquid environment at 400 K. The striped bars indicate the regions ($\pm 14^\circ$ around the maxima of the distributions at $\pm 54^\circ$ and $\pm 126^\circ$, see Fig. 15d) used to count the transitions between the states (see text). Snapshots of typical conformations are included with the $C^{(2)}$ and $C^{(2')}$ carbon atoms that are used to define the CCNN dihedral angles marked in red.

of the CCNN dihedral angle (see Fig. 15c and d). The conformations in *trans* AB are not affected by the liquid environment (solid line in Fig. 15c), whereas in the case of the *cis* isomer the amplitude of the ring motion (compared to a planar structure) is slightly larger in the isolated molecule than in the bulk liquid at the same temperature (see Fig. 15d). Additionally, it was found that the distribution functions of the CNNC dihedral angle are not affected by the liquid environment – neither in the case of the *trans* nor in the case of the *cis* isomer (data not shown), and it was verified that the system does not undergo thermal *cis* $\leftrightarrow$*trans* isomerisation, which is a rare event that indeed should not occur on the timescale presently accessible by such classical molecular dynamics simulations.

Fig. 15d shows that the distribution of the CCNN dihedral angle in *cis* AB has four chemically equivalent maxima around $\pm 54^\circ$ and $\pm 126^\circ$ and consequently two types of transitions between these states. As indicated in the figure, there is one "fast" type of transition where the phenyl ring is intermediately standing perpendicular to the plane spanned by the $C^{(1)}$ (or the $C^{(1')}$) carbon and the two N atoms, and one "slow" type of transition, where the ring is intermediately in–plane with the $C^{(1)}$ and the two N atoms (to avoid steric hindrance in this planar conformation during the "slow" transition, the second phenyl ring has to "make way" by adopting a conformation perpendicular to the plane). These transitions are observed in the classical simulations of *cis* AB, whereas the timescale of QM simulations of a few ps are too short to sample these transitions systematically. Fig. 16 shows one example of such a process by monitoring the dynamics of the CCNN dihedral angles of one AB unit in a simulation of liquid *cis* AB at 400 K, where both types of transitions are observed. In addition, snapshots of representative *cis* AB conformations

are shown to illustrate the conformational changes during the transitions. In order to get a rough estimate for the timescale of these ring flips the transitions of both types are counted for all CCNN dihedrals in a simulation of 343 *cis* AB molecules at 400 K. Since the separation between the states, in particular between the states involved in the "fast" transitions, is ambiguous, narrow regions ($\pm 14^\circ$) around the maxima of the distributions at $\pm 54^\circ$ and $\pm 126^\circ$ were defined (as marked in Fig. 16) and only transitions between these regions were counted. This results in transition times of approximately 20 ps for the "fast" and 200 ps for the "slow" ring flips. By Boltzmann inverting the dihedral distribution in Fig. 15d, one obtains an effective barrier for the "fast" transition of the order of about 3 kJ/mol ($\approx 1\ k_{\mathrm{B}}T$, where k_{B} is the Boltzmann constant) and for the "slow" transition a barrier of about 12 kJ/mol ($\approx 4\ k_{\mathrm{B}}T$), which approximately reproduces the relative magnitude of the two transition rates extracted from the dynamics.

3.2.2 QM/MM Photoisomerisation Simulations

Cis→trans The photoisomerisation of AB in the bulk liquid has been studied using the nonadiabatic QM/MM simulation method introduced in Sec. 2.5. Condensed phase effects are investigated by comparing results for the liquid with those for the gas phase (Sec. 3.1). Fig. 12b shows the time evolution of $\psi^{NN'}$ (i.e. the CNNC dihedral) for typical $cis \rightarrow trans$ trajectories in the liquid together with the corresponding time evolution of ψ^{N} and $\psi^{N'}$. Similar to the gas phase (Fig. 12a) $\psi^{NN'}$ changes rapidly, in about 30 fs, to a value of $\approx 90^\circ$ upon photoexcitation at $t = 0$; after the $S_1 \rightarrow S_0$ transition to the ground state, $\psi^{NN'}$ reaches a value of $\approx 180^\circ$, thus indicating a successful $cis \rightarrow trans$ isomerisation in both cases. Inspection of the order parameters ψ^{N} and $\psi^{N'}$ (cf. Fig. 12) reveals that the total change in $\psi^{NN'}$ is due to equal contributions from the two coordination planes at N and N′ in opposite directions. Note that the analysis presented in Fig. 12 shows no significant differences between the liquid and the gas phase. The liquid environment obviously does not impose any major constraints on the dynamics of the CNNC moiety. As discussed above, this is due to the fact that photoisomerisation proceeds through a pedal motion of the CNNC group which does *not* involve large amplitude rotation of the phenyl rings, as illustrated by the inset of Fig. 12b) which shows the rmsd's of the N atoms and the phenyl ring centres R. Again, during the first 30 fs it is mainly the translocation of the N atoms that is responsible for the change in $\psi^{NN'}$ by $\approx 90^\circ$, while the phenyl rings remain largely fixed in space.

While the "hula-twist" motion of the CNNC moiety during $cis \rightarrow trans$ photoisomerisation is practically unaffected by the bulk environment, we have observed a pronounced hindrance of the rotation of the phenyl rings about the C-N bonds, measured by the rotation angles ψ^{RN} and $\psi^{R'N'}$[64]. In the liquid, relaxation of the molecular structure to the ground state equilibrium following a $cis \rightarrow trans$ switch of the CNNC group is seen to be much slower than in the gas phase.

Trans→cis Analogous to the above discussion, we now analyze the $trans \rightarrow cis$ photoisomerisation mechanism in the liquid as obtained from nonadiabatic QM/MM simulations[64]. The time evolution of $\psi^{NN'}$ in the $trans \rightarrow cis$ case is displayed in the bottom panel of Fig. 14b. We observe a much slower decrease in $\psi^{NN'}$ compared to the gas phase (Fig. 14a). For the trajectory shown, a value of $\approx 120^\circ$ is reached after 450 fs, still far

from the 90° value where the $S_1 \rightarrow S_0$ hop took place in the gas phase. In fact none of the trajectories in the liquid reached 90°, and the average time to reach 120° is 672 fs, compared to 259 fs in the gas phase (see Fig. 13). In contrast to the gas phase, there is no sustained sign-change in either ψ^N or $\psi^{N'}$, explaining the fact that changes in $\psi^{NN'}$ are smaller in the liquid.

It may seem surprising at first that $trans \rightarrow cis$ photoisomerisation is strongly hindered in the bulk, while $cis \rightarrow trans$ is essentially unaffected. This can be easily rationalised, however, in terms of the S_1 potential landscape (see Sec. 3.1.2, Fig. 11). Vertical excitation of the *cis*-AB isomer promotes the system to a steep Franck-Condon region in the S_1, providing a large driving force for isomerisation. The Franck-Condon region for *trans*-AB, on the other hand, is comparatively flat and the resulting forces small, rendering $trans \rightarrow cis$ photoisomerisation much more vulnerable to any environmental disturbances. A recent nonadiabatic simulation study of the photoisomerisation of AB in various organic solvents using a force field derived *ab initio* by Tiberio et al.[150] has confirmed the pedal mechanism and the strong impact of the solvent on decay times and quantum yields.

3.3 The 8AB8 Liquid Crystal

3.3.1 Atomistic Classical Simulations

As described in Sec. 3.1.1, we extended and partly reparameterized an existing classical atomistic forcefield to be able to simulate the liquid crystalline compound 8AB8 which consists of a central azobenzene unit and two octodecane chains connected to the oxygens. We studied the phase behaviour of this compound by classical simulations using replica exchange techniques[46]. Even though it is possible to use these techniques to equilibrate preset liquid crystalline structures and to draw some conclusions about the stability of certain phases in the atomistic model, the time and length scales required to properly cover liquid crystalline phase transition processes and to extensively investigate the phase behaviour of 8AB8 can only be reached with a coarse grained simulation approach. Atomistic simulation then again becomes important after a backmapping procedure, to "hand over" well equilibrated structures of the liquid crystalline system obtained from CG simulations to the further classical and QM/MM investigation of the photoisomerization in dense, ordered systems.

3.3.2 Coarse Grained Classical Simulations

In Ref. 44 a coarse graining scheme originally developed for amorphous polymers was applied to liquid crystalline 8AB8. It is known that the behaviour of the polymer melts strongly depends on chain connectivity and excluded volume interactions of the polymeric beads. Consequently, it is often not essential for a correct prediction of the melt structure and dynamics at the mesoscale to introduce attractive (nonbonded/intermolecular) interactions. For liquid crystalline 8AB8 we find that specific (attractive) nonbonded interactions between the different units are required to obtain a CG model that is capable of reproducing the correct liquid crystal behaviour and that is closely linked to the atomistic level. This makes switching between the levels of resolution possible. Fig. 17 shows how the atomistic structure of 8AB8 was mapped onto the coarse grained beads. Intramolecular (bonded) CG

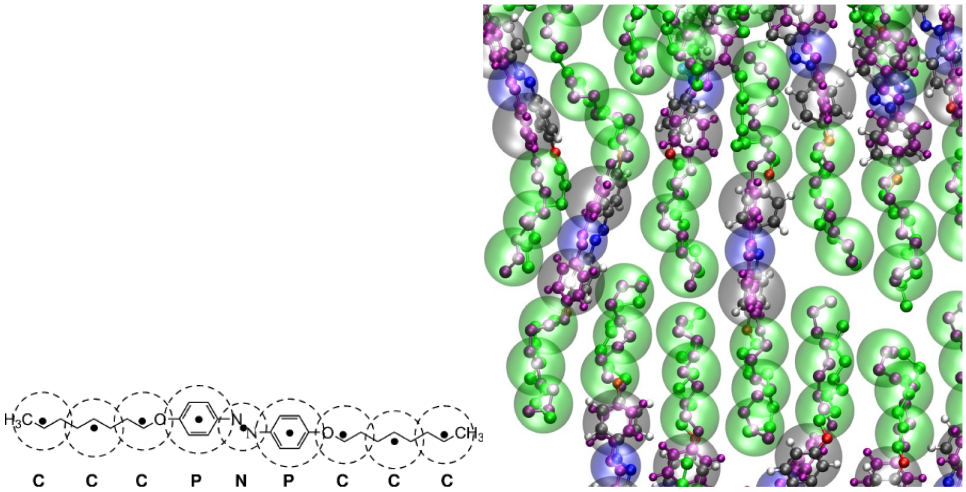


Figure 17. Left: Chemical structure of (trans) 8AB8 and MM↔CG mapping scheme (CG beads are denoted C = alkyl; P = phenyl; N = azo). Right: Snapshot of 8AB8 molecules in a backmapped liquid-crystalline structure. Large spheres: CG beads (green: C; gray: P; blue: N); Purple small spheres: re-inserted atomistic coordinates after equilibration with restraining to CG structure; Green, gray, blue (+white and red) small spheres: re-inserted atomistic coordinates after 5 ps free MD simulation without restraining potential.

potentials were obtained from simulations of an all-atom single 8AB8 molecule while intermolecular potentials were developed based on all-atom simulations of isotropic liquids of fragments of the 8AB8 molecule. Liquid benzene, liquid azobenzene (in its *trans* and in its *cis* form), liquid octadecane and various mixtures of these compounds were used for the intermolecular part. Based on the structure of these liquids (radial distribution functions), nonbonded interaction potentials were determined, both using analytical potential functions and the iterative Boltzmann inversion method.

An overview of all analytical and tabulated interaction functions obtained with this procedure can be found in the Supplementary Material of Ref. 44. As an illustration of this structure-based coarse graining method, Fig. 18 shows the radial distribution functions characteristic of liquid *trans* and *cis* azobenzene. The figure shows the corresponding structure functions from atomistic simulations (mapped onto CG degrees of freedom) together with the corresponding CG simulations with (numerical) interaction functions obtained from iterative Boltzmann inversion, specifically for each isomer. The figure also shows the result of CG simulations with an averaged interaction function which serves as a compromise to be able to use a single set of CG potentials both for *trans* and *cis* azobenzene. The resulting interaction functions obtained for isotropic liquids were then put together to simulate liquid (*trans*) 8AB8 to study the liquid crystalline phase behaviour. We found that the use of (soft) analytical potentials which are purely repulsive (in the spirit of the previous coarse graining examples of polymeric systems) did not yield the correct mesophase behaviour of 8AB8, in fact no long-range ordering was observed for the model chosen, even with a rather wide scan of temperatures and pressures. With potentials generated with the iterative Boltzmann inversion method, i.e. numerical (tabulated) potentials which are also partly attractive, it is however possible to observe liquid crystalline structures. Thus, for

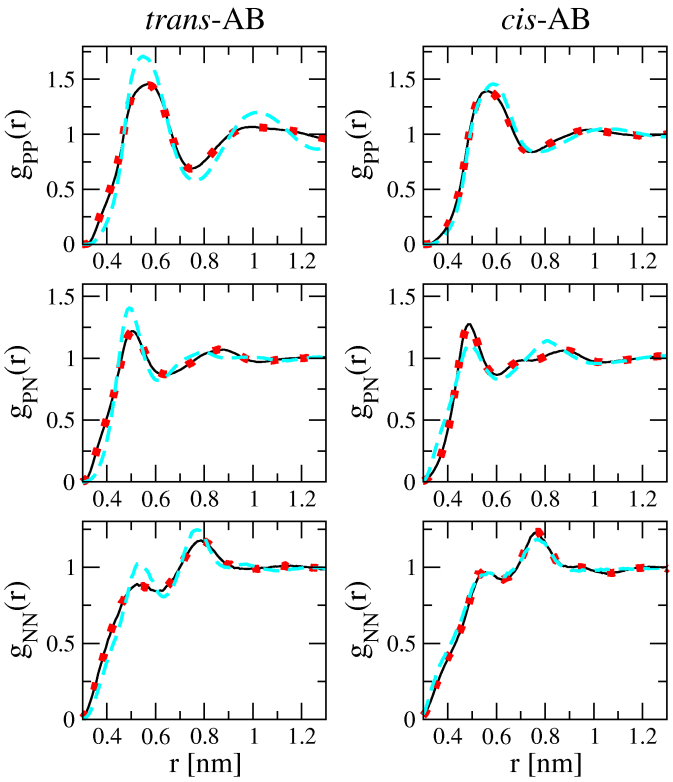


Figure 18. Radial distribution functions, $g(r)$. for liquid *trans*-AB (left panels) and *cis*-AB (right panels). Top panels: phenylphenyl $g_{\mathrm{PP}}(r)$. Middle panels: phenylazo $g_{\mathrm{PN}}(r)$. Lower panels: azoazo $g_{\mathrm{NN}}(r)$. Black solid lines: g(r) from atomistic simulation. Red, fat dotted lines: CG simulation with potentials through iterative Boltzmann inversion specifically for the respective trans and cis compounds. Cyan dashed lines: CG simulations with average potential as compromise for *trans* / *cis* AB.

the given molecule, i.e. the given size and shape of the mesogen and the given molecular flexibility of the alkoxy tails, it seems to be important to account for attractions between the different beads in the CG model in order to reproduce liquid crystalline phases of 8AB8. By varying simulation temperature and density it was possible to distinguish structures with smectic layers and more disordered structures, which are however not truly nematic but still exhibit a varying degree of positional order with partly interdigitated smectic layers, as illustrated in Fig. 20 which shows the z-positional order in the liquid crystal systems. Several snapshots of the corresponding 8AB8 liquid crystal simulations at different densities and system sizes can be seen in Fig. 19. At this point it should be noted that with this approach of building a CG model on an atomistic forcefield description of the molecule, possible weaknesses of the atomistic model will be automatically transferred to the CG model. With the given approach the mesoscale simulations maintain an important link to the chemical structure, and through the inverse mapping procedure it is possible to obtain atomistic coordinates of the system as illustrated in Fig. 17. This is important for passing down the CG configuration and velocities to the MM level below, which, in turn, serves as

a starting point for a na-QM/MM simulation during which one of the 8AB8 molecules is photoswitched from *trans* to *cis*.

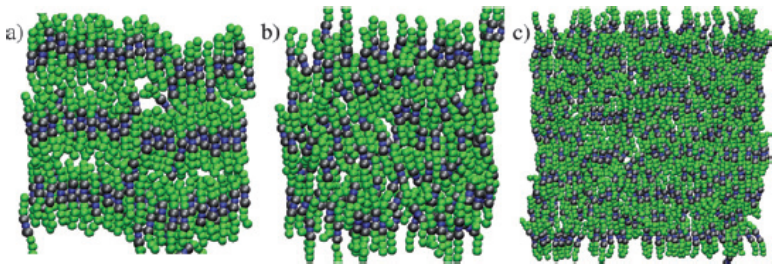


Figure 19. Snapshots of slices through the liquid crystal system from simulations at T = 0.8 with the average-8AB8 FF. a) 1323 8AB8 molecules, density = 1.44 molecules per nm^3; b) 1323 8AB8 molecules, density = 1.62 molecules per nm^3; c) 6125 8AB8, density = 1.52 molecules per nm^3.

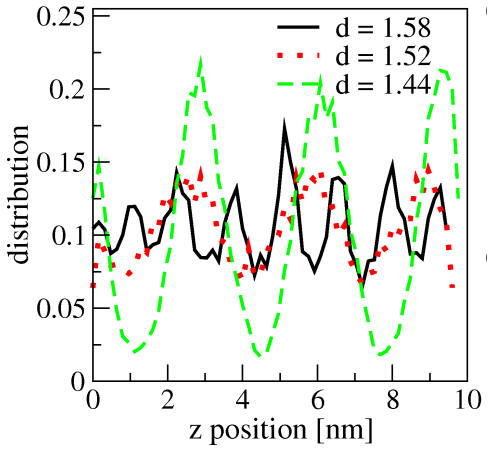


Figure 20. Distribution of centres of 8AB8 molecules (N beads) along z-direction of the simulation box of the liquid crystal system at T = 0.8 and various densities (see legend, densities are in molecules per nm^3).

4 Summary and Outlook

A multiscale simulation approach has been presented which links nonadiabatic *ab initio* molecular dynamics, atomistic classical molecular dynamics, and coarse grained classical molecular dynamics, to make possible the simulation of mesoscopic processes triggered by quantum-mechanical events highly local in space and time. As a test bed for our combined approach we chose the liquid crystalline system 8AB8, consisting of an azobenzene chromophore embedded in a hydrocarbon chain. By switching a fraction of AB units from

trans-AB to *cis*-AB using light of a suitable wavelength, transitions between ordered and disordered phases can be induced in the 8AB8 liquid crystal.

First, an atomistic force field for AB and AB chain derivatives was derived from AIMD data in the gas phase, and was then shown to be suitable for bulk liquid AB and liquid crystalline 8AB8 as well. The atomistic force field was the prerequisite to be able to perform nonadiabatic QM/MM simulations of the photoisomerisation of AB in the bulk, and it formed the basis for the development of a coarse grained representation of 8AB8, which was required to study the phase behaviour of the liquid crystal. We have indeed demonstrated that the CG force field is capable of reproducing several ordered phases of the 8AB8 liquid crystal and phase transitions between them.

We have gained unprecedented insights into the photoisomerisation mechanism of AB from na-QM simulations. A clear relationship between the molecular structure of a particular AB-based photoswitch and its relevant properties, such as photoisomerisation efficiency, has been established. This could pave the way for the rational design of improved photoswitches for light-controllable nano-devices and materials. Nonadiabatic QM/MM simulations of the photoisomerisation in liquid AB have revealed that the $cis \rightarrow trans$ reaction is large unaffected by the environment, while $trans \rightarrow cis$ is strongly hindered in the bulk. The respective behaviours of *cis*-AB and *trans*-AB could be traced back to their photoisomerisation mechanisms and rationalised in terms of the potential energy landscape.

All the necessary parts to perform a na-QM/MM/CG multiscale simulation are now available and can be applied to study photoisomerisation in 8AB8 and its effect on the liquid crystalline order. The most straightforward way of doing this is in a sequential fashion. Having built the multiscale model from the bottom up, i.e. based on first principles, the best way to tackle the liquid crystal is to apply the individual tools at the different scales in a cyclic manner, starting from the top. This means that first an ordered liquid crystalline phase needs to be produced in a CG simulation. The CG system then has to be back-mapped onto the atomistic representation using the procedure presented here and an MM simulation needs to be performed to reequilibrate the system. To simulate a photoisomerisation event, an na-QM/MM simulation has to be carried out subsequently. From this bottom layer, the information is then passed back to the MM and eventually to the CG level to complete the cycle.

Since, to induce a phase transition in 8AB8, a significant fraction of molecules need to be switched, it is desirable to develop a purely analytical switching potential based on the knowledge gained from na-QM/MM simulations about the mechanism. Such a switching potential is currently being designed and tested.

At present, the multiscale simulations involve different codes for the different layers. A great challenge consists in implementing a simultaneous na-QM/MM/CG simulation method, which would facilitate applications to very extended systems, albeit at the expense of a limited timescale. This means that all the methods outlined in these notes can be potentially combined in a single software package. Therefore, parts of the multiscale strategy outlined here, such as the adaptive resolution scheme, are currently being extended to be able to treat – with a single program – several scales, including the quantum level, concurrently in a robust way.

Other desirable extensions are, for instance, an adaptive QM/MM partitioning, beyond the path integral representation of atomic nuclei (without treating electrons explicitly), allowing for particle exchange between the QM and the MM regions, multiple QM regions,

and a theoretically sound method of switching "on" and "off" a QM region. These are issues that need to be solved not only in the context of the project introduced here, but for hydrid simulation schemes in general.

The multiscale strategy and methodology developed here presents a powerful tool, first and foremost in the ever growing field of light-addressable azo-materials, but at the same time it is transferable to a plethora of other applications, including (photo)biochemistry and -biophysics.

Acknowledgments

I am indebted to D. Marx, K. Kremer, L. Delle Site, C. Peter, and M. Böckmann who have all contributed to this project. We are grateful to the Volkswagen Stiftung for supporting our project "Adaptive Multiscale Simulation: Connecting the Quantum to the Mesoscopic Level" within the framework of the program "New Conceptual Approaches to Modeling and Simulation of Complex Systems – Computer Simulation of Molecular and Cellular Biosystems as well as Complex Soft Matter". The simulations were performed using resources from NIC Jülich, BOVILAB@RUB and Rechnerverbund–NRW.

References

1. R. Car and M. Parrinello, Phys. Rev. Lett., **55**, 2471, 1985.
2. D. Marx and J. Hutter, "Ab initio molecular dynamics: Theory and implementation", in: Modern Methods and Algorithms of Quantum Chemistry, J. Grotendorst, (Ed.). NIC, Jülich, 2000, `www.theochem.rub.de/go/cprev.html`, (accessed August 2010).
3. D. Marx and J. Hutter, *Ab Initio Molecular Dynamics: Basic Theory and Advanced Methods*, Cambridge University Press, Cambridge, 2009.
4. J. Aqvis and A. Warshel, Chem. Rev., **93**, 2523, 1993.
5. J. L. Gao and D. G. Truhlar, Annu. Rev. Phys. Chem., **53**, 467, 2002.
6. P. Carloni, U. Rothlisberger, and M. Parrinello, Acc. Chem. Res., **35**, 455, 2002.
7. P. Sherwood, A. H. de Vries, M. F. Guest, G. Schreckenbach, C. R. A. Catlow, S. A. French, A. A. Sokol, S. T. Bromley, W. Thiel, A. J. Turner, S. Billeter, F. Terstegen, S. Thiel, J. Kendrick, S. C. Rogers, J. Casci, M. Watson, F. King, E. Karlsen, M. Sjovoll, A. Fahmi, A. Schäfer, and C. Lennartz, J. Mol. Stuct. THEOCHEM, **632**, 1, 2003.
8. G. Tresadern, P. F. Faulder, M. P. Gleeson, Z. Tai, G. MacKenzie, N. A. Burton, and I. H. Hillier, Theor. Chem. Acc., **109**, 108, 2003.
9. H. M. Senn and W. Thiel, *QM/MM Methods for Biomolecular Systems*, Angew. Chem. Int. Ed., **48**, 1198–1229, 2009.
10. G. A. Voth, (Ed.), *Coarse-Graining of Condensed Phase and Biomolecular Systems*, Chapman and Hall/CRC Press, Taylor and Francis Group, 2009.
11. M. L. Klein and W. Shinoda, *Large-scale molecular dynamics simulations of self-assembling systems*, Science, **321**, 798–800, 2008.
12. A. P. Lyubartsev and A. Laaksonen, *Calculation of effective interaction potentials from radial-distribution functions - a reverse Monte-Carlo approach*, Phys. Rev. E, **52**, no. 4, 3730 – 3737, 1995.

13. F. Müller-Plathe, *Coarse-graining in polymer simulation: From the atomistic to the mesoscopic scale and back*, ChemPhysChem, **3**, no. 9, 754 – 769, 2002.
14. S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. de Vries, *The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations*, J. Phys. Chem. B, **111**, no. 27, 7812–7824, 2007.
15. M. E. Johnson, T. Head-Gordon, and A. A. Louis, *Representability problems for coarse-grained water potentials*, J. Chem. Phys., **126**, no. 14, 144509, 2007.
16. L. Monticelli, S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman, and S. J. Marrink, *The MARTINI Coarse-Grained Force Field: Extension to Proteins*, J. Chem. Theor. Comput., **4**, no. 5, 819–834, 2008.
17. W.G. Noid, J.W. Chu, G.S. Ayton, and G.A. Voth, *Multiscale coarse-graining and structural correlations: connections to liquid state theory*, J. Phys. Chem. B, **111**, 4116–4127, 2007.
18. W.G. Mullinax, J.W.; Noid, *Generalized Yvon-Born-Green Theory for Molecular Systems*, Phys. Rev. Lett., **103**, 198104, 2009.
19. H. Wang, C. Junghans, and K. Kremer, *Comparative atomistic and coarse-grained study of water: what do we lose by coarse-graining?*, Eur. Phys. J. E, **28**, 221–229, 2009.
20. Alexander Lyubartsev, Alexander Mirzoev, LiJun Chen, and Aatto Laaksonen, *Systematic coarse-graining of molecular models by the Newton inversion method*, Faraday Discuss., **144**, 43–56, 2010.
21. J. Baschnagel, K. Binder, P. Doruker, A. A. Gusev, O. Hahn, K. Kremer, W. L. Mattice, F. Muller-Plathe, M. Murat, W. Paul, S. Santos, U. W. Suter, and V. Tries, *Bridging the gap between atomistic and coarse-grained models of polymers: Status and perspectives*, Adv. Polym. Sci., **152**, 41 – 156, 2000.
22. W. Tschöp, K. Kremer, J. Batoulis, T. Burger, and O. Hahn, *Simulation of polymer melts. I. Coarse-graining procedure for polycarbonates*, Acta Polym., **49**, no. 2-3, 61 – 74, 1998.
23. C. F. Abrams and K. Kremer, *Combined coarse-grained and atomistic simulation of liquid bisphenol A-polycarbonate: Liquid packing and intramolecular structure*, Macromolecules, **36**, no. 1, 260 – 267, 2003.
24. V. A. Harmandaris, N. P. Adhikari, N. F. A. van der Vegt, and K. Kremer, *Hierarchical modeling of polystyrene: From atomistic to coarse-grained simulations*, Macromolecules, **39**, no. 19, 6708 – 6719, 2006.
25. Q. Sun and R. Faller, *Systematic coarse-graining of a polymer blend: Polyisoprene and polystyrene*, J. Chem. Theory Comput., **2**, no. 3, 607 – 615, 2006.
26. Theodora Spyriouni, Christos Tzoumanekas, Doros Theodorou, Florian Mueller-Plathe, and Giuseppe Milano, *Coarse-grained and reverse-mapped united-atom simulations of long-chain atactic polystyrene melts: Structure, thermodynamic properties, chain conformation, and entanglements*, Macromolecules, **40**, no. 10, 3876–3885, 2007.
27. V. A. Harmandaris, D. Reith, N. F. A. van der Vegt, and K. Kremer, *Comparison between Coarse-Graining Models for Polymer Systems: Two Mapping Schemes for Polystyrene*, Macromol. Chem. Phys., **208**, 2109 – 2120, 2007.
28. V. A. Harmandaris and K. Kremer, *Dynamics of polystyrene melts through hierarchical multiscale simulations*, Macromolecules, **42**, 791–802, 2009.

29. D. Fritz, V. A. Harmandaris, K. Kremer, and N. F. A. van der Vegt, Macromolecules, **42**, 7579, 2009.
30. M. Muller, K. Katsov, and M. Schick, *Biological and synthetic membranes: What can be learned from a coarse-grained description?*, Physics Reports, **434**, no. 5-6, 113 – 176, 2006.
31. B. J. Reynwar, G. Illya, V. A. Harmandaris, M. M. Müller, K. Kremer, and M. Deserno, *Aggregation and vesiculation of membrane proteins by curvature-mediated interactions*, Nature, **447**, no. 7143, 461 – 464, 2007.
32. M. Deserno, *Mesoscopic Membrane Physics: Concepts, Simulations, and Selected Applications*, Macromol. Rapid Comm., **30**, no. 9-10, 752–771, 2009.
33. H. D. Nguyen and C. K. Hall, *Molecular dynamics simulations of spontaneous fibril formation by random-coil peptides*, Proc. Natl. Acad. Sci. USA, **101**, no. 46, 16180 – 16185, 2004.
34. G. Bellesia and J. E. Shea, *Self-assembly of beta-sheet forming peptides into chiral fibrillar aggregates*, J. Chem. Phys., **126**, no. 24, 245104, 2007.
35. I. F. Thorpe, J. Zhou, and G. A. Voth, *Peptide folding using multiscale coarse-grained models*, J. Phys. Chem. B, **112**, 13079–13090, 2008.
36. A. Villa, C. Peter, and N. F. A. van der Vegt, *Self-assembling dipeptides: conformational sampling in solvent-free coarse-grained simulation*, Phys. Chem. Chem. Phys., **11**, 2077, 2009.
37. A. Villa, N. F. A. van der Vegt, and C. Peter, *Self-assembling dipeptides: including solvent degrees of freedom in a coarse-grained model*, Phys. Chem. Chem. Phys., **11**, 2068, 2009.
38. T. Head-Gordon and S. Brown, *Minimalist models for protein folding and design*, Curr. Opin. Struct. Biol., **13**, no. 2, 160 – 167, 2003.
39. C. Clementi, *Coarse-grained models of protein folding: toy models or predictive tools?*, Curr. Op. Struct. Biol., **18**, no. 1, 10–15, 2008.
40. Valentina Tozzini, *Multiscale Modeling of Proteins*, Acc. Chem. Res., **43**, no. 2, 220–230, 2010.
41. C. Peter and K. Kremer, Faraday Discuss., **144**, 9, 2010.
42. B. Hess, S. Leon, N. van der Vegt, and K. Kremer, *Long time atomistic polymer trajectories from coarse grained simulations: bisphenol-A polycarbonate*, Soft Matter, **2**, no. 5, 409 – 414, 2006.
43. G. Santangelo, A. Di Matteo, F. Muller-Plathe, and G. Milano, *From mesoscale back to atomistic models: A fast reverse-mapping procedure for vinyl polymer chains*, J. Phys. Chem. B, **111**, no. 11, 2765 – 2773, 2007.
44. C. Peter, L. Delle Site, and K. Kremer, *Classical simulations from the atomistic to the mesoscale: coarse graining an azobenzene liquid crystal*, Soft Matter, **4**, 859–869, 2008.
45. X. Chen, P. Carbone, G. Santangelo, A. Di Matteo, G. Milano, and F. Muller-Plathe, *Backmapping coarse-grained polymer models under sheared nonequilibrium conditions.*, Phys. Chem. Chem. Phys., **11**, no. 12, 1977 – 88, 2009.
46. M. Böckmann, C. Peter, L. Delle Site, N. L. Doltsinis, K. Kremer, and D. Marx, J. Chem. Theory Comput., **3**, 1789, 2007.
47. G. S. Kumar and D. C. Neckers, *PHOTOCHEMISTRY OF AZOBENZENE-CONTAINING POLYMERS*, Chem. Rev., **89**, no. 8, 1915 – 1925, 1989.

48. T. Ikeda and O. Tsutsumi, Science, **268**, 1873, 1995.
49. B. L. Feringa, (Ed.), *Molecular Switches*, Wiley-VCH, Weinheim, 2001.
50. Z Sekkat and W. Knoll, (Eds.), *Photoreactive organic thin films*, Academic Press, San Diego, 2002.
51. A. Natansohn and P. Rochon, *Photoinduced motions in azo-containing polymers*, Chem. Rev., **102**, 4139–4175, 2002.
52. A. Stolow, Ann. Rev. Phys. Chem., **54**, 89, 2003.
53. S. Spörlein, H. Carstens, H. Satzger, C. Renner, R. Behrendt, L. Moroder, P. Tavan, W. Zinth, and J. Wachtveitl, Proc. Nat. Acad. Sci., **99**, 7998, 2002.
54. Y. Yu, M. Nakano, and T. Ikeda, Nature, **425**, 145, 2003.
55. I. Banerjee, L. Yu, and H. Matsui, J. Am. Chem. Soc., **125**, 9542, 2003.
56. W. R. Browne and B. L. Feringa, Nature Nanotechnology, **1**, 25, 2006.
57. T. Hugel, N. B. Holland, A. Cattani, L. Moroder, M. Seitz, and H. E. Gaub, Science, **296**, 1103, 2002.
58. A. Khan, C. Kaiser, and S. Hecht, Angew. Chem. Int. Ed., **45**, 1878, 2006.
59. M. Böckmann, D. Marx, C. Peter, L. Delle Site, K. Kremer, and N. L. Doltsinis, Phys. Chem. Chem. Phys., **13**, 7604–7621, 2011.
60. N. L. Doltsinis and D. Marx, Phys. Rev. Lett., **88**, 166402, 2002.
61. N. L. Doltsinis and D. Marx, J. Theor. Comp. Chem., **1**, 319–349, 2002.
62. A. Laio, J. VandeVondele, and U. Rothlisberger, J. Chem. Phys., **116**, 6941, 2002.
63. M. Böckmann, N. L. Doltsinis, and D. Marx, Phys. Rev. E, **78**, 036101, 2008.
64. M. Böckmann, N. L. Doltsinis, and D. Marx, J. Phys. Chem. A, **114**, 745, 2010.
65. H. Nieber, A. Hellweg, and N. L. Doltsinis, J. Am. Chem. Soc., **132**, 1778–1779, 2010.
66. T. Kerdcharoen, K. R. Liedl, and B. M. Rode, Chem. Phys., **211**, 313, 1996.
67. T. S. Hofer, A. B. Pribil, B. R. Randolf, and B. M. Rode, J. Am. Chem. Soc., **127**, 14231, 2005.
68. T. Kerdcharoen and K. Morokuma, Chem. Phys. Lett., **355**, 257, 2002.
69. R. Bulo, B. Ensing, J. Sikkema, and L. Visscher, J. Chem. Theory Comput., **5**, 2212, 2009.
70. A. Heyden and D. G. Truhlar, J. Chem. Theory Comput., **4**, 217, 2008.
71. G. Csanyi, T. Albaret, M. C. Payne, and A. De Vita, Phys. Rev. Lett., **93**, 175503, 2004.
72. A. B. Poma and L. Delle Site, Phys. Rev. Lett., **104**, 250201, 2010.
73. M. Praprotnik, L. Delle Site, and K. Kremer, *Adaptive resolution molecular-dynamics simulation: Changing the degrees of freedom on the fly*, J. Chem. Phys., **123**, 224106, 2005.
74. M. Praprotnik, L. Delle Site, and K. Kremer, *Adaptive resolution scheme for efficient hybrid atomistic-mesoscale molecular dynamics simulations of dense liquids*, Phys. Rev. E, **73**, 066701, 2006.
75. M. Praprotnik, K. Kremer, and L. Delle Site, *Adaptive molecular resolution via a continuous change of the phase space dimensionality*, Phys. Rev. E, **75**, 017701, 2007.
76. M. Praprotnik, K. Kremer, and L. Delle Site, J. Phys. A: Math. Gen., **40**, 017701, 2007.
77. E. E. Nikitin, in: Chemische Elementarprozesse, H. Hartmann, (Ed.). Springer, Berlin, 1968.

78. E. E. Nikitin and L. Zülicke, *Theory of Chemical Elementary Processes*, Springer, Berlin, 1978.
79. L. Salem, *Electrons in Chemical Reactions: First Principles*, Wiley, New York, 1982.
80. L. Salem, C. Leforestier, G. Segal, and R. Wetmore, J. Am. Chem. Soc., **97**, 479, 1975.
81. J. C. Tully, in: Classical and Quantum Dynamics in Condensed Phase Simulations, B. J. Berne, G. Ciccotti, and D. F. Coker, (Eds.). World Scientific, Singapore, 1998.
82. J. C. Tully, in: Modern Methods for Multidimensional Dynamics Computations in Chemistry, D. L. Thompson, (Ed.). World Scientific, Singapore, 1998.
83. J. C. Tully, M. Gomez, and M. Head-Gordon, J. Vac. Sci. Technol., **A11**, 1914, 1993.
84. J. C. Tully and R. K. Preston, J. Chem. Phys., **55**, 562, 1971.
85. J. C. Tully, J. Chem. Phys., **93**, 1061, 1990.
86. U. Müller and G. Stock, J. Chem. Phys., **107**, 6230, 1997.
87. S. Hammes-Schiffer and J. C. Tully, J. Chem. Phys., **101**, 4657, 1994.
88. D. F. Coker and L. Xiao, J. Chem. Phys., **102**, 496, 1995.
89. F. Webster, P. J. Rossky, and R. A. Friesner, Comp. Phys. Comm., **63**, 494, 1991.
90. F. J. Webster, J. Schnitker, M. S. Friedrichs, R. A. Friesner, and P. J. Rossky, Phys. Rev. Lett., **66**, 3172, 1991.
91. E. R. Bittner and P. J. Rossky, J. Chem. Phys., **103**, 8130, 1995.
92. E. R. Bittner and P. J. Rossky, J. Chem. Phys., **107**, 8611, 1997.
93. M. D. Hack and D. G. Truhlar, J. Chem. Phys., **114**, 2894, 2001.
94. M. D. Hack and D. G. Truhlar, J. Chem. Phys., **114**, 9305, 2001.
95. Y. L. Volobuev, M. D. Hack, M. S. Topaler, and D. G. Truhlar, J. Chem. Phys., **112**, 9716, 2000.
96. R. G. Parr and W. Yang, *Density Functional Theory of Atoms and Molecules*, Oxford University Press, Oxford, 1989.
97. R. M. Dreizler and E. K. U. Gross, *Density–Functional Theory*, Springer, Berlin, 1990.
98. N. L. Doltsinis, "Nonadiabatic dynamics: mean-field and surface hopping", in: Quantum Simulations of Complex Many-Body Systems: From Theory to Algorithms, J. Grotendorst, D. Marx, and A. Muramatsu, (Eds.). NIC, FZ Jülich, 2002, `www.fz-juelich.de/nic-series/volume10/doltsinis.pdf`.
99. N. L. Doltsinis, "Molecular dynamics beyond the born-oppenheimer approximation: mixed quantum–classical approaches", in: Computational Nanoscience: Do it Yourself!, J. Grotendorst, S. Blügel, and D. Marx, (Eds.). NIC, FZ Jülich, 2006, `www.fz-juelich.de/nic-series/volume31/doltsinis1.pdf`.
100. I. Frank, J. Hutter, D. Marx, and M. Parrinello, J. Chem. Phys., **108**, 4060, 1998.
101. S. Grimm, C. Nonnenberg, and I. Frank, J. Chem. Phys., **119**, 11574, 2003.
102. S. Grimm, C. Nonnenberg, and I. Frank, J. Chem. Phys., **119**, 11585, 2003.
103. S. Grimm, C. Nonnenberg, and I. Frank, J. Chem. Phys., **119**, 11574, 2003, *ibid.* **119**, (2003) 11585.
104. N. L. Doltsinis, Mol. Phys., **102**, 499, 2004.
105. H. Langer and N. L. Doltsinis, Phys. Chem. Chem. Phys., **6**, 2742, 2004.
106. H. Langer, N. L. Doltsinis, and D. Marx, ChemPhysChem, **6**, 1734, 2005.
107. P. R. L. Markwick and N. L. Doltsinis, J. Chem. Phys., **126**, 175102, 2007.

108. N. L. Doltsinis, P. R. L. Markwick, H. Nieber, and H. Langer, in: Radiation Induced Molecular Phenomena in Nucleic Acid, M. K. Shukla and J. Leszczynski, (Eds.). Springer, Netherlands, 2008.
109. H. Nieber and N. L. Doltsinis, Chem. Phys., **347**, 405, 2008.
110. S. R. Billeter and D. Egli, J. Chem. Phys., **125**, 224103, 2006.
111. W.F. van Gunsteren and H.J.C Berendsen, BIOMOS B.V., Zürich/Groningen, (1996).
112. R. Baron, D. Trzesniak, A. H. de Vries, A. Elsener, S. J. Marrink, and W. F. van Gunsteren, *Comparison of thermodynamic properties of coarse-grained and atomic-level simulation models*, Chemphyschem, **8**, no. 3, 452 – 461, 2007.
113. V. Ruehle, C. Junghans, A. Lukyanov, K. Kremer, and D. Andrienko, *Versatile Object-oriented Toolkit for Coarsegraining Applications (VOTCA)*, J. Chem. Theory Comput., **5**, 3211, 2009.
114. F. Ercolessi and J. B. Adams, *Interatomic Potentials from First-Principles Calculations: The Force-Matching Method*, Europhys. Lett., **26**, no. 8, 583, 1994.
115. S. Izvekov and G. A. Voth, *A multiscale coarse-graining method for biomolecular systems*, J. Phys. Chem. B, **109**, no. 7, 2469 – 2473, 2005.
116. G. S. Ayton, W. G. Noid, and G. A. Voth, *Multiscale modeling of biomolecular systems: in serial and in parallel*, Curr. Opin. Struct. Biol., **17**, no. 2, 192 – 198, 2007.
117. A. Villa, C. Peter, and N. F. A. van der Vegt, *Transferability of Nonbonded Interaction Potentials for Coarse-Grained Simulations: Benzene in Water*, J. Chem. Theory Comput., **6**, 2434–2444, 2010.
118. A. Chaimovich and M. S. Shell, *Anomalous waterlike behavior in spherically-symmetric water models optimized with the relative entropy*, Phys. Chem. Chem. Phys., **11**, 1901–1915, 2009.
119. J. R. Silbermann, S. H. L. Klapp, M. Schoen, N. Chennamsetty, H. Bock, and K. E. Gubbins, *Mesoscale modeling of complex binary fluid mixtures: Towards an atomistic foundation of effective potentials*, J. Chem. Phys., **124**, 074105, 2006.
120. D. Reith, M. Putz, and F. Muller-Plathe, *Deriving effective mesoscale potentials from atomistic simulations*, J. Comp. Chem., **24**, no. 13, 1624 – 1636, 2003.
121. M. Praprotnik, L. Delle Site, and K. Kremer, Annu.Rev.Phys.Chem., **59**, 545, 2008.
122. B. Ensing, S. O. Nielsen, P. B. Moore, M. L. Klein, and M. Parrinello, J. Chem. Theory Comput., **3**, 1100, 2007.
123. S. Izvekov and G. A. Voth, J. Chem. Theory Comput., **5**, 3232, 2009.
124. Simon Poblete, Matej Praprotnik, Kurt Kremer, and Luigi Delle Site, *Coupling different levels of resolution in molecular simulations*, J.Chem.Phys., **132**, no. 11, 114101, 2010.
125. L. Delle Site, *Some fundamental problems for an energy conserving adaptive resolution molecular dynamics scheme*, Phys. Rev. E, **76**, 047701, 2007.
126. M. Praprotnik, L. Delle Site, and K. Kremer, *A macromolecule in a solvent: Adaptive resolution molecular dynamics simulation*, J. Chem. Phys., **126**, 134902, 2007.
127. S. Matysiak, M. Praprotnik, L. Delle Site, K. Kremer, and C. Clementi, *Adaptive resolution simulation of liquid water*, J. Phys. Condens. Matter, **19**, 292201, 2007.
128. S. Matysiak, C. Clementi, M. Praprotnik, K. Kremer, and L. Delle Site, *Modeling diffusive dynamics in adaptive resolution simulation of liquid water*, J. Chem. Phys., **128**, 024503, 2008.

129. R. Delgado-Buscalioni, K. Kremer, and M. Praprotnik, J.Chem.Phys., **128**, 114110, 2008.
130. R. Delgado-Buscalioni, K. Kremer, and M. Praprotnik, J. Chem. Phys., **131**, 244107, 2009.
131. B. Lambeth, C. Junghans, K. Kremer, C. Clementi, and L Delle Site, Submitted, 2010.
132. C. I. Bayly, P. Cieplak, W. Cornell, and P. A. Kollman, J. Phys. Chem., **97**, 10269, 1993.
133. D. A. Case, D. A. Pearlman, J. W. Caldwell, T. E. Cheatham III, J. Wang, W. S. Ross, C. L. Simmerling, T. A. Darden, K. M. Merz, R. V. Stanton, A. L. Cheng, J. J. Vincent, M. Crowley, V. Tsui, H. Gohlke, R. J. Radmer, Y. Duan, J. Pitera, I. Massova, G. L. Seibel, U. C. Singh, P. K. Weiner, and P. A. Kollman, AMBER 7, University of California, San Francisco (2002).
134. A. Cembran, F. Bernardi, M. Garavelli, L. Gagliardi, and G. Orlandi, J. Am. Chem. Soc., **126**, 3234, 2004.
135. A. W. Adamson, A. Vogler, H. Kunkely, and R. Wachter, J. Am. Chem. Soc., **100**, 1300, 1978.
136. J. Harada, K. Ogawa, and S. Tomoda, *Molecular Motion and Conformational Interconversion of Azobenzenes in Crystals as Studied by X-ray Diffraction*, Acta Cryst. B, **53**, 662–672, Feb. 1997.
137. T. Pancur, F. Renth, F. Temps, B. Harbaum, A. Krüger, R. Herges, and C. Näther, *Femtosecond fluorescence up-conversion spectroscopy of a rotation-restricted azobenzene after excitation to the S_1 state*, Phys. Chem. Chem. Phys., **7**, 1985–1989, Apr. 2005.
138. C. M. Stuart, R. R. Frontiera, and R. A. Mathies, *Excited-State Structure and Dynamics of* cis- *and* trans-*Azobenzene from Resonance Raman Intensity Analysis*, J. Phys. Chem. A, **111**, 12072–12080, Nov. 2007.
139. C.-W. Chang, Y.-C. Lu, T.-T. Wang, and E. W.-G. Diau, *Photoisomerization Dynamics of Azobenzene in Solution with S_1 Excitation: A Femtosecond Fluorescence Anisotropy Study*, J. Amer. Chem. Soc., **126**, 10109–10118, July 2004.
140. H. Satzger, S. Spörlein, J. Wachtveitl, W. Zinth, and P. Gilch, Chem. Phys. Lett., **372**, 216, 2003.
141. R. Siewertsen, H. Neumann, B. Buchheim-Stehn, R. Herges, C. Näther, F. Renth, and F. Temps, J. Amer. Chem. Soc., **131**, 15594, 2009.
142. M. Böckmann, N. L. Doltsinis, and D. Marx, Angew. Chem. Int. Ed., **49**, 3382, 2010.
143. H. Flyvbjerg and H. G. Petersen, *Error estimates on averages of correlated data*, J. Chem. Phys., **91**, 461, 1989.
144. A. W. Jasper, S. Nangia, C. Zhu, and D. G. Truhlar, *Non-Born–Oppenheimer Molecular Dynamics*, Acc. Chem. Res., **39**, 101, 2006.
145. M. D. Hack, A. M. Wensmann, D. G. Truhlar, M. Ben-Nun, and T. J. Martínez, J. Chem. Phys., **115**, 1172, 2001.
146. A. Toniolo, C. Ciminelli, M. Persico, and T. Martinez, J. Chem. Phys., **123**, 234308, 2005.
147. R. Turanský, M. Konôpka, N. L. Doltsinis, I. Štich, and D. Marx, *Optical, Mechanical, and Opto–Mechanical Switching of Anchored Dithioazobenzene Bridges*, ChemPhysChem, **11**, 345, 2010.
148. R. Turanský, M. Konôpka, N. L. Doltsinis, I. Štich, and D. Marx, *Switching of functionalized azobenzene suspended between gold tips by mechanochemical, photochemical, and opto-mechanical means*, Phys. Chem. Chem. Phys., **12**, 13922, 2010.

149. O. Carstensen, J. Sielk, J. B. Schönborn, G. Granucci, and B. Hartke, J. Chem. Phys., **133**, 124305, 2010.
150. G. Tiberio, L. Muccioli, R. Berardi, and C. Zannoni, ChemPhysChem, **11**, 1018, 2010.
151. M. Barbatti, G. Granucchi, M. Persico, M. Ruckenbauer, M. Vazdar, M. Eckert-Maksic, and H Lischka, J. Photochem. Photobiol. A, **190**, 228–240, 2007.
152. E. Tapavicza, I. Tavernelli, and U. Rothlisberger, Phys. Rev. Lett., **98**, 023001, 2007.
153. M. Barbatti, Wiley Int. Rev., **1**, 620–633, 2011.

Transition Path Sampling of Phase Transitions – Nucleation and Growth in Materials Hard and Soft

Michael Grünwald[1], Swetlana Jungblut[2], and Christoph Dellago[2]

[1] Department of Chemistry
University of California at Berkeley, Berkeley, CA 94720, USA
E-mail: michael.gruenwald@berkeley.edu

[2] Faculty of Physics
University of Vienna, Boltzmanngasse 5, 1090 Vienna, Austria
E-mail: {swetlana.jungblut, christoph.dellago}@univie.ac.at

In this article, we give an introduction to transition path sampling, a computer simulation methodology developed to investigate rare but important events between known long-lived stable states. Such rare event processes play an important role in many areas of biology, chemistry, physics, and, in particular, materials science. Here, we focus on nucleation phenomena such as the freezing transition of a liquid or the structural transformation of a crystalline solid, in which the rare event is related to the formation of a critical nucleus of the thermodynamically favored phase embedded in the metastable phase. Due to the arising free energy barrier, typical nucleation times can exceed the basic time scale of particle motions by many orders of magnitude. Here, we will first lay out the general ideas of transition path sampling and explain how this technique circumvents the problem of widely disparate time scales. We will then discuss how transition path can be implemented and used to determine rate constants and reveal the transition mechanics. Finally, we will demonstrate the practical application of transition path sampling using the pressure induced structural transition of CdSe nanocrystals and the freezing of a supercooled soft particle fluid as examples.

1 Introduction

Many processes occurring in materials are characterized by widely disparate time scales occurring simultaneously. Consider, for instance, the diffusion of atoms or molecules adsorbed on a surface[1,2]. At sufficiently low temperatures, the atom typically resides at adsorption sites caused by the interactions of the adatom with the surface atoms. Due to thermal fluctuations, the adatom oscillates about the potential energy minimum on a time scale of picoseconds. Rarely, the adatom crosses the potential energy barrier separating the potential energy minima and jumps from one adsorption site to another. Note that this motion can occur through the jump of a single particle, but more complex mechanisms involving the motion of several atoms, for instance, the exchange with a sub-surface atom, are possible as well[1,3]. Since the jumps are thermally activated, these barrier crossing events occur rarely at low temperatures with typical time scales that can exceed those of basic atomic oscillations by many orders of magnitude. Nevertheless, jumps between adsorption sites are very important as they determine the rate at which adatoms diffuse on the surface. Another class of processes dominated by rare but important events are first order phase transitions such as the freezing of a supercooled liquid or the structural transition of a crystalline material under pressure. Away from the regime of spinodal decomposition, first order phase transitions occur through a nucleation and growth mechanism in which a nucleus of the stable phase forms in the metastable phase. This process involves the cross-

ing of a barrier related to the free energetic cost of creating an interface between the two phases. Once the system has crossed this barrier, i.e., the nucleus has reached a critical size, the transformation process proceeds by further growth of the nucleus of the stable phase, rapidly transforming the system in a barrier-less fashion. In this contribution, we will concentrate on nucleation and growth processes occurring at first order phase transitions, focusing on their computer simulation using transition path sampling, a methodology designed to circumvent computational difficulties associated with rare barrier crossing events.

A qualitative picture of nucleation processes occurring in metastable phases at first order phase transitions is provided by classical nucleation theory[4,5] (CNT). This theory asserts that transitions such as the freezing of a liquid or structural transformations in solids proceed via the formation of a localized nucleus of the stable phase in the metastable phase. Due to the free energetic cost associated with the creation of the interface between the two phases, the free energy increases as a function of nucleus size in the early nucleation stages, opposing rapid growth of the nucleus. Assuming that the nucleation occurs in the bulk of the metastable phase (a scenario commonly referred to as *homogeneous nucleation*) and the growing nucleus is spherical, the excess Gibbs free energy $\Delta G(r)$ of the system as a function of the nucleus radius r is given by

$$\Delta G(r) = 4\pi r^2 \gamma + \frac{4}{3}\rho_s \pi r^3 \Delta\mu, \tag{1}$$

where γ is the surface free energy per unit area, $\Delta\mu < 0$ is the difference in chemical potential between the two phases, and ρ_s is the number density of the stable phase. Note that if the nucleation occurs near impurities or at surfaces rather than in the bulk, i.e., in the case of heterogeneous nucleation, similar expressions for the free energy as a function of nucleus size can be derived[6]. While for small crystallite sizes the surface term dominates, for larger sizes the volume term prevails. This results in a barrier of height

$$\Delta G^* = \frac{16\pi\gamma^3}{3\rho_s^2 \Delta\mu^2} \tag{2}$$

which becomes very high compared to the thermal energy $k_{\mathrm{B}}T$ close to coexistence, where the difference in the chemical potential between the two phases approaches zero, $\Delta\mu \approx 0$. As a consequence, depending on the external conditions such as pressure or temperature, nucleation occurs very rarely on the time scale of basic molecular motions. Indeed, undercooled liquids can exist for almost arbitrary periods of time in the metastable state before they eventually crystallize[6]. Only if a rare thermal fluctuation drives the nucleus past the critical size, corresponding to the top of the free energy barrier, will the nucleus continue to grow rapidly transforming the entire system into the stable phase. Thus, the formation of the critical nucleus can be viewed as the decisive moment in the phase transition determining the rate at which the transformation occurs. While the concepts underlying classical nucleation theory are qualitatively reasonable, quantitative discrepancies arise when this theory is applied to specific models. For instance, the nucleation free energy of a freezing soft sphere fluid, shown in Fig. 1, displays a barrier as predicted by CNT, but its shape deviates from the CNT-form. In this case, the assumption of a spherical nucleus of the stable phase is not strictly valid, as indicated by the critical cluster shown in the right hand side panel of Fig. 1.

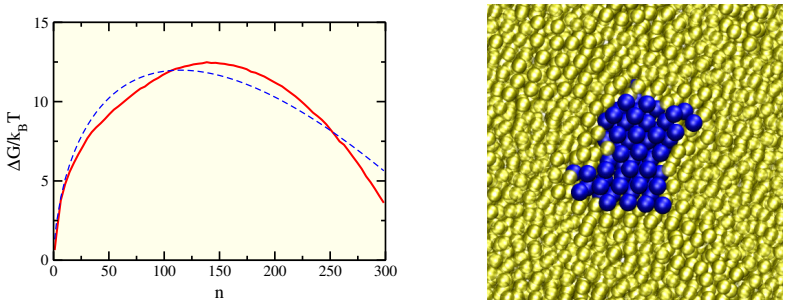


Figure 1. Free energy ΔG as a function of the number n of particles in the crystalline cluster (left) and cross section through a critical cluster (right) for a freezing soft sphere fluid. These results were obtained from a simulation of 6668 Lennard-Jones particles at a temperature about 30% under the freezing temperature. In the free energy plot on the left, the red line denotes simulation results and the dashed blue line is a fit of the classical nucleation theory formula for the free energy. For further details on how these results were obtained see Ref. 7.

While experimental studies yield detailed information on the thermodynamics and kinetics of first order transitions and can place some constraints on the possible atomic motions that carry the system from one structure to the other, they lack the time and space resolution required to infer the exact atomic rearrangement mechanism. In principle, this information can be provided by computer simulations such as molecular dynamics[8,9], which permit to follow the detailed motion of individual atoms, or molecules, as the transition occurs. The high free energy barriers opposing the rapid transformation from the metastable to the thermodynamically preferred stable phase, however, represent a huge challenge for particle based computer simulations due to the wide gap between the time scales of atomic motions and those of the nucleation events. While the height of free energy barriers can be reduced by driving the system sufficiently strongly away from coexistence, such conditions are usually unrealistic with respect to the situation studied in experiments. For more realistic circumstances, straightforward computer simulation methods are not applicable due to prohibitive demands on computing resources originating from small nucleation rates and the resulting long waiting times before nucleation occurs.

Several approaches have been suggested recently to overcome the computational challenge posed by wide time scale gaps including metadynamics[10,11], coarse molecular dynamics[12,13] or temperature accelerated dynamics[14,15]. These methods lead to a speed-up with respect to regular molecular dynamics by introducing a bias acting on a set of predefined collective variables or by raising the temperature in a controlled way, both with the effect of promoting the crossing of free energy barriers and enhancing the rate at which configuration space is sampled. In many cases, however, no information is available on the specific mechanism of the transition. If the initial and the final state of the transition are known and well characterized, transition path sampling (TPS)[16–19], a computational methodology based on the statistical description of all possible pathways connecting two given stable states, is applicable. This method is based on the definition of the transition path ensemble consisting of all dynamical trajectories connecting the initial with the final state. This ensemble of pathways is then sampled using a Monte Carlo procedure designed to harvest trajectories according to their likelihood to occur. Analysis of the collected pathways can then yield important information on the transition mechanism, for instance,

in form of a reaction coordinate capable of quantifying the progress of the transition. In the framework of transition path sampling it is furthermore possible to calculate rate constants.

In the following sections, we will first introduce the fundamentals of the theory and implementation of transition path sampling. Then, we will discuss the application of transition path sampling to the specific case of transitions occurring via nucleation and growth. For an in-depth treatment of the transition path sampling methodology and related methods we refer the reader to the original publications[16–28] and recent review articles[29–36]. Other rare event methods such as metadynamics[10,11], coarse molecular dynamics[12,13], temperature accelerated dynamics[14,15], hyperdynamics[37,38], parallel replica dynamics[39–41], the string method[42–44], and forward flux sampling[45–49] are not treated in this article.

The remainder of this article is organized as follows. In the next section, we will introduce the transition path sampling methodology and explain efficient algorithms to sample the transition path ensemble and calculate reaction rate constants. Then, in Sec. 4, we will discuss statistical tools that can be used to analyze transition pathways collected with transition path sampling and recover the transition mechanism, i.e., to identify the collective coordinates that capture the important physics of the transition. Application of these methods will be demonstrated in Sec. 5 using the pressure-induced wurtzite to rocksalt transition in CdSe nanocrystals and the freezing of supercooled liquids as illustrative examples.

2 Fundamentals of Transition Path Sampling

The typical situation to which transition path sampling can be profitably applied is illustrated in Fig. 2. The landscape has many local maxima, minima and saddle points and represents the potential energy (or free energy) surface of a complex particle system. The landscape has two wide basins, A and B, each consisting of several local minima, separated by a high and rough barrier. While barriers small compared to the thermal energy $k_{\mathrm{B}}T$ can be crossed easily leading to rapid local equilibration within the basins, surmounting the higher barrier between A and B occurs only rarely. Hence, the time evolution of the system, which is governed by some stochastic or deterministic equations of motion, consists of long periods spent within the stable states A and B punctuated by rapid, but rare transitions between them. We assume that states A and B can be characterized easily as regions in configuration space, for instance, by specifying a certain range of an order parameter, while the nature of the barrier separating A and B is unknown. In other words, we know the initial and final states of the transition, but do not know the mechanism followed by the system as it moves from A to B. Finding this mechanism (or possibly several mechanisms) and identifying a reaction coordinate that captures the progress of the reaction is exactly the goal of a transition path sampling simulation.

2.1 Transition Path Ensemble

The transition path sampling method is based on the definition of the transition path ensemble, consisting of all pathways starting in state A at time $t = 0$ and reaching state B within time $\mathcal{T}$. Each pathway, or trajectory, is described as a sequence of microscopic states,

$$x(\mathcal{T}) \equiv \{x_0, x_{\Delta t}, x_{2\Delta t}, \ldots, x_{\mathcal{T}}\}, \tag{3}$$

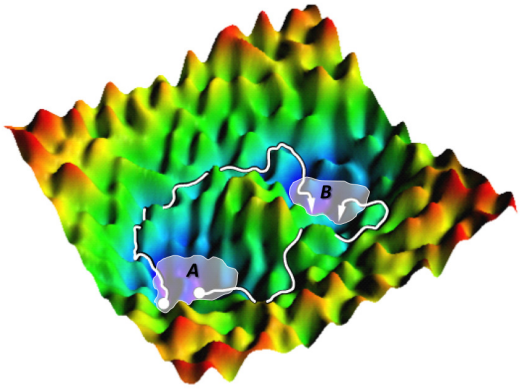


Figure 2. Hypothetical free energy landscape with two long-lived stable states, A and B. These states are stable in the sense that the system spends extended amounts of time in these regions of configuration space. The rough barrier between the stable states is crossed rarely at low temperatures and the system can travel from A to B along different trajectories, two of which are shown as white lines.

where x_t denotes the state of the system at time t specifying the positions and, if required, the momenta of all particles in the system. Thus, each path consists of an ordered series of snapshots of the system separated by a time increment Δt, for instance, the time step (or a multiple of it) of a molecular dynamics simulation. Assuming that the dynamics is Markovian, i.e., that the probability of the future time evolution of the system is fully determined by its current microscopic state, the probability (density) to observe a particular trajectory $x(\mathcal{T})$ is given by

$$\mathcal{P}[x(\mathcal{T})] = \rho(x_0) \prod_{i=0}^{\mathcal{T}/\Delta t-1} p(x_{i\Delta t} \to x_{(i+1)\Delta t}), \tag{4}$$

where $\rho(x_0)$ is the distribution of initial conditions x_0 and $p(x_{i\Delta t} \to x_{(i+1)\Delta t})$ is the transition probability from $x_{i\Delta t}$ to $x_{(i+1)\Delta t}$ within the short time Δt. The particular form of the distribution of initial conditions depends on the particular situation one studies. In typical transition path sampling applications, $\rho(x_0)$ is an equilibrium distribution, such as the canonical or microcanonical distribution[50], but non-equilibrium distributions of initial conditions have been considered as well[51]. Similarly, the specific form of the short time transition probability $p(x_{i\Delta t} \to x_{(i+1)\Delta t})$ depends on the type of underlying dynamics governing the time evolution of the system. For popular types of Markovian dynamics often used in molecular simulations[52], including Newtonian dynamics, thermostatted molecular dynamics in various ensembles, Langevin dynamics, Brownian dynamics or even Monte Carlo dynamics, the short time transition probabilities can easily be written down explicitly (as we will see below) and then be used to construct the probability of an entire trajectory.

The path probability specified in Eq. 4 describes unconstrained pathways and does not include any condition on where the path starts or ends. In a typical transition path sampling simulation one is, however, interested only in those short stretches of the time evolution, in which the rare barrier crossing event takes place. In fact, transition path sampling can be

viewed as a strategy to avoid the simulation of the system during the long and uninteresting waiting times between transitions. To include only reactive trajectories in the ensemble of pathways, i.e., pathways starting in A and ending in B, we restrict the path probability of Eq. 4 by multiplication with appropriate characteristic functions acting on the initial and final points of each trajectory,

$$\mathcal{P}_{AB}[x(\mathcal{T})] \equiv h_A(x_0)\mathcal{P}[x(\mathcal{T})]h_B(x_\mathcal{T})/Z_{AB}(\mathcal{T}). \tag{5}$$

The characteristic function $h_A(x)$ for region A is unity if x is in A and vanishes otherwise,

$$h_A(x) = \begin{cases} 1 & \text{if} \quad x \in A, \\ 0 & \text{if} \quad x \notin A. \end{cases} \tag{6}$$

The characteristic function for region B is defined analogously. The definition of appropriate characteristic functions which include all typical configurations of the initial and final stable states is often non-trivial and usually requires some trial and error. It is important that the definition of A does not include configurations that belong to the basin of attraction of B or vice-versa, as in this case effectively non-reactive trajectories may become part of the transition path ensemble. Frequently, the stable states can be specified by requiring that certain order parameters $q_A(x)$ and $q_B(x)$ are within certain limits, $\lambda_A^{(\min)} < q_A(x) < \lambda_A^{(\max)}$ and $\lambda_B^{(\min)} < q_B(x) < \lambda_B^{(\max)}$, respectively. To study the freezing of a supercooled liquid, for instance, one could define the liquid as initial state A by requiring that the number of particles with a local crystalline environment is below a certain threshold selected to accommodate typical thermal fluctuations, but which excludes any significant crystalline region. Analogously, for the crystalline state, region B, one could require that the number of locally crystalline particle exceeds another threshold far beyond the size of the critical cluster[7].

The path distribution of Eq. 5 is normalized by the path integral

$$Z_{AB}(\mathcal{T}) \equiv \int \mathcal{D}x(\mathcal{T})\, h_A(x_0)\mathcal{P}[x(\mathcal{T})]h_B(x_\mathcal{T}), \tag{7}$$

which can be viewed as a path partition function. In this equation, the notation

$$\int \mathcal{D}x(\mathcal{T}) \equiv \int \cdots \int \mathrm{d}x_0 \mathrm{d}x_{\Delta t} \mathrm{d}x_{2\Delta t} \cdots \mathrm{d}x_\mathcal{T} \tag{8}$$

implies an integration over all time slices of the path. The path probability density of Eq. 5 assigns a statistical weight to all pathways connecting A to B and defines the transition path ensemble (TPE). Sampling this ensemble with methods that will be discussed in subsequent sections yields a collection of transition pathways generated according to their probability to occur in a hypothetical, extremely long straightforward computer simulation. Analysis of these pathways can then provide information to uncover the transition mechanism (or mechanisms).

To complete the definition of the transition path ensemble for a particular application, it is necessary to specify the short time transition probabilities appearing in the product on the right hand side of Eq. 4. This transition probability depends on the dynamics chosen to model the time evolution of the system under study; analytical expressions for various kinds of dynamics are available[16]. For instance, consider a many-particle system evolving

according to Newton's equations of motion,

$$\dot{r} = \frac{\partial \mathcal{H}(r,p)}{\partial p}, \qquad \dot{p} = -\frac{\partial \mathcal{H}(r,p)}{\partial r}. \tag{9}$$

Here, r includes the positions of all particles in the system, p refers to their momenta and $\mathcal{H}(r,p)$ is the total energy of the system. In this case, the time evolution is deterministic and the initial state $x_0 = \{r_0, p_0\}$ completely determines the state $x_t = \{r_t, p_t\}$ of the system a time t later,

$$x_t = \phi_t(x_0). \tag{10}$$

The corresponding propagator $\phi_t(x)$ is a function that uniquely maps x_0 into x_t. Since the time evolution is deterministic, no stochastic spread occurs and the transition probability is given as a Dirac delta function,

$$p(x_t \to x_{t+\Delta t}) = \delta[x_{t+\Delta t} - \phi_{\Delta t}(x_t)]. \tag{11}$$

The transition path ensemble is then given by

$$\mathcal{P}_{AB}[x(\mathcal{T})] = \frac{\rho(x_0)}{Z_{AB}(\mathcal{T})} h_A(x_0) \prod_{i=0}^{\mathcal{T}/\Delta t - 1} \delta[x_{(i+1)\Delta t} - \phi_{\Delta t}(x_{i\Delta t})] h_B(x_{\mathcal{T}}), \tag{12}$$

where the normalizing factor reduces to

$$Z_{AB}(\mathcal{T}) = \int dx_0 \, \rho(x_0) h_A(x_0) h_B(x_{\mathcal{T}}) \tag{13}$$

due to the properties of the delta function. The transition path ensemble of Eq. 12 also describes the statistics of pathways arising from other forms of deterministic dynamics such as Nosé-Hoover dynamics or Gaussian isokinetic dynamics.

For stochastic dynamics, the short time transition probability is spread out rather than singular as a consequence of the noise acting on the system. Consider, for instance, a particle evolving stochastically in the presence of a viscous solvent as described by the Langevin equation in the high friction limit,

$$m\gamma\dot{r} = -\frac{\partial V(r)}{\partial r} + \mathcal{F}, \tag{14}$$

where $V(r)$ is the potential energy as a function of the particle position r, m is the mass of the particle and γ is the friction coefficient. In the above equation, $\mathcal{F}$ is an uncorrelated Gaussian random force with zero mean and a variance given by

$$\langle \mathcal{F}(t)\mathcal{F}(0)\rangle = 2m\gamma k_{\mathrm{B}} T \delta(t). \tag{15}$$

In this case, the short time transition probability takes a Gaussian form[16]

$$p(r_t \to r_{t+\Delta t}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(r_{t+\Delta t} - r_t + \frac{\Delta t}{\gamma m}\frac{\partial V}{\partial r})^2}{2\sigma^2} \right\}, \tag{16}$$

with variance

$$\sigma^2 = \frac{2k_{\mathrm{B}}T}{m\gamma}\Delta t. \tag{17}$$

Thus, this transition probability describes the motion of a particle as consisting of a systematic drift that depends on the external force $-\partial V/\partial r$ and a Gaussian spread describing the diffusion of the particle under the influence of the random noise. Transition probabilities for other types of dynamics can be easily derived[16]. Note that the particular type of dynamics is not a choice made in the framework of transition path sampling, but rather depends on the physical properties of the particular model one intends to study. It is therefore good practice to choose the dynamics that most closely represents the underlying physical situation rather than the one that supposedly offers advantages in the implementation of transition path sampling.

2.2 Sampling the Transition Path Ensemble

The central idea of transition path sampling is to harvest reaction trajectories according to their weight in the transition path ensemble of Eq. 5. This can be achieved by sampling pathways with a Monte Carlo approach that corresponds to carrying out a biased random walk in the space of trajectories. The basic step of this procedure consists of generating a new pathway $x^{(\mathrm{n})}(\mathcal{T})$ from a given pathway $x^{(\mathrm{o})}(\mathcal{T})$. This new trajectory is then accepted or rejected depending on the ratio of the statistical weights of the new and old trajectory in the transition path ensemble. If the newly generated trajectory is accepted, it becomes the current one. In the case of a rejection, however, the old trajectory remains the current one. Iterating this procedure using an appropriate acceptance rule (see below) will generate reactive trajectories with frequencies proportional to their weight in the transition path ensemble. If the sampling is ergodic, this Monte Carlo algorithm will find all important pathways, which can then be further analyzed to identify the reaction mechanism. To start the Monte Carlo procedure one needs an initial reactive pathway that must be generated by other means. The most convenient way to construct an initial pathway depends on the problem under study. In some cases, a high temperature or high pressure molecular dynamics simulation might be used to obtain such a path, while in other cases an initial pathway may be generated according to a postulated transition mechanism. Note that the initial path does not need to be a fully dynamical trajectory satisfying the underlying equations of motion. This freedom often facilitates the initialization of the transition path sampling procedure.

An acceptance/rejection rule that guarantees the desired path distribution is sampled can be derived from the detailed balance condition

$$\begin{aligned} \mathcal{P}_{AB}[x^{(\mathrm{o})}(\mathcal{T})]\pi[x^{(\mathrm{o})}(\mathcal{T}) \to x^{(\mathrm{n})}(\mathcal{T})] = \\ \mathcal{P}_{AB}[x^{(\mathrm{n})}(\mathcal{T})]\pi[x^{(\mathrm{n})}(\mathcal{T}) \to x^{(\mathrm{o})}(\mathcal{T})], \end{aligned} \tag{18}$$

where $\pi[x^{(\mathrm{o})}(\mathcal{T}) \to x^{(\mathrm{n})}(\mathcal{T})]$ is the probability to move from the old path $x^{(\mathrm{o})}(\mathcal{T})$ to the new path $x^{(\mathrm{n})}(\mathcal{T})$ in one Monte Carlo step. The detailed balance condition guarantees stationarity of the path ensemble under the action of the Monte Carlo procedure and requires that the average flux from the old path to the new one is exactly balanced by the flux in reverse direction[52]. Reflecting the two-step character of the basic Monte Carlo move, the transition probability $\pi[x^{(\mathrm{o})}(\mathcal{T}) \to x^{(\mathrm{n})}(\mathcal{T})]$ can be written as a product of the probability to generate the new path from the old one and the probability to accept this newly generated

pathway,

$$\pi[x^{(\mathrm{o})}(\mathcal{T}) \to x^{(\mathrm{n})}(\mathcal{T})] = \\ P_{\mathrm{gen}}[x^{(\mathrm{o})}(\mathcal{T}) \to x^{(\mathrm{n})}(\mathcal{T})] \times P_{\mathrm{acc}}[x^{(\mathrm{o})}(\mathcal{T}) \to x^{(\mathrm{n})}(\mathcal{T})]. \quad (19)$$

Inserting this form of the transition probability into the detailed balance equation 18 yields a condition on the acceptance probability, which can be satisfied with the celebrated Metropolis rule[53] leading to[29]

$$P_{\mathrm{acc}}[x^{(\mathrm{o})}(\mathcal{T}) \to x^{(\mathrm{n})}(\mathcal{T})] = h_A[x_0^{(\mathrm{n})}]h_B[x_{\mathcal{T}}^{(\mathrm{n})}] \\ \times \min\left\{1, \frac{\mathcal{P}[x^{(\mathrm{n})}(\mathcal{T})]P_{\mathrm{gen}}[x^{(\mathrm{n})}(\mathcal{T}) \to x^{(\mathrm{o})}(\mathcal{T})]}{\mathcal{P}[x^{(\mathrm{o})}(\mathcal{T})]P_{\mathrm{gen}}[x^{(\mathrm{o})}(\mathcal{T}) \to x^{(\mathrm{n})}(\mathcal{T})]}\right\}. \quad (20)$$

According to this criterion, only reactive trajectories for which $h_A[x_0^{(\mathrm{n})}] = 1$ and $h_B[x_{\mathcal{T}}^{(\mathrm{n})}] = 1$ can have a non-vanishing acceptance probability ensuring that reactivity is maintained throughout the course of the transition path sampling simulation. From the general form of the acceptance probability as stated in Eq. 20 one can derive specific expressions for the acceptance probability valid for particular forms of path ensemble and trajectory generation[29].

The Monte Carlo procedure described in the previous paragraphs provides a general framework for the development of specific algorithms for the generation of a new pathway from a given one. The efficiency of the transition path sampling simulation, i.e., the rate at which trajectory space is explored, depends crucially on this central part of the Monte Carlo procedure. Exploiting the freedom given by the Monte Carlo approach, several path generation algorithms have been proposed with acceptance probabilities derived from the detailed balance condition[30]. Here, we discuss the so-called shooting move[17,18], because it has proven particularly efficient and is generally applicable.

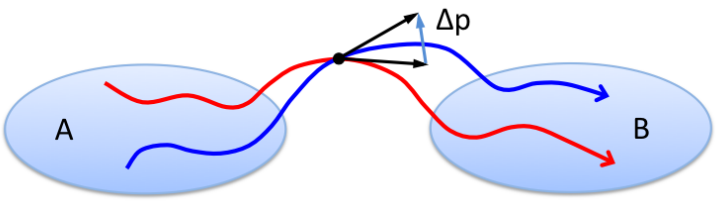


Figure 3. In a shooting move, a new pathway (blue) is generated from an old one (red) by randomly selecting one time slice on the path, changing the momenta at that time slice by a displacement Δp generated from an appropriate probability distribution in momentum space, and finally integrating the equations of motion forward and backward with the new momenta. In the case of stochastic dynamics, the momentum displacement may be useful but is not strictly necessary, because the new and old trajectories diverge naturally due to the different noise histories used in the generation of the trajectories. If the new trajectory is still connecting the stable states, it is accepted with a probability that depends on the ensemble under study. Non-reactive trajectories are rejected.

The basic idea of the shooting move is depicted schematically in Fig. 3. In this move, which makes use of the natural tendency of the system to relax into the stable states, a time

slice $x_{t'}^{(\mathrm{o})}$, called the shooting point, is selected randomly from the given path, shown in red in the figure. This time slice is then altered, for instance, by adding a small perturbation Δp, yielding the new time slice $x_{t'}^{(\mathrm{n})}$. Starting from this time slice one then integrates the equation of motion forward to time $\mathcal{T}$ and backward to time 0 obtaining a complete new path $x^{(\mathrm{n})}(\mathcal{T})$. Since two initially close points in phase space diverge under the action of chaotic dynamics, the new trajectory will differ from the old one. The magnitude of the difference can be controlled by choosing a perturbation with an appropriate amplitude. Thus, new trajectories can be generated with a high probability of being reactive, which is necessary to obtain a satisfactory average acceptance probability and hence a good sampling efficiency[17–19]. Note that for stochastic dynamics the perturbation step can be omitted since, due to the stochastic noise, two trajectories will diverge even if they are initiated at the same point in phase space. Modifying the shooting point can, however, increase the efficiency of a transition path sampling simulation also in the case of stochastic dynamics.

Since the generation of trajectories with the shooting method is done using the natural dynamics of the system, most factors of the acceptance probability in Eq. 20 cancel, leading to a very simple form of the acceptance probability[31],

$$P_{\mathrm{acc}}[x^{(\mathrm{o})}(\mathcal{T}) \to x^{(\mathrm{n})}(\mathcal{T})] = h_A[x_0^{(\mathrm{n})}]h_B[x_{\mathcal{T}}^{(\mathrm{n})}] \min\left[1, \frac{\rho(x_{t'}^{(\mathrm{n})})}{\rho(x_{t'}^{(\mathrm{o})})}\right], \tag{21}$$

where $\rho(x)$ is the stationary distribution of the dynamics. For equilibrium systems the stationary distribution is the equilibrium distribution, but in non-equilibrium situations other steady-state distributions may be appropriate. The acceptance probability of Eq. 21 is valid for any dynamics that is microscopically reversible, including Newtonian dynamics as well as stochastic dynamics such as Langevin dynamics. According to the above acceptance probability, non-reactive pathways are rejected while trajectories connecting A with B are accepted with a probability that depends on how the perturbation applied to the shooting point has changed its probability in the stationary distribution. If the shooting point is left unchanged or the dynamics is Newtonian with a microcanonical distribution of initial conditions, the acceptance probability further simplifies to

$$P_{\mathrm{acc}}[x^{(\mathrm{o})}(\mathcal{T}) \to x^{(\mathrm{n})}(\mathcal{T})] = h_A[x_0^{(\mathrm{n})}]h_B[x_{\mathcal{T}}^{(\mathrm{n})}], \tag{22}$$

implying that the new trajectory is accepted if it is reactive and rejected otherwise.

Applying the shooting algorithm, usually combined with other path generation moves, one can perform transition path sampling simulations that efficiently sample reactive trajectories. If the transition of interest can occur through different and unconnected classes of pathways, exploration of trajectory space may be slow. This situation is quite analogous to sampling problems in conventional Monte Carlo simulations arising from large barriers in configuration space. To improve the ergodicity of Monte Carlo simulations, a number of approaches has been put forward, which can be applied also in the transition path sampling scheme. For instance, parallel tempering[54] carried out at the path level[55] can dramatically improve the sampling, as can the application of the Wang-Landau flat histogram approach[28]. Another complication arises if broad barriers are crossed diffusively. In this case, achieving a good acceptance probability is difficult, because, due to the long barrier crossing times, the trajectory divergence is uncontrollably large. A solution, based on introducing small stochastic elements in an otherwise deterministic dynamics combined with

the generation of partial trajectories has been put forward in Ref. 56. Another solution to this problem consists in using linearized equations of motion for small displacements[57].

3 Kinetics

Determining reaction rates from computer simulations is important, because often this offers a useful way to link simulation results with experimental observations providing important information on the reaction mechanism. For instance, analysis of the kinetics of pressure induced phase transitions in nanocrystals as a function of particle size shed light on the geometry of the critical nucleus and, hence, on the nucleation and growth mechanism governing the transition[58,59]. The central quantity used to characterize the kinetics of homogeneous nucleation in macroscopic systems is the nucleation rate J, defined as the number of nuclei that grow beyond critical size per unit time and unit volume. Alternatively, one may specify the transformation rate constant k for the entire system, related to the nucleation rate J simply by a multiplication with the volume, $k = JV$. Since a single successful nucleation event transforms the entire system irreversibly into the stable phase, the conditional probability $P_{AB}(t)$ for a system of volume V, prepared in the metastable phase (state A) at time 0, to be observed in the stable phase (state B) at time t is governed by the differential equation

$$\frac{\mathrm{d}P_{AB}(t)}{\mathrm{d}t} = k[1 - P_{AB}(t)]. \tag{23}$$

In a system of macroscopic size the transformation is effectively irreversible, because the stable phase is overwhelmingly more stable than the metastable phase. Therefore the barrier for the backward transformation from the stable to the metastable phase is excessively high such that the transformation never occurs in this direction. The solution of the above equation, in which the factor $[1 - P_{AB}(t)]$ takes into account that nucleation can only occur if the system has not transformed yet, yields an exponential approach to the asymptotic value of unity with a characteristic time $\tau = 1/k = 1/JV$,

$$P_{AB}(t) = 1 - \exp(-t/\tau). \tag{24}$$

Thus, provided one waits long enough, any metastable state will eventually convert to the thermodynamically stable state. (The conversion time may be exceedingly long; for instance, at room temperature a diamond crystal would not convert into the more stable graphite form even on time scales comparable to the age of the universe.) Note that for heterogeneous nucleation occurring at surfaces it is appropriate to consider a nucleation rate that specifies the number of nucleation events per unit time and unit surface area rather than unit volume. Furthermore, in small systems the transformation might be reversible and a dynamical equilibration between the two phases might be observed, in particular close to coexistence[60,61]. In this case, transformation in the backward direction must be explicitly taken into account in the kinetic description of the process[36] and the equation for the probability $P_{AB}(t)$ becomes

$$\frac{\mathrm{d}P_{AB}(t)}{\mathrm{d}t} = k_{AB}[1 - P_{AB}(t)] - k_{BA}P_{AB}(t). \tag{25}$$

Here k_{AB} and k_{BA} are the transformation rate constants for the forward transformation from A to B and the backward transformation from B to A, respectively. Equation 25 is

solved by

$$P_{AB}(t) = P_B[1 - \exp(-t/\tau)], \tag{26}$$

where the characteristic time τ for the approach to the asymptotic value, the relaxation time, is given by $\tau = (k_{AB} + k_{BA})^{-1}$ and P_B is the equilibrium probability to find the system in state B. The probabilities P_A and P_B to find the system in state A and B, respectively, are related to the forward and backward transformation rate constants by $P_A k_{AB} = P_B k_{BA}$, as follows by requiring stationarity. Thus, in this case, the probability $P_{AB}(t)$ behaves exponentially as well, but with a relaxation time that depends both on the forward and backward transformation rate constants and is dominated by the larger one of the two. For nucleation in a macroscopic system, the backward transformation never occurs and can be neglected, corresponding to the case $k_{AB} = k = JV$, $k_{BA} = 0$, and $P_B = 1$.

According to Eq. 24, the nucleation rate J can be determined experimentally[62,63] or in molecular simulations[64,65] by determining the average time $\langle t \rangle$ one has to wait for a nucleation event starting from a system prepared in the metastable state,

$$J = \frac{1}{V\langle t \rangle}. \tag{27}$$

While this approach is usually feasible in experiments, nucleation rates can be calculated from straightforward molecular dynamics simulations only in exceptional cases when the nucleation barrier is made sufficiently low by selecting conditions far away from coexistence. Under typical experimental conditions, however, nucleation is unlikely to occur even once on the time scales accessible to the simulations. Homogeneous crystallization of supercooled liquid water, for instance, occurs with nucleation rates of only $J = 10^4 - 10^9\,\mathrm{cm}^{-3}\mathrm{s}^{-1}$ even at temperatures as low as -35 to $-37\,^\circ$C, as is known from freezing experiments carried out on levitated water droplets[63]. Even at such extreme conditions far below freezing a molecular dynamics simulation of about 1000 water molecules in a box with a side length of about 3 nm would yield only one nucleation event every $10^{10} - 10^{15}$ seconds of real time requiring $10^{25} - 10^{30}$ molecular dynamics steps, which is far beyond the capabilities of any currently available or imaginable computer system. The origin for such long nucleation times can be understood in terms of classical nucleation theory, as briefly discussed in the Introduction. Since the creation of the critical nucleus implies a free energetic cost ΔG^* associated with the creation of the interface between the metastable and stable phase at the surface of the growing nucleus, configurations corresponding to the barrier top have a small likelihood to occur proportional to the Boltzmann factor $\exp\{-\beta\Delta G^*\}$. Here, $\beta = 1/k_\mathrm{B}T$ is the reciprocal temperature. Since formation of the critical nucleus is the crucial (and least likely) event of the whole nucleation process, this factor appears also in the nucleation rate,

$$J \propto \exp(-\beta\Delta G^*). \tag{28}$$

Due to this exponential dependence of the nucleation rate on the barrier height ΔG^*, nucleation time can be exceedingly long even if ΔG^* is only a moderately large multiple of the thermal energy $k_\mathrm{B}T$.

Such long time scales do not represent a difficulty for transition path sampling, which was specifically designed to circumvent the problem of widely separated time scales. But while transition rate constants can be determined in the general framework of transition

path sampling, their computation requires procedures that go beyond those laid out in previous section. Transition path sampling focuses on reactive trajectories, i.e., on trajectories belonging to the transition path ensemble. From these trajectories alone, however, transition rate constants cannot be determined. The reason is that the transition path ensemble specifies only the relative weight of reactive trajectories compared to each other, but not the absolute likelihood of these trajectories. To compute transition rate constants it is, thus, necessary to compute the probability of observing a reactive trajectory compared to the probability of a trajectory without this condition. In other words, one needs to determine the combined statistical weight $Z_{AB}(\mathcal{T})$ of the transition path ensemble of Eq. 5 with respect to the weight of the ensemble of unrestricted pathways described by the path distribution of Eq. 4. This general procedure lies at the core of all transition path sampling approaches for the calculation of rate constants as well as other methods such as forward flux sampling[45,46,48].

3.1 Rate Constants from Path Free Energies

One TPS-approach for the calculation of transition rate constants[16,17,31] exploits the fact that, for short times, the conditional probability $P_{AB}(t)$ is linear with slope k_{AB},

$$P_{AB}(t) \approx k_{AB} t. \tag{29}$$

This equation is valid for times $t \ll \tau$ after short time transients, related to the specific way the barrier is crossed, have decayed. To determine the conditional probability $P_{AB}(t)$, one expresses it as ratio of two path ensemble averages,

$$P_{AB}(t) = \frac{\langle h_A(x_0) h_B(x_t) \rangle}{\langle h_A(x_0) \rangle} = \frac{\int \mathcal{D}x(t) \mathcal{P}[x(t)] h_A(x_0) h_B(x_t)}{\int \mathcal{D}x(t) \mathcal{P}[x(t)] h_A(x_0)}. \tag{30}$$

Here, the numerator depends on time but the denominator is time independent and equals the equilibrium probability of the initial state, P_A. This ratio can be viewed as a ratio of two partition functions belonging to two distinct ensembles of pathways. The path integral in the denominator corresponds to the partition function of all pathways having their initial point x_0 in A without any requirement on the location of the path endpoint x_t. The path integral in the numerator, on the other hand, is the partition function of all pathways starting in A and ending in B. The ratio of partition functions can, thus, be calculated by determining the reversible work, or free energy, in path space, required to transform between these two path ensembles[66,67].

The path free energy required to change from a path ensemble with unconstrained endpoint to one in which the endpoint is required to be in region B can be determined by introducing an order parameter $\lambda(x)$ in such a way that region B corresponds to a specific range of this parameter, $\lambda_B^{(\min)} < \lambda(x) < \lambda_B^{(\max)}$, and the entire configuration space including region A corresponds to $-\infty < \lambda(x) < \infty$. From a path sampling simulation of pathways starting in A without condition on their endpoint one could then, in principle, calculate the probability distribution $P_A(\lambda, t)$ to find the path endpoint at a particular value λ of the order parameter. Using this distribution, the sought conditional probability $P_{AB}(t)$ can be expressed as an integral over region B,

$$P_{AB}(t) = \int_{\lambda_B^{(\min)}}^{\lambda_B^{(\max)}} \mathrm{d}\lambda \, P_A(\lambda, t). \tag{31}$$

Since we are dealing with rare events, the distribution $P_A(\lambda, t)$ can be very small in the range of interest such that a direct calculation is impractical. One can solve this problem by setting up the path equivalent of an umbrella sampling simulation[68] and carry out a series of independent path sampling simulations in which the order parameter $\lambda(x_t)$ at the path endpoint is required to be inside a particular window. Order parameter distributions calculated separately for a set of overlapping windows can then be connected to form $P_A(\lambda, t)$, from which the conditional probability $P_{AB}(t)$ can be determined according to Eq. 31. Combined with a particular and convenient factorization of $P_{AB}(t)$ this basic procedure yields a practical algorithm[31] for the calculation of transition rate constant.

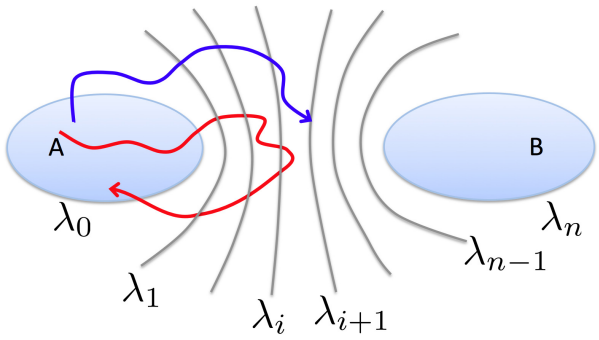


Figure 4. In a transition interface sampling simulation one considers a set of interfaces defined as iso-surfaces of an appropriate order parameter, $\lambda(x) = \lambda_i$, where the index i runs from 0 to n and numbers the interfaces. The boundaries of regions A and B correspond to the values λ_0 and λ_n, respectively. Trajectories belonging to the path ensemble of interface λ_i are required to start in region A and cross interface λ_i. While the red trajectory crosses interface λ_i and then returns to A before reaching interface λ_{i+1}, the blue trajectory, also crossing interface λ_i, reaches interface λ_{i+1} first.

3.2 Transition Interface Sampling

Another transition path sampling approach to determine transition rate constants, called transition interface sampling (TIS)[25–27,69,70], is based on the definition of a set of interfaces spanning the space between region A and region B, as illustrated in Fig. 4. Akin to the order parameter windows of the previous paragraph, these interfaces can be defined to be the iso-surfaces of the order parameter λ for a set of $n + 1$ values λ_0, λ_1, $\cdots$, λ_n increasing monotonically. In this setting, we define region A by $\lambda(x) \leq \lambda_0$ and region B as $\lambda(x) \geq \lambda_n$. The theoretical basis of transition interface sampling is the expression of the rate constant as a product of the effective positive flux $\Phi_{1,0}$ out of region A and the conditional probability $P_A(\lambda_n|\lambda_1)$ that trajectories that cross interface λ_1 reach region B,

$$k_{AB} = \Phi_{1,0} P_A(\lambda_n|\lambda_1). \tag{32}$$

The effective positive flux $\Phi_{1,0}$ is defined as the number of times interface λ_1 is crossed by trajectories originating in state A. Here, "effective" means that only trajectories coming directly from A are counted (i.e., additional crossings of interface λ_1 without return to

A are not considered) and "positive" implies that only crossings of interface λ_1 towards B contribute. For an appropriate definition of interfaces λ_0 (i.e., the boundary of region A) and λ_1, the effective positive flux $\Phi_{1,0}$ can be computed in straightforward molecular dynamics simulation of state A by counting the number, per unit time, of first crossings of interface λ_1 after the system has left A.

The second factor in Eq. 32, the crossing probability $P_A(\lambda_n|\lambda_1)$ that a trajectory coming from A and crossing interface λ_1 reaches the final region B, is more difficult to calculate, because typically it is a very small number due to the high free energy barrier separating the stable states. This difficulty can be circumvented by expressing the crossing probability as a product of conditional crossing probabilities depending only on adjacent interfaces,

$$P_A(\lambda_n|\lambda_1) = \prod_{i=1}^{n-1} P_A(\lambda_{i+1}|\lambda_i). \tag{33}$$

Here, the conditional probability $P_A(\lambda_{i+1}|\lambda_i)$ denotes the probability that a trajectory coming from A and crossing interface λ_i reaches interface λ_{i+1} rather than returning to A. If interfaces λ_i and λ_{i+1} are sufficiently close together, the conditional probability $P_A(\lambda_{i+1}|\lambda_i)$ is a number of order 1, which can be computed in a transition path sampling simulation of trajectories that are required to start in A and to cross interface λ_i. Trajectories in this ensemble either go on to cross interface λ_{i+1} or go back to A without reaching λ_{i+1} first. The local crossing probability $P_A(\lambda_{i+1}|\lambda_i)$ is then simply estimated as the fraction of trajectories of the former type. (Note that trajectories sampled in this scheme have variable length as the integration of the equations of motion can be stopped as soon as interface λ_{i+1} or the boundary of region A are reached.) Combining the results of path sampling simulations separately carried out for all interfaces λ_1 to λ_{n-1} with the effective positive flux $\Phi_{1,0}$ computed in a molecular dynamics simulation, one finally obtains the transition rate constant k_{AB} for the entire process. How to place the interfaces in order to optimize the efficiency of the TIS-simulation has been discussed in Ref. 71.

For processes involving slow, diffusive barrier crossing events, the efficiency of the simulation can be dramatically increased by exploiting the memory loss along the corresponding long transition pathways[26]. In this approach, called partial path transition interface sampling (PPTIS), one considers short trajectory segments (partial paths) that only cross one or two adjacent interfaces and are not required to start in stable state A. Several other enhancements of the TIS-formalism are presented and discussed in Ref. 69. Combining TIS with replica exchange moves can considerably improve the efficiency of a transition interface simulation[72,73]. Furthermore, by applying a recently developed approach[74] to the pathways sampled in a transition interface sampling simulation, one can remove the bias introduced by requiring that certain interfaces are crossed and reconstruct an unbiased ensemble of pathways. From this reweighted path ensemble (RPE) it is possible to calculate equilibrium free energies with respect to arbitrary variables and to identify complex non-linear reaction coordinates using a maximum likelihood approach[75,76].

3.3 Activation Energies from Transition Path Sampling

As an alternative, transition rate constants can be determined by a generalization of the thermodynamic integration method[77] to the space of trajectories. Here, the basic idea is to

use the transition path sampling method to calculate the derivative of the transition rate constant with respect to a control parameter, rather than the transition rate constant itself[24,28]. This control parameter could, for instance, be the temperature, the pressure, or an interaction parameter. Starting from conditions for which the transition rate constant is known, i.e., from a reference state, one can then determine the transition rate constant under other conditions by integrating its derivative with respect to the control parameter[24,28]. The advantage of this procedure is that the derivative of the transition rate constant with respect to the control parameter can be calculated from a straightforward TPS-simulation without the need to partition space with interfaces or other special arrangements. Furthermore, the derivative itself can be of interest as it yields the activation energy or activation volume, in the case of the temperature and pressure derivatives, respectively. Also, no definition of interfaces or any a priori knowledge of the transition mechanism is necessary. The drawback of the method is that the transition rate constant derivative, which is expressed in terms of path averages taken in the transition path ensemble, may be affected by considerable statistical uncertainties[24,28], such that enhancements of this general approach are necessary in order to develop it into an efficient method for the calculation of transition rate constant.

4 Identifying the Transition Mechanism

A transition path sampling simulation typically yields in full atomistic detail many examples of transition trajectories along which the system moves from the initial to the final state. While inspection of these trajectories with a molecular visualization program may reveal some interesting features of the transition, it does not automatically yield a detailed understanding of the underlying mechanism in terms of a small number of collective degrees of freedom. For instance, watching a crystalline nucleus grow in a supercooled liquid does neither tell us at which stage the nucleus has reached critical size nor does it reveal if other variables besides the nucleus size play an important role in the transition. Such information can only be extracted by subjecting harvested pathways to further statistical analysis, ideally resulting in a reaction coordinate.

In general, the reaction coordinate is a function $q(r)$, usually defined in configuration space, which quantifies the progress of a reaction. For the freezing of a supercooled liquid, the size of the largest crystalline nucleus may serve for this purpose in some situations, but in other cases it might be necessary to include other parameters such as the shape, the structure, and the surface properties of the crystallite into the reaction coordinate. Finding an appropriate reaction coordinate for a transition occurring in a complex molecular system, however, is usually quite difficult. Furthermore, it is important to realize that there is some arbitrariness in the definition of the reaction coordinate, as often different variables may by used for this purpose. What, then, is a criterion for a good reaction coordinate $q(r)$? It is reasonable to require that the reaction coordinate tells us, in a quantitative way, how far a particular transition has proceeded and what will most likely happen next. In particular, by looking at the reaction coordinate $q(r)$ of a particular configuration r one should be able to tell whether r is a transition state or if it rather belongs to the basins of attraction of A and B.

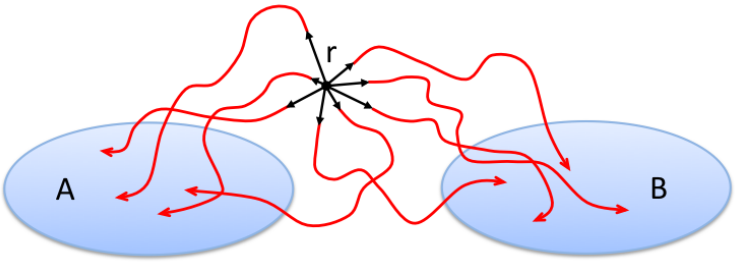


Figure 5. For a particular configuration r the committor $p_B(r)$, or commitment probability, is defined as the probability that a trajectory started at r with random initial momenta will relax into state B rather than state A. The committor can be estimated by initiating N trajectories at r and determining the number N_B of trajectories that reach B before they reach A, $p_B \approx N_B/N$.

4.1 Committor

A practical criterion to gauge the quality of a postulated reaction coordinate can be established by considering the commitment probability $p_B(r)$, also called committor. The commitment probability $p_B(r)$ for a given configuration r is defined as the probability that a trajectory started from r will reach stable state B rather than stable state A. The general concept of the commitment probability goes back at least to Onsager[78], who introduced it under the name of splitting probability to analyze ionic dissociation, and was more recently applied in the context of protein folding[79], where it is known as p_{fold}. In practice, the committor of a particular configuration r can be estimated by initiating a certain number of trajectories at r and counting the fraction that reach B rather than A if followed in time. As a probability, the committor is a number ranging from 0 to 1, which specifies how "committed" a particular configuration is to region B. Configurations close to region A will have a committor of $p_B \approx 0$, while configurations near B correspond to $p_B \approx 1$. The committor also provides a natural definition of transition states[80–84] as configurations with $p_B \approx 1/2$, i.e., intermediate configurations that relax to A and B with equal probability. This statistical transition state criterion, which generalizes the concept of the transition state as a saddle point on the potential energy surface, as familiar from chemical dynamics, also provides a generally applicable definition of the critical nucleus. Accordingly, the critical nucleus is that nucleus which shrinks and grows with the same probability. It is worth noting that, in the context of nucleation and growth, the $p_B = 1/2$ criterion for the critical nucleus was introduced by Honeycutt and Andersen already in 1984 to study the freezing of a supercooled atomic liquid[85,86].

The committor can also be used to quantify the quality of a reaction coordinate. As discussed above, a good reaction coordinate should tell us the likely fate of a trajectory passing through a particular configuration. This information is exactly given by the committor, based on which one learns how far the transition has proceeded and what is likely to happen next. As such, the committor may be viewed as the perfect reaction coordinate[44,87–89] encoding all the relevant information. While the definition of the reaction coordinate as the committor is generally valid, it is unspecific and not particularly useful because one desires to express the reaction coordinate in terms of variables with a physically transparent meaning that can also be probed or even controlled in experiments. The committor,

however, provides the basis for distinguishing between a good and a poor reaction coordinate. From a good reaction coordinate $q(r)$ we require that its value for configuration r determines the committor at r. In other words, a good reaction coordinate should parameterize the committor, $p_B(r) = p_B[q(r)]$. A poor reaction coordinate, on the other hand, is not sufficient to specify the value of the committor. In the following, we will briefly discuss several committor-based approaches for identifying good reaction coordinates and extracting detailed information about the mechanism underlying the rare transition under study.

4.2 Transition State Ensemble

One way to analyze transition pathways harvested by transition path sampling consists in determining and comparing configurations with given committor values. In particular, it is often useful to inspect the properties of the transition state ensemble (TSE), defined as the set of all configurations on the collected transition pathways that have a committor of $1/2$. Members of the transition state ensemble can be determined by calculating the committor for regularly spaced configurations along the transition pathways. Since the committor is small for configurations near A and approaches unity near B, at least one but possibly several transition states exist on each transition pathway. Comparison of such TSE-configurations with ensembles of configurations corresponding to other committor values has revealed important features of the transition mechanism in many cases, including ionic dissociation in water[20], biomolecular isomerization[21], transitions in nanocrystals[58,59] and the liquid-solid transition[7,90–93]. In the case of the freezing transition, the wide distribution of cluster sizes observed in the transition state ensemble indicates that the cluster size is not sufficient to capture all essential properties of the transition[7,90–93].

4.3 Committor Distributions

As explained above, a good reaction coordinate parameterizes the committor, i.e., for any configuration r the reaction coordinate $q(r)$ completely determines the committor. This requirement can be used to test the quality of a postulated reaction coordinate $q(r)$ by plotting the committor as a function of this particular reaction coordinate for a set of configurations taken from the sampled transition pathways. In the case of a good reaction coordinate, one expects all points to be roughly on a smooth curve expressing the unique dependence of the committor on the reaction coordinate. Deviations from the smooth curve should be only due to the statistical inaccuracy of the committor arising from the finite number of trajectories used to determine it. In the case of a poor reaction coordinate, the reaction coordinate does not fully determine the committor and configurations with the same reaction coordinate can have different committor values. This lack of a unique relation between reaction coordinate and committor leads to considerable spread of the p_B-vs.-q plot that is symptomatic for a poor choice of the reaction coordinate. For the freezing of a supercooled liquid, for instance, the committor $p_B(n)$ takes values ranging from 0 to 1 for configurations with the same value n of the largest cluster size, indicating that the size of the crystallite does not adequately capture the complete transition mechanism and other parameters must be taken into account[7,90,94].

While a large spread in the p_B-vs.-q plot implies a poor choice of reaction coordinate, a small spread is not a sufficient condition for a good reaction coordinate. The reason is

that transition pathways cover only a small portion of configuration space, in which the postulated reaction coordinate might just adiabatically follow the true reaction coordinate, but considerably deviate from it in other parts of configuration space. Such a reaction coordinate, while passing the p_B-vs.-q test, would not be suitable to control the transition of the system from the initial to the final state. A more stringent committor-based path analysis, which is also able to detect such imperfect reaction coordinates, consists in calculating committor distributions $P(p_B)$ for equilibrium ensembles of configurations with a given value q^* of the postulated reaction coordinate $q(r)$[20],

$$P(p_B) = \langle \delta[p_B - p_B(r)] \rangle_{q(r)=q^*} \tag{34}$$

Here the angular brackets $\langle \cdots \rangle_{q(r)=q^*}$ denote an equilibrium average restricted to $q(r) = q^*$. In the case of a good reaction coordinate, i.e., if the reaction coordinate determines the committor, the committor distribution $P(p_B)$ will be strongly peaked around the value $p_B(q^*)$. (The residual spread is due to the statistical error in the committor, as pointed out by Peters[95].) For a poor reaction coordinate, on the other hand, the committor distribution may be broad and even have more than one maximum.

An analysis based on committor distributions can be particularly revealing if it is carried out for a value q^* of the postulated reaction coordinate that corresponds to the maximum of the free energy curve determined as a function of q. In this case and provided that $q(r)$ is a good reaction coordinate, configurations with $q(r) = q^*$ are expected to be transition states with $p_B = 1/2$. Correspondingly, the committor distribution computed at $q(r) = q^*$ should have a single narrow peak at $p_B = 1/2$. Committor distributions deviating from the unimodal form indicate that additional variables must be included for a complete description of the transition mechanism. In the case of the dissociation of two ions in liquid water[20], for instance, the interionic distance turned out to be an insufficient reaction coordinate. Deceptively, the free energy profile computed as function of this variable displayed a barrier separating the associated state, where the ions are close, from the dissociated state, where the ions are separated by one water molecule. However, the committor distribution, determined with the interionic distance constrained at a distance corresponding to the top of the barrier, displayed two peaks, one at 0 and one at 1, implying that most configurations at the barrier top were not transition states but rather clearly belonged to the basins of attraction of the stable states. Committor distributions have been used in several other circumstances to determine the quality of a guessed reaction coordinate[21,76,87,90,92,93,96–98].

4.4 Extracting the Reaction Coordinate

While determining the transition state ensemble and committor distributions may yield valuable insights into the transition mechanism, ideally one would like to have a more systematic approach to extract knowledge about the reaction coordinate from the information stored in committor values. In particular, it would be very useful to identify the relevant variables that capture the physics underlying the reaction mechanism and to separate them from other variables which may be treated as random noise. Two approaches have been developed recently for this purpose and we will discuss them briefly in the following.

In the method proposed by Ma and Dinner[87], genetic neural networks (GNN)[99,100] are used to screen large pools of possible reaction coordinates and to single out the combinations of a few variables that best reproduce the functional dependence of the committor

on the configurational variables. For this purpose, one first carries out a transition path sampling simulation and collects a sufficient number of configurations, say a few thousand of them, from the harvested transition pathways. For these configurations one then computes the committor as well as a long list of collective variables which may contribute to the reaction coordinate. This list of variables, compiled based on any prior knowledge and/or intuition one might have about the transition mechanism, can be rather long with thousands of entries. The weight coefficients of the neural networks are then optimized for combinations of a few collective variables and a genetic algorithm is used to search for the variable set that leads to the smallest deviation between the predicted and measured committor values. Applied to the isomerization of alanine dipeptide in vacuum and explicit solvent[87], the genetic neural network procedure has produced a set of collective variables that include internal as well as solvent degrees of freedom and point to the importance of long-range electrostatic interactions for the isomerization process.

Another approach to process commitment data is to determine the optimum reaction coordinate by likelihood maximization as proposed by Peters and Trout[76,97,101]. In this method, a maximum likelihood[102] procedure is used to identify the parameters of a postulated reaction coordinate model that best explain the observed commitment probabilities. A nice feature of this approach is that it does not necessarily require explicit calculation of committors, but can also make use of the information on acceptances and rejections acquired during a transition path sampling simulation, which are viewed as single instances of a committor calculation. (Note that this can be done only if the shootings moves are carried out without bias following the aimless shooting procedure.) The maximum likelihood method is based on the construction of a model that stipulates how the reaction coordinate depends on a list of collective variables. In the simplest case, a linear combination of the collective variables fed into a sigmoidal switching function, for instance a hyperbolic tangent, is used for this purpose[76]. Likelihood maximization has been used to study the mechanism of magnetization reversal in the Ising model[76] and structural transitions in solid terephtalic acid[97]. Recently, a procedure to carry out likelihood maximization with a non-linear reaction coordinate model has been proposed by Bolhuis and collaborators[75] by using ideas of the string method[42–44]. Application of this flexible maximum likelihood algorithm to the freezing of a soft sphere liquid[92,93] has shown that the solid nucleus is embedded in a cloud of highly correlated yet non-crystalline surface particles.

5 Applications

To date, transition path sampling has been applied to investigate a variety of processes involving rare but important events including chemical reactions[103–112], solvation processes[20,113–116], the dynamics of liquids and clusters[23,117–121], glassy dynamics[122–124], transport and diffusion[125], single-file water dynamics[71], biomolecular isomerizations[21,87,126,127], protein folding[128–130], DNA dynamics[131–135], membrane dynamics[136,137], chaotic dynamics[138], and non-equilibrium processes[51,121,139–145]. In particular, transition path sampling has been extensively employed to investigate nucleation and growth processes occurring at first order phase transitions ranging from magnetization reversal in the Ising model[96], pressure induced phase transitions in semiconductor nanoclusters[146], the freezing of soft sphere systems[90–93] and of mixtures[7], crystallization of a supersaturated solution[147], demixing of a binary mixture[148], phase separation and crystallization

from the melt[149], the solid-solid transition of terephtalic acid[97], the liquid-vapor transition of methane[150], the wurtzite to rocksalt transition in bulk CdSe[151], and the boiling of water[152] to pressure induced transitions of alkali halides[153–158] and heterogeneous nucleation around a tiny seed[94]. In the following, we will illustrate the application of transition path sampling to nucleation using the pressure induced wurtzite to rocksalt transition in CdSe nanocrystals[9,58,59,146,159,160] and the freezing of supercooled liquids[7,94] as examples.

5.1 Pressure Induced Structural Transitions in Nanocrystals

When high pressure is applied, many solid materials undergo a first order phase transition from a low-density crystal structure to a structure with higher density and coordination. Predicting the occurrence and stability of different crystal structures from first principles is an important and challenging research field[161–163]. However, it is often the kinetics of a structural transformation, rather than the relative thermodynamic stability of the crystal structures, that dominates the phase behavior of crystalline solids. A well-known example is carbon, whose diamond phase persists on astronomical time scales under ambient conditions despite the thermodynamic stability of graphite. To understand the origin of such structural metastability, a fully dynamic view of the transformation process on the atomic scale is indispensable.

Despite major advances of the time and space resolution of electron microscopes[60], experiments can only provide a coarse-grained view of atomistic rearrangements in solids. Molecular dynamics computer simulations have therefore emerged as the main tool for revealing the microscopic mechanisms of structural transformations. However, the computer simulation of the nucleation of a structural transformation is plagued by the very time scale problem discussed in Sec. 1. Under experimental conditions, the free energy barrier associated with the transformation is typically large and prevents observation of the process with straightforward molecular dynamics simulation. Of course, the transformation becomes observable when much higher pressures are used that practically render the low-pressure structure mechanically unstable. The disadvantages of such an approach are twofold: First, mechanisms can be quite different compared to experimental conditions. Second, observables that link experiment and simulation, like the rate constant or its derivatives, cannot be directly compared with experimental values. These drawbacks can be avoided by the use of transition path sampling methods.

Transition path sampling has been successfully applied to the study of the homogeneous nucleation of pressure-induced transformations in bulk materials[151,158]. In nanocrystals, however, structural transformations can proceed strikingly different from the corresponding bulk transformation due to the strong influence of surface free energies. The phase diagrams and kinetics that govern the transformation process typically depend on particle size, shape, and surface composition. This sensitivity to surface properties offers the exciting possibility of stabilizing structures in nanocrystals that are unstable in the bulk through suitable surface modifications. This prospect is particularly intriguing in semiconductor nanocrystals, whose opto-electronic properties change dramatically during structural transformations.

Driven by a comprehensive experimental study by Alivisatos and coworkers[62,164–167], the semiconductor CdSe has emerged as the prototypical material for the study of pressure-induced transformations in nanocrystals. Alivisatos and coworkers showed that the ther-

modynamic transition pressure of the wurtzite to rocksalt transformation decreases with increasing crystal size[164], while activation enthalpies and volumes increase[62,165]. This size-dependence results in metastability of the high-pressure rocksalt structure at ambient conditions for crystals larger than a certain threshold size[167]. The microscopic explanation for these results was later revealed with TPS computer simulations[59].

In a typical experiment, an ensemble of millions of nanocrystals, each consisting of thousands of atoms, is pressurized simultaneously in a diamond anvil cell[164]. Even on today's fastest computers, simulating the dynamics of such a vast number of atoms is hopelessly out of reach and simulations focus on trajectories of single nanocrystals. An important ingredient in such a simulation is the specific way by which pressure is applied on the nanocrystal. Experimentally, the choice of pressure medium is dictated by chemical suitability and the requirement that no substantial solidification should occur up to the highest applied pressure; the solvent should deliver *hydrostatic* pressure at all times. While little is actually known about the structure and dynamics of the interface between a nanocrystal and the surrounding solvent at high pressures, different methods have been proposed that aim at modeling a perfectly hydrostatic pressure bath in molecular dynamics computer simulations[168,169]. One approach consists in the use of a modified Lennard-Jones liquid, whose parameters are chosen to avoid crystallization[170,171]. Constant pressure and temperature are then applied by means of a Nosé-Hoover barostat[52]. A different method, which leaves the equations of motion of the system unchanged, makes use of a soft sphere pressure medium whose equation of state is known[172,173]. In this method, the pressure is adjusted by simply changing the interaction parameters of the soft spheres.

A third method, proposed by the authors, makes use of an ideal gas pressure medium which is modeled in a grand-canonical approach[159,160]. Such an ideal gas of non-interacting particles guarantees ideally hydrostatic pressure, is conceptually simple and computationally cheap. Since the particles interact only with the nanocrystal, it is only necessary to follow their dynamics within close proximity of the crystal. Practically, this is achieved by modeling the stochastic flow of ideal gas particles through a surface surrounding the nanocrystal. It is the statistics of this flow, the number of particles and their velocity distribution, that determines both the temperature and pressure applied on the nanocrystal. A snapshot taken from a simulation using the ideal gas barostat is shown in Fig. 6. In all the methods described above, the interaction between pressure medium and nanocrystal can be chosen freely and must only be sufficiently repulsive to avoid diffusion of the pressure medium into the nanocrystal. It should be pointed out, however, that the details of this interaction can have a profound influence on the surface free energy of the nanocrystal and can also change the kinetics of structural transformations[165].

Straightforward molecular dynamics simulations have shown that the mechanism of the wurtzite to rocksalt transformation in CdSe nanocrystals depends strongly on the shape of the nanoparticle[9]. While spherical crystals transform directly from wurtzite to rocksalt, rod-shaped particles with low-energy surface facets transform in two steps, via a 5-coordinated intermediate structure (space group 194, sometimes referred to as *h*-MgO, BN, or stacked honeycomb structure). The study also revealed that at the strongly elevated pressures used in straightforward molecular dynamics simulations, transformations can proceed quite violently; different atomic rearrangement patterns were observed, as well as simultaneous nucleation from different sites in the crystal, and the formation of grain boundaries.

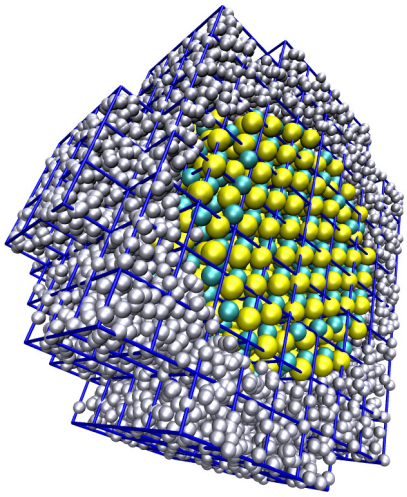


Figure 6. Cross section of a simulation box holding a CdSe nanocrystal (blue and yellow atoms) surrounded by a pressure medium of ideal gas particles (gray). The blue grid illustrates a simple implementation of the ideal gas barostat using cell lists[160].

To obtain an unbiased view of the transformation mechanism, TPS was used to study the transformation of rod-shaped particles at much lower pressures, close to experimental conditions[160]. It had been assumed that CdSe nanocrystals under a certain size would transform not through nucleation and growth, but rather through a simultaneous, concerted rearrangement of all crystal atoms[164]. Indeed, a similar mechanism was observed in a straightforward molecular dynamics simulation[9]. To investigate the prevalence of this mechanism, TPS was started with a seed trajectory showing the concerted transformation mechanism. The evolution of trajectories in this TPS simulation is illustrated in Fig. 7. Within a few hundred iterations of the shooting algorithm, a profound change of mechanism was observed. Instead of simultaneous atom rearrangements, the relaxed pathway is characterized by a continuous growth of the rocksalt structure through sequential sliding of parallel crystal planes. Further iterations of the algorithm only produced different realizations of the same mechanism. Since the frequency of occurrence of different pathways in a TPS simulation reflects their relative statistical weights in the transition path ensemble, this result strongly suggests that at experimental conditions a nucleation and growth scenario is highly favored over concerted transformation mechanisms.

The simulation of thousands of trajectories of systems consisting of many thousands of atoms requires the use of efficient interaction potentials to model materials like CdSe[174,175]. These potentials are typically optimized to reproduce a selected number of bulk material properties and their predictive power for transformations in nanocrystalline systems needs to be firmly established. This can only be achieved through a direct comparison of quantities observed in both experiment and simulation. Rate constants, and derivatives like activation energies and volumes, are quantities that closely reflect the underlying transformation mechanism and can be obtained within the transition path sampling framework.

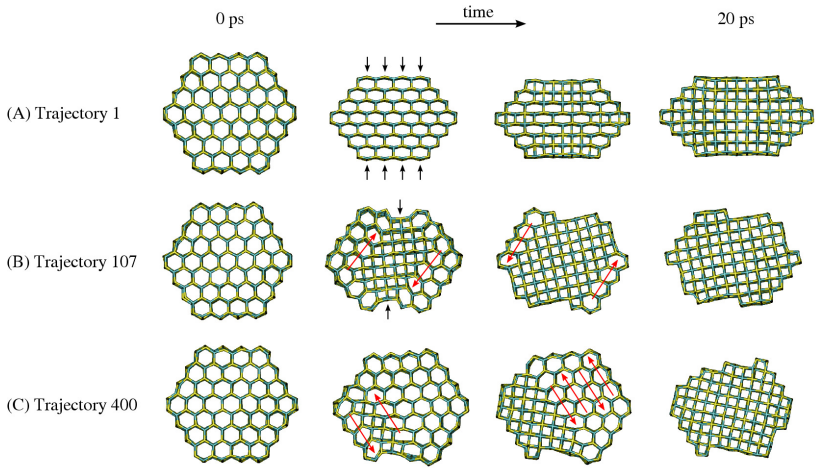


Figure 7. Relaxation of the transformation mechanism in a TPS simulation of a CdSe nanocrystal[160]. Rows (A) trough (C) show snapshots along single trajectories within the TPS run, with time evolving from left to right. At 0 ps, the crystal is in the low pressure structure, at 20 ps it is in the high pressure structure. Crystals are viewed along the wurtzite c-axis. (A) The first trajectory shows a transformation that proceeds through a concerted motion of all atoms. All unit cells are transformed and new bonds are formed almost simultaneously. (B) A hundred iterations of the shooting algorithm later, a second mechanism appears, characterized by a sequential sliding of (100) crystal planes. (C) At iteration 400, the concerted mechanism is lost and the crystal transforms through the pure sliding-planes mechanism.

Activation enthalpies and volumes, in particular, are characteristics of the central part of a transformation, the transition state which holds the critical nucleus of the high-pressure phase. Furthermore, no potentially costly free energy calculations are necessary to obtain these quantities. In a systematic TPS study, the authors have used committor analysis (as discussed in Sec. 4) to identify the transition states during the transformation of CdSe nanocrystals of different sizes[59,58]. The corresponding critical rocksalt nuclei are illustrated in Fig. 8. While classical nucleation theory predicts the occurrence of compact, roughly spherical nuclei, the presence of a surface can drastically alter this picture. The critical rocksalt nuclei in CdSe nanocrystals are elongated in shape and originate at the crystal surface. With increasing crystal size, they predominantly grow in one direction, along the long axis of the crystal. This particular size-dependence of the shape of the critical nucleus directly reflects in a linear size-dependence of activation enthalpies and volumes[59]. A comparison with experimentally determined values[62,165] showed a good qualitative agreement and thus confirmed the observed nucleation mechanism.

In summary, TPS is a suitable tool for the discovery of the atomistic mechanisms of pressure-induced structural transformations in nanocrystals. The nucleation of these transformations can be observed under experimental conditions, allowing direct contact with experiments to be established via the calculation of activation enthalpies and volumes. This can be done following the systematic procedure proposed in Ref. 58, or directly by calculating ensemble averages as discussed in Sec. 3.3.

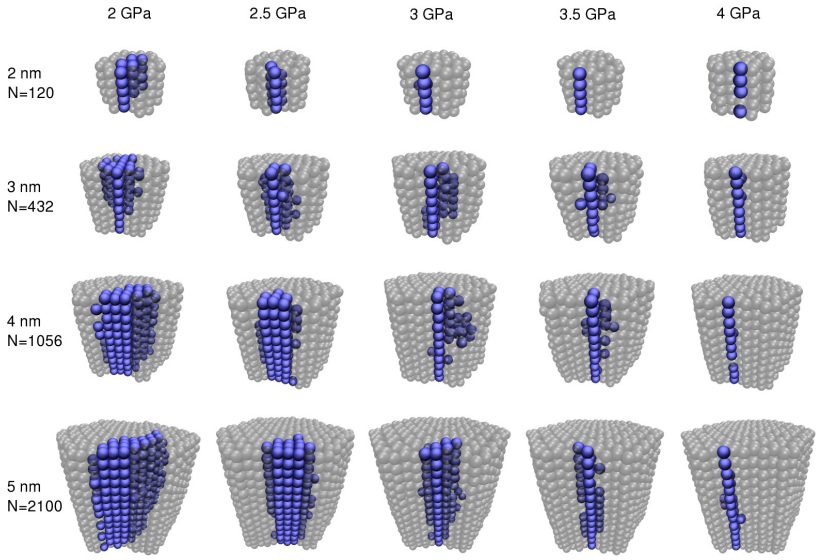


Figure 8. Typical transition states of the transformation of CdSe nanocrystals, for different crystal sizes and pressures[58]. Atoms in the rocksalt structure are blue and constitute the critical nucleus, atoms in the low-pressure structure are transparent gray. Note that all nuclei are strongly aspherical, have contact with the crystal surface and span the entire length of the crystals. With increasing crystal size, these nuclei predominantly grow in one direction.

5.2 Freezing of Supercooled Fluids

Transition path sampling methods have also proven very valuable in computational investigations of the freezing transition of supercooled liquids. Recently, interest in this particularly fundamental phase transition has been revived, sparked by freezing experiments carried out on colloidal suspensions[176]. Taking advantage of modern optical microscopy such systems can be studied in great detail with a time and space resolution that permits to follow the motion of individual particles as the nucleation and growth process occurs. From a computational point of view, colloidal particles can often be modeled accurately using soft sphere interactions such as the Lennard-Jones[177] or Gaussian pair potentials[92]. Based on these interactions, large systems of soft colloidal particles can be simulated for long times to study basic condensed matter phenomena ranging from the glass transitions, gas-liquid-coexistence or the effective interactions of polymer coils[178]. In such colloidal systems crystallization is a rare event on the time scale of the basic particle motions. Thus, in straightforward simulations as well as in experiments the time spent waiting for the transition is long in comparison to the transition time. The application of TPS allows to restrict the computational effort to the fraction of time when freezing actually takes place. In this section we review recent transition path sampling results on the mechanism of the crystallization of a supercooled Lennard-Jones fluid. In doing that, we place the emphasis on the importance of the particular crystalline structures found in the crystalline nucleus, both for homogeneous crystallization in the bulk as well as for heterogeneous crystallization near a small crystalline template. The crystallization of a Lennard-Jones liquid has been

studied recently, using TIS, by Moroni and coworkers[90] , who confirmed previous findings such as Ostwald's step rule[179,180], which states that the undercooled liquid transforms first into a state closest in free energy to the initial state, even if other states are thermodynamically most stable. Their results also provide new insights into the nature of the transition pointing to the importance of the shape and structure of the formed crystallites[90,91].

Transition path sampling techniques allow to concentrate the analysis on the transition itself without making any a priori assumptions about the way it takes place. In conventional methods like umbrella sampling, the progress of the transition is monitored by the value of a postulated order parameter (or several order parameters), and the system is forced to follow a path on which the order parameter increases. In contrast, TPS requires only the definition of the initial and the final states and allows the system to evolve freely between them. In order to define the initial and final states for the freezing transition (and to analyze the harvested pathways) it is necessary to be able to determine local crystal structures. Individual particles are identified as crystalline on the basis of the Steinhardt order parameters[181] following the scheme proposed by ten Wolde and coworkers[177]. For each particle, the structure of the local environment is first described by a complex vector composed by the spherical harmonics of the bonds between the particle and its neighbors. The crystallinity of a particle is then defined by the degree of correlations between the local structures surrounding the particle itself and its neighbors. A cluster analysis, in which neighboring particles with the same local structure are considered to belong to the same cluster, finally groups the crystalline particles in clusters surrounded by liquid particles. The size of the largest of these crystalline clusters is used as the order parameter to define the initial and final region.

Recent results obtained from TIS simulations[7,90,91] showed that the size of the largest solid cluster is insufficient for an accurate description of the crystallization process, and that the structure of the clusters also plays an important role. These findings triggered a renewed discussion about the best definition of the crystallinity order parameter. Currently, there are several approaches for the determination of the crystalline order. Schilling and coworkers[182] proposed to look at the strength of the correlation between the local structures around a particle and its neighbors to define the degree of ordering. Lechner and coworkers[92] introduced a criterion to identify solid particles based on spatially averaged bond order parameters[183]. Kawasaki and Tanaka[184] combined these averaged bond order parameters with temporal averaging.

Although the definitions of the crystallization order parameters vary, most of the new approaches agree on a two-step freezing mechanism. The new parameters allow to identify a pre-ordered phase in the first stage of the transition before the actual crystallization takes place. In a recent investigation of this mechanism using the Gaussian core model, Lechner and coworkers[92,93] applied a non-linear generalization[75] of the likelihood maximization approach[76] to identify the best reaction coordinate and showed that the transition is best described in terms of the number of particles in the largest solid cluster and of the averaged order parameters. These spatially averaged order parameters present more stringent conditions to the local ordering of the surroundings of a particle than the Steinhardt bond order parameters. Thus, the degree of ordering of the crystallites appears to be an important factor which was not included in the traditional definition of crystallinity. The main conclusion of this work[92,93] was that the growing crystalline cluster is embedded in a cloud of pre-structured surface particles that are highly correlated but not manifestly crystalline yet.

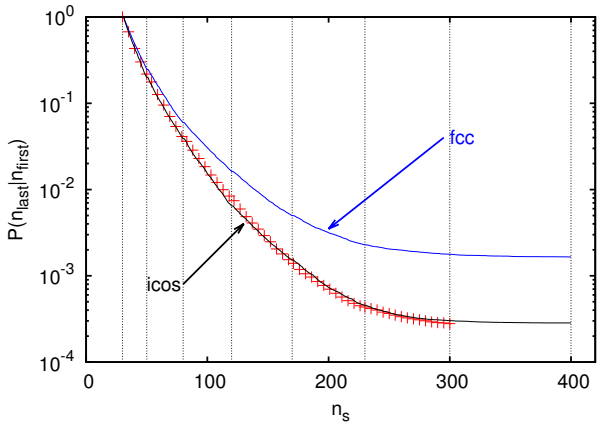


Figure 9. Crossing probabilities as a function of the largest cluster size in the presence of a face-centered cubic (blue line), an icosahedral (black line) and without a seed (red crosses), vertical lines indicate the positions of the TIS interfaces. The size of the cluster, n_s, is identified within the scheme proposed by ten Wolde and coworkers[177] (Figure adapted from Ref. 94.)

The role of the crystalline structure was also demonstrated by another recent study[94], that considered heterogeneous nucleation of an undercooled Lennard-Jones fluid near small crystalline seeds of different structures but equal size. As explained in Sec. 3.2, within the TIS method, the crystallization rate is expressed as a product of the flux factor, or the rate (per volume) of leaving the initial state, and the probability to reach the final state. To improve the sampling of the last term, the region between the initial and the final state is divided into smaller portions by introducing interfaces, as indicated by the vertical lines in Fig. 9. In this case, the interfaces have been defined using the size n_s of the largest crystalline cluster as the order parameter. On a particular interface, all configurations have the same value of the order parameter, i.e., they all contain a crystalline cluster of the same size. According to Eq. 33, the total crossing probability is written as the product of crossing probabilities between neighboring interfaces, which are much easier to calculate in separate path sampling simulations. Figure 9 displays the crossing probability accumulated up to interface j,

$$P_A(\lambda_j|\lambda_1) = \prod_{i=1}^{j-1} P_A(\lambda_{i+1}|\lambda_i). \tag{35}$$

When the transition state has been crossed, this crossing probability saturates, because essentially all trajectories that have made it beyond the transition state eventually reach the final region and the local crossing probabilities become unity. In other words, a crystallite that has reached critical size will continue to grow leading to the crystallization of the entire system. For this reason, the sampling does not have to be performed for the whole region between the initial and the final states, but should be concentrated on the interfaces for which $P_A(\lambda_{i+1}|\lambda_i)$ differs from unity[71].

The crossing probabilities, and hence the nucleation rates, of a fluid are strongly af-

fected by the presence of small templates. Figure 9 shows the crossing probabilities obtained for systems with small crystalline seeds of 13 particles kept at fixed positions and also includes the respective result for the bulk system without a seed. Two situations are considered: a seed with face-centered cubic (fcc) structure, which corresponds to the stable bulk structure, and a seed with icosahedral structure, incommensurate with the preferred bulk structure (see Fig. 10). As can be seen in Fig. 9, the conditional probabilities to reach a fully crystallized state differ for the two structures by one order of magnitude (corresponding flux factors are comparable), although the number of fixed particles and the radius of the seeds are similar[94]. In the presence of the incommensurate icosahedral structure the crystallization rate ($J = (2.6 \pm 0.6) \times 10^{-8}$, in reduced Lennard-Jones units) is similar to the homogeneous rate ($J = (2.5 \pm 0.6) \times 10^{-8}$). For the commensurate seed with fcc order the crystallization rate is one order of magnitude larger ($J = (1.4 \pm 0.2) \times 10^{-7}$).

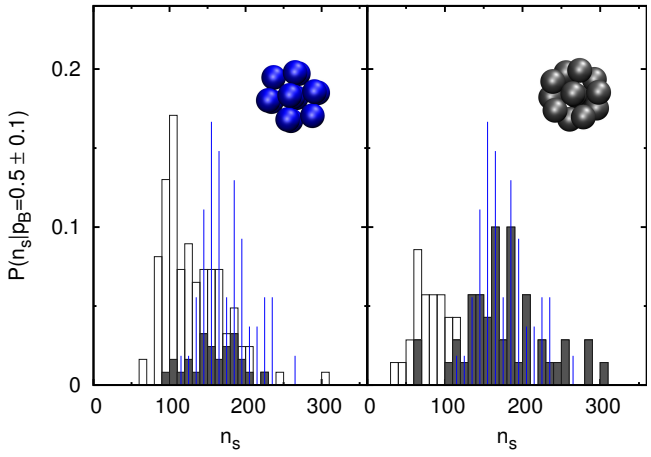


Figure 10. Cluster size distributions in the transition state ensemble in the presence of a face-centered cubic (left) and an icosahedral (right) seed. Vertical lines indicate the results for homogeneous nucleation. The size of the cluster, n_s, is identified within the scheme proposed by ten Wolde and coworkers[177]. The dark filling of the histogram bins indicate the fraction of configurations in which the crystalline cluster is separated from the seed. (Figure adapted from Ref. 94.)

A committor analysis of configurations taken from crystallization pathways obtained from the TIS simulations allows to explain the reason for the observed enhancement of the nucleation rate caused by the fcc seed. The transition state ensemble of the transition, consisting of configurations with probability of $p_B = 0.5$ to end in the final state, can be divided into two distinct classes – with the crystalline clusters formed close to the seed and far away from it. Figure 10 shows the projection of the transition state ensemble on the standard reaction coordinate, the number of solid particles in the largest cluster. For a system with an icosahedral seed the distribution is clearly double-peaked, and also for a system with an fcc seed the wide shoulder in the distribution indicates the onset of a second peak. This peak coincides with the cluster size distribution found for the homogeneous crystallization as well as with the distributions found if only configurations in which the crystalline cluster is separated from the seed are considered for both heterogeneous systems. Thus, two different crystallization pathways can be identified and their relative im-

portance is determined by the structure of the seed. For an icosahedral seed, the crystalline clusters are typically formed away from the seed, and the bulk crystallization scenario is recovered. In contrast, the crystals mostly attach to the seed with a face-centered cubic structure and the crystallization observed is heterogeneous.

In summary, the application of the transition path sampling to the crystallization of soft spheres contributed to a better understanding of the process and highlighted the importance of the crystalline structures for the description of the transition.

6 Conclusion and Outlook

As discussed in the previous sections of this article, transition path sampling is a generally applicable and efficient computational methodology to study nucleation and growth occurring at first order phase transitions. By concentrating on the rare barrier crossing events, this method circumvents the computational problem of widely disparate time scales enabling the collection of many, possibly very different transition pathways. Transition path sampling can be used to calculate transition rate constants and, perhaps more importantly, to identify reaction coordinates, i.e., variables that describe the progress of the transition and provide a handle to probe or even control the transition in experiments. To date, transition path sampling has been applied to many nucleation processes yielding new insights but also pointing out open problems. For instance, it would be very useful to have a transition path sampling algorithm for the calculation of transition rate constants that does not rely on the definition of an order parameter to transform the ensemble of free trajectories into that of reactive trajectories. An approach in the spirit of thermodynamic integration as outlined in Sec. 3.3 might lead to an advance in this respect. Other examples of worthwhile enhancement of the TPS method include the development of efficient algorithms for ergodic sampling as well as more systematic procedures to find reaction coordinates and infer the transition mechanism. Further research is required to bring about these and other advances, that can be used to improve our understanding of nucleation and other rare event processes occurring in materials hard and soft.

Acknowledgments

We acknowledge financial support from Austrian Science Fund (FWF) within the SFB ViCoM (F 41) and project P20942-N16. MG was supported by the Austrian Science Fund (FWF) under Grant No. J 3106-N16.

References

1. G. Antczak and G. Ehrlich, *Surf. Sci. Rep.* **62**, 39 (2007).
2. H. Jónsson, *Proc. Natl. Acad. Sci. USA* **108**, 944 (2011).
3. P. J. Feibelman, *Phys. Rev. Lett.* **65**, 729 (1990).
4. R. Becker and W. Döring, *Ann. Phys.* **24**, 719 (1935).
5. M. Volmer, *Kinetik der Phasenbildung*, Steinkopff: Dresden, Deutschland (1939).
6. P. G. Debenedetti, *Metastable Liquids: Concepts and Principles*, Princeton University Press, Princeton (1996).

7. S. Jungblut and C. Dellago, *J. Chem. Phys.* **134**, 104501 (2011).
8. R. Martonak, L. Colombo, C. Molteni, and M. Parrinello, *J. Chem. Phys.* **117**, 11329 (2002).
9. M. Grünwald, E. Rabani, and C. Dellago, *Phys. Rev. Lett.* **96**, 255701 (2006).
10. A. Laio and M. Parrinello, *Proc. Natl. Acad. Sci. USA* **99**, 12562 (2002).
11. A. Laio and M. Parrinello, in *Computer Simulations in Condensed Matter: from Materials to Chemical Biology*, edited by M. Ferrario, G. Ciccotti, and K. Binder, Springer, Berlin (2006).
12. I. G. Kevrikidis, C. W. Gear, J. M. Hyman, P. G. Kevrekidis, O. Runborg, and K. Theodoropoulos, *Comm. Math. Sciences* **14**, 715 (2003).
13. S. Sriraman, I. G. Kevrekidis, and G. Hummer, *Phys. Rev. Lett.* **95**, 130603 (2005).
14. A. F. Voter, F. Montalenti, and T. C. Germann, *Annu. Rev. Mater. Res.* **32**, 321 (2002).
15. M.R. Sørensen and A. F. Voter, *J. Chem. Phys.* **112**, 9599 (2000).
16. C. Dellago, P. G. Bolhuis, F. S. Csajka, and D. Chandler, *J. Chem. Phys.* **108**, 1964 (1998).
17. C. Dellago, P. G. Bolhuis, and D. Chandler, *J. Chem. Phys.* **108**, 9263 (1998).
18. P. G. Bolhuis, C. Dellago, and D. Chandler, *Faraday Discuss.* **110**, 421 (1998).
19. C. Dellago, P. G. Bolhuis, F. S. Csajka, and D. Chandler, *J. Chem. Phys.* **110**, 6617 (1999).
20. P. L. Geissler, C. Dellago, and D. Chandler, *J. Phys. Chem. B* **103**, 3706 (1999).
21. P. G. Bolhuis, C. Dellago, and D. Chandler, *Proc. Natl. Acad. Sci. USA* **97**, 5877 (2000).
22. P. L. Geissler, C. Dellago, and D. Chandler, *Phys. Chem. Chem. Phys.* **1**, 1317 (1999).
23. D. Laria, J. Rodriguez, C. Dellago, and D. Chandler, *J. Phys. Chem. A* **105**, 2646 (2001).
24. C. Dellago and P. G. Bolhuis, *Mol. Sim.* **30**, 795 (2004).
25. T. S. van Erp, D. Moroni, and P. G. Bolhuis, *J. Chem. Phys.* **118**, 7762 (2003).
26. D. Moroni, P. G. Bolhuis, and T. S. van Erp, *J. Chem. Phys.* **120**, 4055 (2004).
27. D. Moroni, T. S. van Erp, and P. G. Bolhuis, *Phys. Rev. E* **71**, 056709 (2005).
28. E. E. Borrero and C. Dellago, *J. Chem. Phys.* **133**, 134112 (2010).
29. C. Dellago, P. G. Bolhuis, and P. L. Geissler, p. 349, in *Computer Simulations in Condensed Matter: from Materials to Chemical Biology*, edited by M. Ferrario, G. Ciccotti, and K. Binder, Springer Lecture Notes in Physics (2006).
30. P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler, *Ann. Rev. Phys. Chem.* **53**, 291 (2002).
31. C. Dellago, P. G. Bolhuis, and P. L. Geissler, *Adv. Chem. Phys.* **123**, 1 (2002).
32. C. Dellago and D. Chandler, in *Molecular Simulation for the Next Decade*, p. 321, edited by P. Nielaba, M. Mareschal, and G. Ciccotti, Springer, Berlin (2002).
33. C. Dellago, in *Handbook of Materials Modeling*, p. 1585, edited by S. Yip, Springer, Berlin (2005).
34. C. Dellago, in *Free energy calculations: Theory and applications in chemistry and biology*, edited by A. Pohorille and C. Chipot, Springer, Berlin (2007).
35. C. Dellago and P. G. Bolhuis, *Topics in Current Chemistry* **268**, 291, edited by M. Reiher, Springer (2007).
36. C. Dellago and P. G. Bolhuis, *Adv. Poly. Sci.* **221**, 167 (2008).

37. A. F. Voter, *J. Chem. Phys.* **106**, 4665 (1997).
38. A. F. Voter, *Phys. Rev. Lett.* **78**, 3908 (1997).
39. A. F. Voter, *Phys. Rev. B* **57**, 13985 (1998).
40. B. P. Uberuaga, S. J. Stuart, and A. F. Voter, *Phys. Rev. B* **75**, 014301 (2007).
41. O. Kum, B. M. Dickson, S. J. Stuart, B. P. Uberuaga, and A. F. Voter, *J. Chem. Phys.* **121**, 9808 (2004).
42. W. E and E. Vanden-Eijnden, *J. Stat. Phys.* **123**, 503 (2006).
43. E. Vanden-Eijnden, in *Computer Simulations in Condensed Matter: from Materials to Chemical Biology*, p. 453, edited by M. Ferrario, G. Ciccotti, and K. Binder, Springer Lecture Notes in Physics (2006).
44. W. E, W. Ren, and E. Vanden-Eijnden, *Chem. Phys. Lett.* **413**, 242 (2005).
45. R. J. Allen, P. B. Warren, and P. R. ten Wolde, *Phys. Rev. Lett.* **94**, 018104 (2005).
46. R. J. Allen, D. Frenkel, and P. R. ten Wolde, *J. Chem. Phys.* **124**, 024102 (2006).
47. R. J. Allen, D. Frenkel, and P. R. ten Wolde, *J. Chem. Phys.* **124**, 194111 (2006).
48. C. Valeriani, R. J. Allen, M. J. Morelli, D. Frenkel, and P. R. ten Wolde, *J. Chem. Phys.* **127**, 114109 (2007).
49. E. E. Borrero and F. A. Escobedo, *J. Chem. Phys.* **125**, 164904 (2006).
50. D. Chandler, *Introduction to Modern Statistical Mechanics*, Oxford University Press (1987).
51. G. E. Crooks and D. Chandler, *Phys. Rev. E* **64**, 026109 (2001).
52. D. Frenkel and B. Smit, *Understanding Molecular Simulation*, Academic Press (1996).
53. N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, *J. Chem. Phys.* **21**, 1087 (1953).
54. C. J. Geyer and E. A. Thompson, *J. Am. Stat. Soc.* **90**, 909 (1995).
55. T. J. H. Vlugt and B. Smit, *Phys. Chem. Comm.* **2**, 1 (2001).
56. P. G. Bolhuis, *J. Phys.: Condens. Matter* **15**, S113 (2003).
57. M. Grünwald, P. L. Geissler, and C. Dellago, *J. Chem. Phys.* **129**, 194101 (2008).
58. M. Grünwald and C. Dellago, *J. Chem. Phys.* **131**, 164116 (2009).
59. M. Grünwald and C. Dellago, *Nano Lett.* **9**, 2099 (2009).
60. H. Zheng, J. B. Rivest, T. A. Miller, B. Sadtler, A. Lindenberg, M. F. Toney, L.-W. Wang, C. Kisielowski, and A. P. Alivisatos, *Science* **333**, 206 (2011).
61. J. B. Rivest, L. K. Fong, P. K. Jain, M. F. Toney, and A. P. Alivisatos, *J. Phys. Chem. Lett.* **2**, 2402 (2011).
62. K. Jacobs, D. Zaziski, E. C. Scher, A. B. Herold, and A. P. Alivisatos, *Science* **293**, 5536 (2001).
63. P. Stöckel, I. M. Weidinger, H. Baumgärtel, and T. Leisner, *J. Phys. Chem. A* **109**, 2540 (2005).
64. M. Matsumoto, S. Saito, and I. Ohmine, *Nature* **416**, 409 (2002).
65. H. Matsubara, T. Koishi, T. Ebisuzaki, and K. Yasuoka, *J. Chem. Phys.* **127**, 214507 (2007).
66. C. Dellago and P. L. Geissler, in Proceedings of *The Monte Carlo Method in the Physical Sciences: Celebrating the 50th anniversary of the Metropolis algorithm*, AIP Conference Proceedings, vol. **690** (2003).
67. P. L. Geissler and C. Dellago, *J. Phys. Chem. B* **108**, 6667 (2004).
68. G. M. Torrie and J. P. Valleau, *J. Comp. Phys.* **23**, 187 (1977).

69. T. S. van Erp and P. G. Bolhuis, *J. Comp. Phys.* **205**, 157 (2005).
70. T. S. van Erp, *Phys. Rev. Lett.* **98**, 268301 (2007).
71. E. E. Borrero, M. Weinwurm, and C. Dellago, *J. Chem. Phys.* **134**, 244118 (2011).
72. T. S. van Erp, *Phys. Rev. Lett.* **98**, 268301 (2007).
73. P. G. Bolhuis, *J. Chem. Phys.* **129**, 114108 (2008).
74. J. Rogal, W. Lechner, J. Juraszek, B. Ensing, and P. G. Bolhuis, *J. Chem. Phys.* **133**, 174109 (2010).
75. W. Lechner, J. Rogal, J. Juraszek, B. Ensing, and P. G. Bolhuis, *J. Chem. Phys.* **133**, 174110 (2010).
76. B. Peters and B. L. Trout, *J. Chem. Phys.* **125**, 054108 (2006).
77. J. Kirkwood, *J. Chem. Phys.* **3**, 300 (1935).
78. L. Onsager, *Phys. Rev.* **54**, 554 (1938).
79. R. Du, V. Pande, A. Y. Grosberg, T. Tanaka, and E. I. Shakhnovich, *J. Chem. Phys.* **108**, 334 (1998).
80. D. Ryter, *Physica A* **142**, 103 (1987).
81. A. Berezhovski and A. Szabo, *J. Chem. Phys.* **125**, 104902 (2006).
82. M. M. Klosek, B. J. Matkowsky, and Z. Schuss, *Ber. Bunsenges. Phys. Chem.* **95**, 331 (1991).
83. E. Pollak, A. M. Berezhkovskii, and Z. Schuss, *J. Chem. Phys.* **100**, 334 (1994).
84. P. Talkner, *Chem. Phys.* **180**, 199 (1994).
85. J. D. Honeycutt and H. C. Andersen, *Chem. Phys. Lett.* **108**, 535 (1984).
86. J. D. Honeycutt and H. C. Andersen, *J. Phys. Chem.* **90**, 1585 (1986).
87. A. Ma and A. R. Dinner, *J. Phys. Chem. B* **109**, 6769 (2005).
88. G. Hummer, *J. Chem. Phys.* **120**, 516 (2004).
89. R. B. Best and G. Hummer, *Proc. Natl. Acad. Sci. USA* **102**, 6732 (2005).
90. D. Moroni, P. R. ten Wolde, and P. G. Bolhuis, *Phys. Rev. Lett.* **94**, 235703 (2005).
91. B. Peters, *J. Phys. Chem. Lett.* **2**, 1133 (2011).
92. W. Lechner, C. Dellago, and P. G. Bolhuis, *Phys. Rev. Lett.* **106**, 085701 (2011).
93. W. Lechner, C. Dellago, and P. G. Bolhuis, *J. Chem. Phys.* **135**, 154110 (2011).
94. S. Jungblut and C. Dellago, *Europhys. Lett.* **96**, 56006 (2011).
95. B. Peters, *J. Chem. Phys.* **125**, 241101 (2006).
96. A. C. Pan and D. Chandler, *J. Phys. Chem. B* **108**, 19681 (2004).
97. G. T. Beckham, B. Peters, C. Starbuck, N. Variankaval, and B. L. Trout, *J. Am. Chem. Soc.* **129**, 4714 (2007).
98. S. L. Quaytman and S. D. Schwartz, *Proc. Natl. Acad. Sci. USA* **104**, 12253 (2007).
99. S.-S. So and M. Karplus, *J. Med. Chem.* **39**, 1521 (1996).
100. A. Dinner, S.-S. So, and M. Karplus, *Adv. Chem. Phys.* **120**, 1 (2002).
101. B. Peters, G. T. Beckham, and B. L. Trout, *J. Chem. Phys.* **127**, 034109 (2007).
102. A. W. F. Edwards, *Likelihood*, Cambridge University Press, Cambridge (1972).
103. P. L. Geissler, C. Dellago, D. Chandler, J. Hutter, and M. Parrinello, *Science* **291**, 2121 (2001).
104. T. J. F. Day, U. W. Schmitt, and G. A. Voth, *J. Am. Chem. Soc.* **122**, 12027 (2000).
105. J. M. Park, A. Laio, M. Iannuzzi, and M. Parrinello, *J. Am. Chem. Soc.* **128**, 11318 (2006).
106. C. S. Lo, R. Radhakrishnan, and B. L. Trout, *Catalysis Today* **105**, 93 (2005).
107. D. Zahn, *Chem. Phys.* **300**, 79 (2004).

108. B. Ensing and E. J. Baerends, *J. Phys. Chem. A* **106**, 7902 (2002).
109. M. Pagliai, S. Raugei, G. Cardini, and V. Schettino, *J. Chem. Phys.* **117**, 2199 (2002).
110. J. Ramirez and M. Laso, *J. Chem. Phys.* **115**, 7285 (2001).
111. T. Bucko, L. Benco, O. Dubay, C. Dellago, and J. Hafner, *J. Chem. Phys.* **131**, 214508 (2009).
112. J. E. Basner and S. D. Schwartz, *J. Am. Chem. Soc.* **127**, 13822 (2005).
113. J. Marti, F. S. Csajka, and D. Chandler, *Chem. Phys. Lett.* **328**, 169 (2000).
114. J. Marti and F. S. Csajka, *J. Chem. Phys.* **113**, 1154 (2000).
115. J. Marti, *Mol. Sim.* **27**, 169 (2001).
116. P. L. Geissler and D. Chandler, *J. Chem. Phys.* **113**, 9759 (2000).
117. J. Y. Lee, *Chem. Phys.* **299**, 123 (2004).
118. F. S. Csajka and D. Chandler, *J. Chem. Phys.* **109**, 1125 (1998).
119. P. G. Bolhuis and D. Chandler, *J. Chem. Phys.* **113**, 8154 (2000).
120. M. Athènes, *Eur. Phys. J. B* **38**, 651 (2004).
121. M. Athènes and M. Marinica, *J. Comp. Phys.* **229**, 7129 (2010).
122. M. Merolle, J. P. Garrahan, and D. Chandler, *Proc. Natl. Acad. Sci. USA* **102**, 10837 (2005).
123. L.O. Hedges, R.L. Jack, J.P. Garrahan, and D. Chandler, *Science* **323**, 1309 (2009).
124. Y.S. Elmatad, R.L. Jack, D. Chandler, and J.P. Garrahan, *Proc. Natl. Acad. Sci. USA* **107**, 12793 (2010).
125. T. J. H. Vlugt, C. Dellago, and B. Smit, *J. Chem. Phys.* **113**, 8791 (2000).
126. T. A. McCormick and D. Chandler, *J. Phys. Chem. B* **107**, 2796 (2003).
127. R. Crehuet and M. J. Field, *J. Phys. Chem. B* **111**, 5708 (2007).
128. P. G. Bolhuis, *Proc. Natl. Acad. Sci. USA* **100**, 12129 (2003).
129. P. G. Bolhuis, *Biophys. J.* **88**, 50 (2005).
130. J. Juraszek and P. G. Bolhuis, *Proc. Natl. Acad. Sci. USA* **103**, 15859 (2006).
131. M. F. Hagan, A. R. Dinner, D. Chandler, and A. K. Chakraborty, *Proc. Natl. Acad. Sci. USA* **100**, 13922 (2003).
132. R. Radhakrishnan and T. Schlick, *Proc. Natl. Acad. Sci. USA* **101**, 5970 (2004).
133. R. Radhakrishnan, L. J. Yang, K. Arora, and T. Schlick, *Biophys. J.* **86**, 34A (2004).
134. R. Radhakrishnan and T. Schlick, *J. Am. Chem. Soc.* **127**, 13245 (2005).
135. Y. L. Wang and T. Schlick, *BMC Struct. Biol.* **7**, 7 (2007).
136. J. Marti and F. S. Csajka, *Phys. Rev. E* **69**, 061918 (2004).
137. J. Marti, *J. Phys.: Condens. Matter* **16**, 5669 (2004).
138. P. Geiger and C. Dellago, *Chem. Phys.* **375**, 309 (2010).
139. S. X. Sun, *J. Chem. Phys.* **118**, 5769 (2003).
140. F. M. Ytreberg and D. M. Zuckerman, *J. Chem. Phys.* **120**, 10876 (2004).
141. H. Oberhofer, C. Dellago, and P. L. Geissler, *J. Phys. Chem. B* **69**, 6902 (2005).
142. W. Lechner and C. Dellago, *J. Stat. Mech.*, P04001 (2007).
143. H. Oberhofer and C. Dellago, *Comp. Phys. Comm.* **179**, 41 (2008).
144. A. Imparato and L. Peliti, *J. Stat. Mech.*, L02001 (2007).
145. A. Imparato and L. Peliti, *Compt. Rend. Physique* **8**, 556 (2007).
146. M. Grünwald, P. L. Geissler, and C. Dellago, *J. Chem. Phys.* **127**, 154718 (2007).
147. D. Zahn, *Phys. Rev. Lett.* **92**, 040801 (2004).
148. E. Schöll-Paschinger and C. Dellago, *J. Chem. Phys.* **133**, 104505 (2010).
149. D. Zahn, *J. Phys. Chem. B* **111**, 5249 (2007).

150. D. Zahn, *J. Phys. Chem. B* **110**, 19601 (2006).
151. D. Zahn, Y. Grin, and S. Leoni, *Phys. Rev. B* **72**, 064110 (2005).
152. D. Zahn, *Phys. Rev. Lett.* **93**, 227801 (2004).
153. D. Zahn and S. Leoni, *J. Phys. Chem. B* **110**, 10873 (2006).
154. D. Zahn, O. Hochrein, and S. Leoni, *Phys. Rev. B* **72** (9), 094106 (2005).
155. D. Zahn, *J. Solid State Chem.* **177**, 3590 (2004).
156. S. Leoni and D. Zahn, *Z. Kristall.* **219**, 339 (2004).
157. D. Zahn and S. Leoni, *Z. Kristall.* **219**, 345 (2004).
158. D. Zahn and S. Leoni, *Phys. Rev. Lett.* **92**, 250201 (2004).
159. M. Grünwald and C. Dellago, *Mol. Phys.* **104**, 3709 (2006).
160. C. Dellago and M. Grünwald, *Chall. Adv. Comp. Chem. Phys.* **9**, 61 (2010).
161. D. Gottwald, G. Kahl, and C. N. Likos, *J. Chem. Phys.* **122**, 204503 (2005).
162. A. R. Oganov and C. W. Glass, *J. Chem. Phys.* **124**, 244704 (2006).
163. S. M. Woodley and R. Catlow, *Nat. Mater.* **7**, 937 (2008).
164. S. H. Tolbert and A. P. Alivisatos, *J. Chem. Phys.* **102**, 4642 (1995).
165. C.-C. Chen, A. B. Herhold, C. S. Johnson, and A. P. Alivisatos, *Science* **276**, 398 (1997).
166. J. N. Wickham, A. B. Herhold, and A. P. Alivisatos, *Phys. Rev. Lett.* **84**, 923 (2000).
167. K. Jacobs, J. Wickham, and A. P. Alivisatos, *J. Phys. Chem. B* **106**, 3759 (2002).
168. S. E. Baltazar, A. H. Romero, J. K. Rodríguez-López, H. Terrones, and R. Martoňák, *Comput. Mater. Sci.* **37**, 526 (2006).
169. R. Martoňák, *Euro. Phys. J. B* **79**, 241 (2011).
170. B. J. Morgan and P. A. Madden, *Nano Lett.* **4**, 1581 (2004).
171. B. J. Morgan and P. A. Madden, *Phys. Chem. Chem. Phys.* **8**, 3304 (2006).
172. R. Martoňák, C. Molteni, and M. Parrinello, *Phys. Rev. Lett.* **84**, 682 (2000).
173. R. Martoňák, L. Colombo, C. Molteni, and M. Parrinello, *J. Chem. Phys.* **117**, 11329 (2002).
174. E. Rabani, *J. Chem. Phys.* **116**, 258 (2002).
175. P. Zapol, R. Pandey, and J. D. Gale, *J. Phys.: Condens. Matter* **9**, 9517 (1997).
176. U. Gasser, E. R. Weeks, A. Schofield, P. N. Pusey, and D. A. Weitz, *Science* **292**, 258 (2001).
177. P. R. ten Wolde, M. J. Ruiz-Montero, and D. Frenkel, *J. Chem. Phys.* **104**, 9932 (1996).
178. C. N. Likos, *Phys. Rep.* **348**, 267 (2001).
179. W. Ostwald, *Z. Phys. Chem.* **22**, 289 (1897).
180. P. R. ten Wolde and D. Frenkel, *Phys. Chem. Chem. Phys.* **1**, 2191 (1999).
181. P. J. Steinhardt, D. R. Nelson, and M. Ronchetti, *Phys. Rev. B* **28**, 784 (1983).
182. T. Schilling, H. J. Schöpe, M. Oettel, G. Opletal, and I. Snook, *Phys. Rev. Lett.* **105**, 025701 (2010).
183. W. Lechner and C. Dellago, *J. Chem. Phys.* **129**, 114707 (2008).
184. T. Kawasaki and H. Tanaka, *J. Phys.: Condens. Matter* **22**, 232102 (2010).

Neural Network Potentials for Efficient Large-Scale Molecular Dynamics

Jörg Behler

Lehrstuhl für Theoretische Chemie
Ruhr-Universität Bochum, 44780 Bochum, Germany
E-mail: joerg.behler@theochem.ruhr-uni-bochum.de

The availability of efficient interatomic potentials is a necessary prerequisite for molecular dynamics studies of large systems. An overwhelming number of potentials has been suggested in the literature, from simple classical force fields to sophisticated potentials for applications in materials science. Still, there is no type of potential, which allows for a unified description of systems as different as covalent molecules, possibly interacting via weak van der Waals forces, semiconductors and metals. Artificial neural networks are a promising new tool to construct accurate potential-energy surfaces (PESs) for a variety of systems on an equal footing. They are very flexible, they do not require any knowledge about the functional form of the potential, and they can be constructed using high-level first principles methods. In this lecture the basic properties of different types of neural network potentials are introduced and the current scope and limitations of the method are discussed.

1 Introduction

In principle, the Hamiltonian of a system is fully defined by the positions of the nuclei $\{\mathbf{R}_i\}$, the nuclear charges $\{Z_i\}$, and the total charge of the system Q. Unfortunately, solving exactly the Schrödinger equation employing the resulting Hamiltonian is impossible for all but the most simple systems. Therefore, a full hierarchy of methods has been developed by using different of levels of approximation, and to date many efficient first-principles methods have become available, like e.g. Hartree Fock (HF) theory constructing the wave function from a single Slater determinant only, or the evaluation of the electronic exchange and correlation energy by an approximate functional in density-functional theory (DFT)[1,2]. In particular the combination of DFT with molecular dynamics, termed *ab initio* MD[3,4], has become a standard tool in theoretical chemistry due to its high computational efficiency.

Still, many problems in chemistry and physics cannot be addressed by these methods, because the large number of atoms and/or the required simulation times result in a prohibitively large amount of computing time. Therefore, a huge number of more efficient, but also more approximate, empirical potentials has been suggested in the literature. In principle, these potentials construct a direct functional relation between the atomic positions and the potential-energy of a system. The resulting "potential-energy surface" (PES) is in general a high-dimensional function yielding the total energy E if the coordinates of the atoms are provided in a suitable form. Apart from the total energy, the PES contains also information on the forces F, which are the negative derivatives of the energy with respect to some coordinate $R_{i,\alpha}$ of atom i in direction $\alpha = \{x, y, z\}$,

$$F_{i,\alpha} = -\frac{\partial E}{\partial R_{i,\alpha}} \quad . \tag{1}$$

In principle the shape of the PES determines also all higher derivatives.

Potentials for atomistic simulations can be classified into two different types. In "physical potentials", a physically reasonable functional form is chosen, and a rather small number of parameters is adjusted to reproduce a given set of first-principles data and/or experimentally observed properties. The approximations made and the rather low flexibility often prevent an equally accurate description of a wide range of properties, but the underlying functional form ensures that an overall reasonable behavior is found in most applications. Examples for such potentials are classical force fields[5–9] for organic molecules, bond-order based potentials like the Tersoff potential[10,11] for semiconductors, and the embedded atom method[12,13] for metals.

In "mathematical potentials", on the other hand, very general and highly flexible functions are used, which do not contain any constraints on the physical properties of the systems and which are usually fitted to first-principles data. They often include a very large number of fitting parameters, and a high numerical accuracy can be achieved, but the fitting process has to be done with great care to ensure that no unphysical results are obtained. Examples are splines[14], interpolating moving least squares (IMLS)[15,16], modified Sheppard interpolation (MSI) based on a Taylor expansion[17,18], genetic programming[19], and Gaussian approximation potentials[20].

Unfortunately, none of these potentials is equally accurate for all types of systems and consequently the development of reliable potentials with a wider range of applicability is still a very active field of research. The "perfect potential" should fulfill the following ten criteria[21]:

1. It should be very accurate.
2. It should be possible to improve the potential systematically.
3. It should be general and applicable to all types of systems.
4. It should be able to describe the making and breaking of bonds.
5. It should be applicable to large systems.
6. It should not require much human effort to construct the potential.
7. It should be transferable.
8. It should be fast to evaluate.
9. It should not require much CPU time to construct the potential.
10. It should be easy to calculate analytic derivatives to obtain the atomic forces.

The goal of this lecture is to discuss, which of these criteria are fulfilled to which extent by a promising, rather new type of potential based on artificial neural networks (NNs). NNs represent a class of very flexible mathematical functions, which have been first developed to study the signal processing in the nervous system. They have a lot of properties, which make them ideal candidates to construct potentials for "difficult systems" that are hard to describe by conventional potentials. In this lecture the advantages and limitations of the NN method will be presented, and the current status with respect to the ten criteria

mentioned above will be discussed in detail. The interested reader is also referred to a few recent reviews on NN potentials[22,23,21], which provide further information and a number of examples.

2 Neural Networks

2.1 Overview

The first artificial NNs have been suggested by McCulloch and Pitts in 1943 to study the signal processing in the brain[24]. Similar to a neuron in the nervous system an artificial neuron first collects incoming signals. If the accumulated signal exceeds a certain threshold, the neuron itself sends a signal to its neighboring neurons. A few years later, in 1958, the perceptron was introduced by Rosenblatt[25], which was the first NN consisting of a set of neurons arranged in an input and an output layer. In 1969, Minsky and Papert have shown that perceptrons have some serious limitations[26], for instance they are not able to represent all logical functions. However, soon this problem could be solved by introducing hidden layers. The optimization of the weight parameters of these extended NNs first posed a significant problem, which was solved in 1974 by Werbos[27]. Another important contribution in the same year was the introduction of non-linear functions for the transformation of the output of perceptrons[28]. This extension enabled the generation continuous output values instead of the binary output of the early NNs. This has been a crucial step for the use of NNs in function fitting.

A general definition of neural networks has been given by Kohonen[29]: "*Artificial neural networks are massively parallel interconnected networks of simple (usually adaptive) elements and their hierarchical organizations, which are intended to interact with the objects of the real world in the same way as biological nervous systems do.*" There are many types of NNs[30,31], but to date only a few of them have found applications in the field of chemistry and physics[32–35]. Most of these applications are related to the ability of NNs to recognize patterns in complex data sets and to classify information.

The topic of this lecture is a different kind of application, the construction of a complicated high-dimensional function, the PES, using NNs. It has been shown by several groups that NNs are able to approximate any real-valued high-dimensional function to arbitrary accuracy by feed-forward NNs[36,37]. This important result is the theoretical foundation for the development of NN PESs. In general, the aim is to construct a functional relation between the atomic configuration and its energy using a discrete set of known points, which have been determined by first-principles. Once available, this function can then be used to provide the energies and forces for arbitrary atomic configurations, which is a mandatory requirement for carrying out molecular dynamics simulations. In the following chapters, the basic properties of NN potentials will be discussed.

2.2 Feed-Forward Neural Networks

The central component of any NN potential is the feed-forward neural network. A small example network is shown schematically in Fig. 1. It consists of artificial neurons also called nodes, which are arranged in layers. They are represented by the grey circles. The goal of the NN is to construct a functional relation between the total energy of a system and

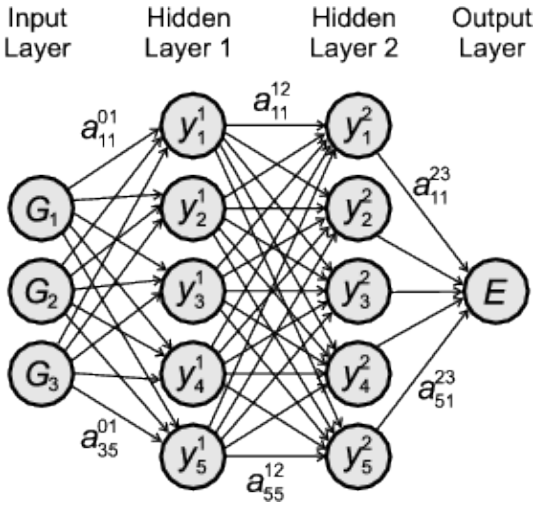


Figure 1. Example for a small feed-forward neural network (NN) with two hidden layers each of which contains five neurons. The hidden layers and their nodes determine the functional relation between the NN output E and the three coordinates G_1, G_2, and G_3, which define the atomic configuration. All nodes in adjacent layers are connected by weight parameters, which are shown as arrows. The bias weights shifting the non-linear regions of the activation functions are not shown. The full functional form of this NN is given by Eq. 3.

the atomic positions. For this purpose, the system is specified by a set of coordinates $\{G_i\}$, and accordingly the example NN in Fig. 1 represents a three-dimensional system, suitable for the construction of a three-atomic molecule, as the number of degrees of freedom of an N-atom system is $3N - 6$. If the set of $\{G_i\}$ is provided, the energy E is obtained in the output node.

The functional form of the NN is given by its architecture, i.e., the number of hidden layers and the number of nodes per layer in between the input and the output layer. The nodes in the hidden layers have no physical meaning, but the more nodes and layers are present, the higher is the flexibility of the NN. All nodes in adjacent layers are connected by weight parameters, which are shown as arrows in Fig. 1. These are the fitting parameters of the NN. Here, we use the symbol a_{ij}^{kl} for a weight connecting node i in layer k with node j in layer $l = k + 1$.

In order to calculate the output of the NN, the values of all nodes are calculated step by step starting at the input layer. In general, the numerical value of a node m in layer n, y_m^n, is given by

$$y_m^n = f_m^n \left(b_m^n + \sum_{i=1}^{M_{n-1}} a_{im}^{n-1,n} \cdot y_i^{n-1} \right) \quad . \tag{2}$$

First, a linear combination of the M_{n-1} values of the nodes in the previous layer $n - 1$ is calculated using the connecting weights as coefficients. Then, a bias weight b_m^n is added. Its purpose is to shift the linear combination to the non-linear region of the activation function f_m^n, which is finally applied to the result (cf. Sec. 2.3). In summary, a number is obtained at each node of the first hidden layer. Then these numbers are multiplied by

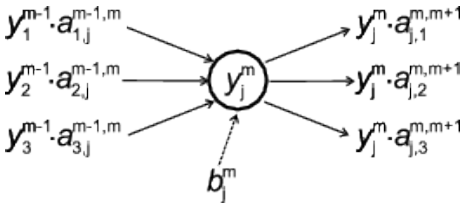


Figure 2. Scheme of an artificial neuron j in layer m. Its numerical value y_j^m is calculated according to Eq. 2 by first constructing a linear combination of the values y_k^{m-1} of the nodes in the previous layer $m-1$ and using the connecting weights $a_{k,j}^{m-1,m}$ as coefficients. Then, a bias weight b_j^m is added to adjust the non-linear region of the activation function f_j^m, which is finally applied to yield y_j^m. This value is then multiplied by the connecting weights $a_{j,l}^{m,m+1}$ passed forward to the nodes in the next layer $m+1$.

the weights connecting the nodes to the nodes in the subsequent layer in order to calculate the values in the next hidden layer. This flow of information in an artificial neuron is summarized in Fig. 2. The procedure is then repeated layer by layer until the output value E is obtained.

The complete functional form of the small example NN in Fig. 1 is then given by

$$E = f_1^3\left(b_1^3 + \sum_{l=1}^{5} a_{l1}^{23} \cdot f_l^2\left(b_l^2 + \sum_{k=1}^{5} a_{kl}^{12} \cdot f_k^1\left(b_k^1 + \sum_{j=1}^{3} a_{jk}^{01} \cdot G_j\right)\right)\right) \quad , \tag{3}$$

and its architecture can be specified by the short notation 3-5-5-1, which provides the number of nodes in each layer.

Neural networks typically contain a very large number of weight parameters N_w, which can be calculated by

$$N_w = \sum_{k=1}^{N_{\mathrm{HL}}+1} \left(M_{k-1} \cdot M_k + M_k\right) \quad , \tag{4}$$

where N_{HL} is the number of hidden layers and M_k the number of nodes in layer k. In practical applications, even NNs of moderate size can contain a few thousand parameters. This has to be kept in mind in the fitting process, since the number of fitting parameters must not exceed the information content of the training points.

Initially, the weight parameters are chosen as random numbers and consequently the NN output does not provide the correct energy of the systems. Still, the "correct" energy for a number of configurations can be determined using electronic structure calculations, and the NN parameters can be "trained" to reproduce these reference energies. Consequently, the NN "learns" the topology of the PES. Once a set of weight parameters has been found, which accurately reproduces all example points, the parameter set is frozen and the NN can be applied to predict the energy of (similar) unknown structures.

2.3 Activation Functions

Neural networks obtain the capability to fit arbitrary functions by employing highly flexible activation functions, which are also called basis functions or transfer functions. In general,

these are non-linear functions, and in the NN method the target function is constructed as a superposition of (nested) activation functions whose shapes are adjusted by the NN weight parameters.

It general, activation functions have a very simple form, and they have the property that they converge to some constant output for very large and very small arguments, while in between they possess a non-linear region. Further, analytic derivatives need to be available in order to optimize the NN parameters using gradient-based optimization algorithms, and to enable the calculation of analytic derivatives of the target function. This is needed for instance to calculate the forces, which are required in applications like geometry optimizations or molecular dynamics.

Many functional forms have been proposed for activation functions, like the sigmoid function

$$f(x) = \frac{1}{1+e^{-x}} \quad , \tag{5}$$

the hyperbolic tangent

$$f(x) = tanh(x) \quad , \tag{6}$$

and the Gaussian function

$$f(x) = e^{-\alpha x^2} \quad . \tag{7}$$

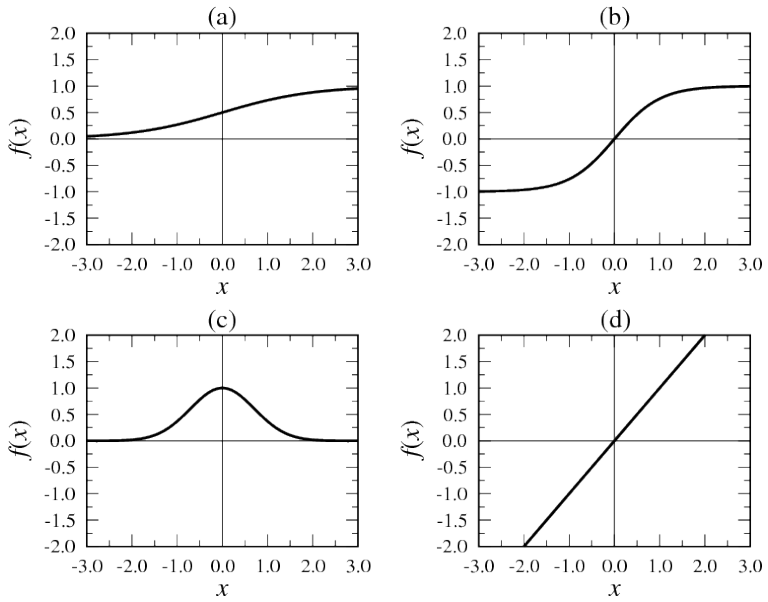


Figure 3. Activation functions typically used for the construction of neural network potentials. For the nodes in the hidden layers usually non-linear functions like the sigmoid function (a), the hyperbolic tangent (b) or the Gaussian function (c) are employed. In the output layer often a linear function (d) is used to avoid any constraint in the range of output values.

For the output node in most cases a linear function

$$f(x) = x \tag{8}$$

is used. This has the advantage that the range of possible output values is not restricted to some finite interval. The most common activation functions are shown in Fig. 3. In Fig. 4 it is demonstrated how a cosine function containing a number of extrema can be constructed as a superposition of hyperbolic tangent activation functions.

There are also some special cases, in which periodic activation functions like the cosine function have been used[38], which is advantageous in particular if the target function is periodic, like e.g. in case of dihedral potentials.

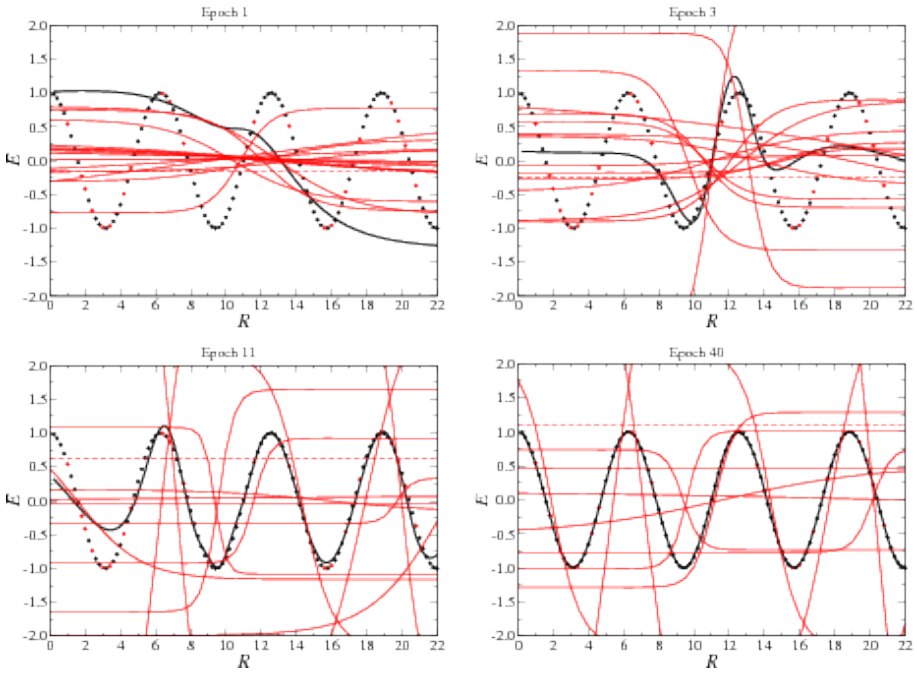


Figure 4. Fit of the cosine function $E = \cos(R)$ in the interval $[0, 22]$ by a 1-15-1 neural network (NN) using a hyperbolic tangent activation function. The black symbols represent the training points to adjust the weights, the red symbols are independent test points. The bias weight is plotted as dashed line, the contributions of the individual nodes $y_m^1 \cdot a_{m1}^{12}$ as red lines. The total NN function is plotted in black. The input coordinate has been preprocessed by shifting the average of all R values to zero. Accordingly, in the initial epoch the centers of the non-linear regions of the activation functions are close to 11. It can be seen that after a few epochs (iterations) the cosine function is well reproduced by a superposition of the red curves.

2.4 Input Coordinates: Symmetry Functions

One of the most important aspects of constructing a NN potential is the choice of a suitable set of input coordinates $\{G_i\}$ to describe the structure. Often, the easiest way to define a structure is to simply specify the Cartesian coordinates of all atoms. Unfortunately, in

case of NN potentials the Cartesian coordinates cannot be used, because their absolute values have no physical meaning. Instead, the relative positions of the atoms are important for the energy of a system. In other words, as the energy of a system is invariant with respect to rotation and translation, the coordinates describing the atomic positions should have the same property. This requirement is not fulfilled for Cartesian coordinates, since their numerical values change for instance if the system is merely shifted in space. As the NN is a purely mathematical approach processing numbers, generally a different output is obtained if the input vector changes numerically, even if the system's internal structure is not modified.

A straightforward solution to avoid this problem is to use internal coordinates like interatomic distances and angles. However, using internal coordinates is also not without problems, because their number grows rapidly with system size, and the choice is not unique. For this reason it has been proposed to use the full distance matrix to describe a system[39], but this is feasible only for small molecules.

Another even more critical problem of internal coordinates is the dependence of the results on their order. As the weight parameters in a NN are generally all different, a different NN energy is obtained if the order of the values in the input nodes is switched. A simple example to illustrate this problem is an isolated water molecule. There are two OH bonds, which under real conditions (finite temperature) will always be different. If we assume that $R_{\mathrm{OH1}} < R_{\mathrm{OH2}}$ and carry out an electronic structure calculation for this configuration, it is not necessary to calculate the same molecule again with exchanged hydrogen atoms ($R_{\mathrm{OH1}} > R_{\mathrm{OH2}}$), because both situations are chemically indistinguishable. Unfortunately, this equivalence is not included in the NN energy expression, because the first input node refers to R_{OH1} and the second to R_{OH2}. Both input nodes are connected to the nodes in the hidden layers by numerically different weight parameters, therefore both chemically equivalent configurations will yield a different energy output. Of course this must be avoided.

For small systems it is possible to solve this problem by a transformation of the coordinates in the spirit of Gassner et al., who have proposed a set of symmetrized coordinates for the H_2O-Al^{3+}-H_2O complex[40]. For a single water molecule, a solution is to use the coordinates

$$G_1 = |R_{\mathrm{OH1}} + R_{\mathrm{OH2}}| \tag{9}$$

and

$$G_2 = |R_{\mathrm{OH1}} - R_{\mathrm{OH2}}| \tag{10}$$

instead of R_{OH1} and R_{OH2}. In this new set of coordinates the order of the hydrogen atoms is arbitrary and the NN output is invariant with respect to the particular choice. The full geometry of the water molecule can be specified by adding a third coordinate R_{HH}, which is unique and thus does not need to be symmetrized.

In general, suitable sets of symmetry-adapted functions are called "symmetry functions". A set of symmetry functions is defined as a set of functions, whose vector of values is the same for any energetically equivalent representation of a system. For instance, the vector of symmetry function values must be the same if any two atoms of the same element in a system exchange their positions. Further, symmetry functions contain also all information on the symmetry of the system. If for example a molecule has mirror symmetry, then both forms yield the same set of symmetry function values.

Two basic solutions to avoid the introduction of symmetry functions are either to sort the input coordinates according to their numerical values before providing them to the NN or to include equivalent structures several times in the training set in each possible atomic order. A drawback of the latter approach is an increased computational fitting effort due to the larger training set. An even more severe problem of this approach is that the different realizations of a given structure are not exactly equivalent, because the NN learns all structures independently. As a result, the symmetry is necessarily broken numerically.

In summary, the approach suggested by Gassner et al.[40] is the best solution for small molecules, and similar functions have also been proposed for the interaction of small molecules with metal surfaces[41,42]. Alternatively, it has also been suggested to deal with the symmetry by using symmetric neurons[43]. Unfortunately, all these approaches are applicable only to small systems containing just a few atoms, because of the rapidly increasing complexity of the resulting functions.

Symmetry functions need to be constructed with care, since artificial symmetries included in these functions necessarily become a property of the NN potential. This can be demonstrated for the periodic function $f(x)$ shown in Fig. 5. The function has a 2π periodicity, which therefore should also be present in the symmetry functions. Thus, either a sine or a cosine function could be used. If, however, just a cosine function is employed, which has the additional property $\cos(x) = \cos(-x)$, also the NN potential will have this symmetry, which is obviously not correct. A similar problem is present if only the sine function is used. The wrong symmetry features can be removed by using a two-dimensional vector of symmetry functions instead of a single function even in this case of a one-dimensional system. The first component of this vector is the sine function, the second is the cosine

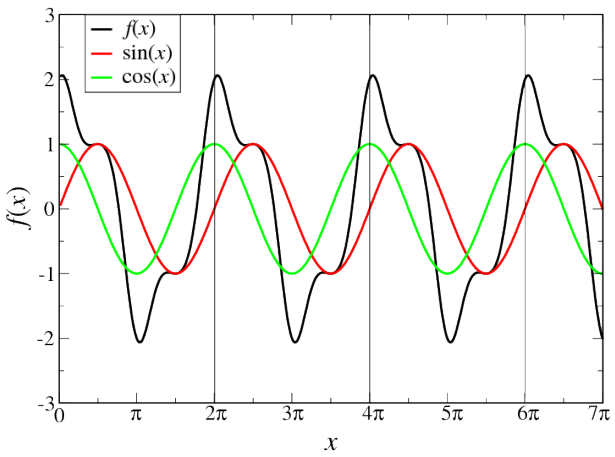


Figure 5. Illustration of the two-dimensional vector of symmetry functions required to fit a general periodic function $f(x)$ with a period length 2π. There is a unique relation between the vector of symmetry function values and the value of the target function. Therefore, using either the sine function (red curve) or the cosine function (green curve) alone is not sufficient, because they have a higher symmetry than $f(x)$. Only the two-dimensional vector containing the sine and the cosine function values for each x have the correct symmetry and allow to construct $f(x)$.

function. The overall periodicity is then still correct, but the wrong additional symmetry is broken. As can be seen in Fig. 5, there is a unique one-to-one correspondence between $f(x)$ and the vector $(\sin(x), \cos(x))$. This example shows that the dimensionality of the systems is not necessarily always the same as the dimensionality of the symmetry function vector. In Sec. 3 we will see how even high-dimensional PESs can be constructed using rather low-dimensional symmetry function vectors.

Symmetry functions contain important features of the target function, but they are much easier to construct than analytic potentials, because they to not have to match the PES in value. Instead, in case of NN potentials the task to construct a functional relation between the (Cartesian) coordinates of the atoms and the total energy of the system is split into two independent problems. First, a coordinate transformation onto symmetry functions is carried out. Then, in a second step the NN is used to associate the vector of symmetry function values with an energy. In conventional empirical potentials, both problems need to be solved in one step, which is a significantly more difficult task.

Symmetry functions are often specific for a given system. Still, there are also very general recipes how more general symmetry functions can be constructed, e.g. for molecule-surface interactions[42]. In Sec. 3 we will see that even for high-dimensional NN potentials very general symmetry functions can be defined, which are basically independent of the system.

Finally it should be noted that symmetry functions can have functional forms, which are difficult to invert. This is possible, because the transformation of the coordinates has to be done always just in one direction, from the Cartesian coordinates of the atoms to the symmetry functions. This is true for the training of NNs as well as for applications. It is not necessary in any situation to determine the Cartesian coordinates for a given set of symmetry function values.

2.5 Fitting the Weight Parameters

The weight parameters are determined by minimizing the error function

$$\Gamma = \sum_{i=1}^{N} \frac{1}{N} (E_{i,\mathrm{ref}} - E_{i,\mathrm{NN}})^2 \tag{11}$$

using a set of known energies reference energies from electronic structure calculations. In general, NNs contain a very large number of fitting parameters. Therefore, in most cases it is impossible to find the global minimum. Still, many local minima are sufficiently accurate to provide a good description of the PES. A wide range of optimization algorithms can be used, like steepest descent, which is called backpropagation in the NN community[44], conjugate gradients[45] and the global extended Kalman filter[46,47], just to give a few examples. For all these algorithms, the weights are adjusted iteratively, and the accuracy of the fit is checked by calculating the root mean squared error (RMSE)

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (E_{i,\mathrm{ref}} - E_{i,\mathrm{NN}})^2} \quad . \tag{12}$$

Since the RMSE strongly emphasizes outliers, also the mean absolute deviation (MAD) is often used,

$$MAD = \frac{1}{N}\sum_{i=1}^{N} |E_{i,\mathrm{ref}} - E_{i,\mathrm{NN}}| \quad . \tag{13}$$

However, monitoring the error of the training set is not sufficient to ensure that a reliable PES is obtained. Since in applications of the NN potential also energies and forces for structures in between the training points are requested, it is important to validate that the potential has the correct shape also for structures not included in the training set. The accuracy of such points can be investigated by the so-called "early stopping" method. Here, not all available reference points are used for the optimization of the NN weight parameters, but some fraction, usually 10 %, is used as an independent test or validation set. During the fit, the error of both, the training and the test set is observed. In the initial stage of the fit, both errors typically decrease rapidly. Then, while the error of the training set is further decreasing, the test set error often exhibits a minimum and rises again. This is the onset of overfitting, which is an improvement of the error of the training points for the price of a reduced accuracy for atomic configurations, which are in between these points. It is not possible to detect overfitting by investigating the error of the training points alone. An example for overfitting is shown in Fig. 6. While the fitted black curve is close to all known training points shown as black symbols, the overall shape of the potential exhibits some additional extrema, which are not justified by the training set. This fit would yield a very small error for the training set, but still the quality of the potential is not good.

Apart from the numerical values of the weight parameters, also the architecture of the NN, i.e., the number of hidden layers and nodes per layeri, is important to obtain an accurate PES. It has been suggested to adjust the number of nodes on-the-fly in the fitting process[48,49], but such approaches are usually computationally very demanding. Often, it

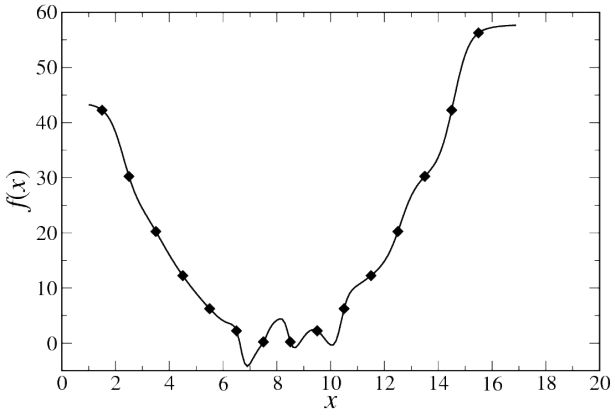


Figure 6. Example for a poor potential showing overfitting. While the potential (black line) represents all training points very accurately, the regions in between the training points exhibit spurious extrema. In order to detect such situations, not all available points should be included for the fitting process, but some test points, which are not used in the weight optimization, should be used to check the reliability of the potential in between the training points.

is easier to employ a trial and error approach and to construct a number of PESs using different NN architectures. Then, the results for different NNs are compared and the best fit is chosen. The optimum choice of the NN architecture can be guided by the errors of the training and the test set. If both errors remain large, the NN is not sufficiently flexible to resolve all subtle features of the PES and some details are averaged out. In this situation the number of nodes should be increased. If, however, the error of the training set is very low, but the error of the test set is clearly larger, overfitting is present, and the flexibility of the NN is too high. The optimal choice for the NN architecture shows a similar error of the training and the test set, which should both be acceptably small.

Finally, it should be noted that the fitting result does not only depend on the NN architecture, but also the initial values of the weight parameters, the choice of the optimization algorithm and the order of the training points can have a strong influence on the obtained potential, because under different conditions different local minima can be found.

2.6 Selection of the Training Points

The choice of the atomic configurations for the reference electronic structure calculations requires special care. They must contain information about all relevant features of the PES, because the NN itself does not have a physically meaningful functional form, and the physical shape of the PES has to be learned from the example points.

For low-dimensional systems like e.g. small molecules, it is possible to map the underlying PES systematically using a dense grid of points. It is not necessary that these points are located on an equidistant grid, and consequently important low-energy regions of the system can be mapped on a denser grid than less-important high-energy regions. Still, this mapping can be done only for very small systems, because the number of points increases exponentially with the number of degrees of freedom. Even if only a very sparse grid with e.g. five points per degree of freedom is used, for a three-atom molecule this results in $5^3 = 125$ reference calculations. For a molecule containing six atoms, this number increases already to $5^6 = 15,625$. Therefore, for larger systems this mapping is not feasible.

Fortunately, often a large fraction of configuration space is chemically not relevant, because the energy is too high and the corresponding points cannot be visited in MD simulations at chemically meaningful temperatures. Many approaches have been proposed in the literature to identify the important points, e.g. by carrying out molecular dynamics simulations to sample configurations[17,50,51] along typical reaction paths.

In general, the construction of NN PESs requires a significantly larger number of training points than empirical potentials. If too few training points are used, the NN can exhibit physically wrong features. However, this drawback of the NN approach can be exploited to identify important structures, which are missing in the training set without carrying out unnecessary electronic structure calculations for configurations, which are already well represented in the training set[52]. This is done in an iterative way. First, several initial potentials with about the same RMSEs are constructed using a few hundred configurations only. These potentials should have different functional forms, which can easily be achieved by selecting different NN architectures. It is then not clear, which of these fits is the best, and for sure they will not provide reliable energies and forces in all situations. Nevertheless, they can be used to suggest new important structures, which should be included in the

training set. This is done by generating a large number of trial structures using geometry optimizations and MD simulations employing one of these fits. Then, the energies and forces of these structures are recalculated using the other fits. If for a given structure all fits predict a similar energy, then this structure if probably very close to a point already included in the training set, and all NNs have been trained to reproduce its energy. If, on the other hand, all fits predict very different energies, then an electronic structure calculation should be carried out for this structure, and the result should be added to the training set to refine the potential. Following this recipe, the NN potential can be improved step by step until a consistent potential is obtained.

3 High-Dimensional Neural Network Potentials

Potentials based on a single feed-forward NN have been constructed for a wide range of small molecules and also the interaction of small, in most cases diatomic, molecules with frozen metal surfaces (cf. Sec. 4). Still, these potentials have a number of limitations and NN potentials will become competitive for applications to large-scale MD simulations only if NN PESs for high-dimensional systems become available.

One serious problem is that in general for each atom in the system there are three degrees of freedom, which have to be provided as input nodes to the NN. Therefore, for a large number of atoms NNs can get very large. Consequently, the evaluation becomes more costly. Further, the more input degrees of freedom are considered, the more training points are required to represent this high-dimensional configuration space. Finally, another consequence of a large number of nodes is a large amount of connecting weight parameters making the fitting process more difficult.

A second problem is the fixed dimensionality of NNs. Even if it would be possible to construct a NN PES for a large system containing e.g. 1000 atoms, this potential could not be applied to systems containing a different number of atoms. This is because for smaller systems the input nodes of the missing atoms are not defined, and for larger systems the required values of the additional connecting weights are not available. For generalized high-dimensional NN PESs it is therefore necessary to find a way to apply a potential, once it has been constructed, to systems containing different numbers of atoms.

Finally, the construction of suitable symmetry functions taking into account the permutation symmetry of all atoms of the same chemical element is a formidable challenge and the recipes developed for low-dimensional systems are not applicable.

A high-dimensional NN approach solving these problems has been suggested by Behler and Parrinello in 2007[53]. In this potential type, which is shown schematically in Fig. 7, the total energy of the system is not constructed using a single feed-forward NN, but a separate NN is introduced for each atom, which provides only the energy contribution of this atom E_i to the total energy. The total energy is then the sum over all atomic energies,

$$E = \sum_i E_i \quad . \tag{14}$$

The E_i depend on the local chemical environments of the atoms, which are defined by a cutoff function

$$f_{\mathrm{c}}\left(R_{ij}\right) = \begin{cases} 0.5 \cdot \left[\cos\left(\frac{\pi R_{ij}}{R_{\mathrm{c}}}\right) + 1\right] \text{ for } R_{ij} \leq R_{\mathrm{c}} \\ 0 \text{ for } R_{ij} > R_{\mathrm{c}} \end{cases} , \tag{15}$$

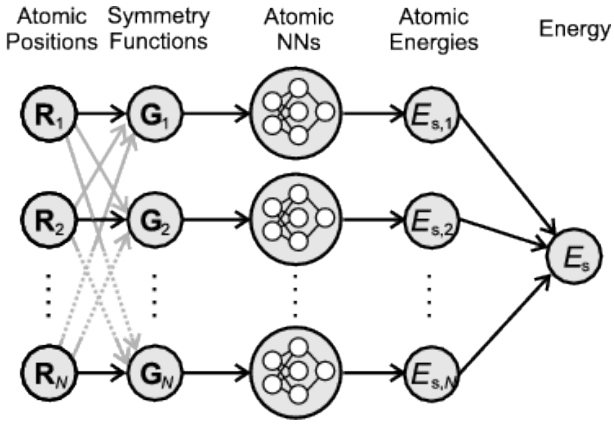


Figure 7. Structure of a high-dimensional neural network potential (NN) illustrated for an N-atom system[53]. The total (short range) energy E_s is a sum of individual atomic energy contributions $E_{s,i}$. These contributions are constructed using a separate atomic NN for each atom. The input for each NN is a vector of symmetry function values $\mathbf{G}_i$ describing the atomic environment up to a cutoff radius R_c. The symmetry functions are many-body functions depending on the Cartesian coordinate vectors $\mathbf{R}$ of all atoms inside the cutoff sphere.

which is similar to the cutoff function employed in the Tersoff potential[10]. The spatial extension of the cutoff function is given by the cutoff radius R_c, which typically has a value of about 6 Å. At R_c the cutoff function has zero value and slope.

The positions of all atoms in the chemical environment inside the cutoff sphere of an atom i are then described by a set of many-body symmetry functions. Several functional forms have been proposed[54]. The radial distribution of neighbors can be described by a "radial function"

$$G_i^1 = \sum_{j \neq i} e^{-\eta R_{ij}^2} \cdot f_c\left(R_{ij}\right) \quad . \tag{16}$$

R_{ij} is the distance between the central atom i and its neighbor j. This distance itself is not a good choice for the symmetry function for two reasons. First, the numerical value of R_{ij} increases with distance, while the physical interaction decreases. Therefore, R_{ij} is replace by a Gaussian, which decays rapidly with increasing separation of the atoms. Each Gaussian is further multiplied by the cutoff function to ensure that G^1 has zero value and slope at the cutoff radius. Second, for each neighbor j there is one R_{ij} value, but the final number of symmetry functions must be independent of the number of neighbors. This is necessary because NNs always have a fixed number of input nodes, but the number of neighbors j inside the cutoff sphere can change in the course of a molecular dynamics simulation. Therefore, the Gaussians for all neighbors j are added to yield a single symmetry function value. The radial distribution of neighbors can be determined by using a set of symmetry functions of type G^1 with different spatial extensions, which are defined by the Gaussian exponent η. This exponent is a parameter defining the shape of the symmetry function, which is not changed during the iterative optimization of the NN weights. In Fig. 8 several radial functions of type G^1 are plotted for different values of η.

Apart from the radial symmetry functions, angular symmetry functions can be used to

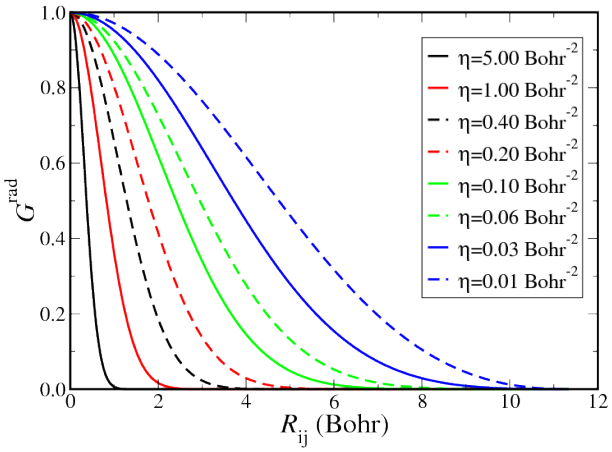


Figure 8. Radial atom-centered many-body symmetry functions used to describe the radial distribution of neighbors in the high-dimensional neural network approach of Behler and Parrinello[54].

specify the angular distribution of neighbors. They have the form

$$G_i^2 = 2^{1-\zeta} \sum_{j,k\neq i} (1+\lambda\cos\theta_{ijk})^\zeta \cdot e^{-\eta\left(R_{ij}^2+R_{ik}^2+R_{jk}^2\right)} \cdot f_c\left(R_{ij}\right)\cdot f_c\left(R_{ik}\right)\cdot f_c\left(R_{jk}\right) \quad . \tag{17}$$

The angle $\theta_{ijk} = \arccos\left(\frac{\mathbf{R_{ij}}\cdot\mathbf{R_{ik}}}{R_{ij}\cdot R_{ik}}\right)$ is centered at atom i, and there is one angle for each atom triple ijk. The parameter λ can have values of $+1$ and -1 and determines if the maxima of the cosine function are centered at $\theta_{ijk} = 0°$ or $\theta_{ijk} = 180°$. The angular resolution is obtained by using several angular functions with different ζ values. More details on these symmetry functions can be found elsewhere[54].

In total, the atomic environments are typically defined by a set of 40 to 100 radial and angular symmetry functions, and for each atom i there is one symmetry function vector $\mathbf{G}_i$. As shown in Fig. 7 this vector is then used as input for an atomic NN yielding the atomic energy contribution E_i. The grey arrows in Fig. 7 indicate that the symmetry functions are many-body functions depending explicitly on the positions of all atoms in the chemical environment. Finally, all E_i are added to provide the total energy E.

The scheme presented in Fig. 7 fulfills all requirements for a high-dimensional NN potential. The total energy is independent of the order of atoms, because the exchange of any two atoms of the same element changes just the order of summation. Further, once the NN parameters, which are constrained to be the same for all atomic NNs referring the to same chemical species, have been determined, the potential is applicable to any system size. If an atom is added, another line is included in the scheme of Fig. 7. If an atom is removed, its atomic NN is deleted. High-dimensional NN potentials of this type have been constructed for a number of systems, like silicon[53,55,56], carbon[57,58], sodium[59], and copper[52].

In case of multicomponent systems an extension is required. Because in systems of arbitrary chemical composition charge transfer occurs, there can be long-range electro-

static interactions, which are not included in the high-dimensional NN scheme discussed above. It has been proposed by Popelier and coworkers to express atom-centered electrostatic multipoles by NNs to improve the description of electrostatics in classical force fields[60,61]. In a similar way also the approach of Behler and Parrinello has been extended by an electrostatic energy term[62],

$$E_{\mathrm{tot}} = E_s + E_{elec} \quad . \tag{18}$$

Here, the short range energy contribution E_s corresponds to the original scheme proposed by Behler and Parrinello. The electrostatic interactions are based on environment-dependent atomic charges, which are constructed by a second set of atomic NNs. These charges can then be used to calculate the electrostatic energy and forces employing established standard algorithms like an Ewald summation[63].

Apart from this high-dimensional NN scheme, there is also another approach with similar capabilities. Already in 1999 Smith and coworkers proposed to improve the accuracy of empirical potentials by expressing some parts in the functional form by NNs. Specifically, for each pair of atoms they replaced the many-body bond order term in the attractive part of the Tersoff potential[10] by neural networks[64,65]. The variable number of neighbors of both atoms in the bond has been taken into account by constructing a chain of atoms including the pair and one neighbor for each atom in the environment. Several numbers are used to characterize the structure of these chains yielding one vector of input coordinates for each neighbor. Since the number of neighbors can be different in each structure, a NN with variable size has been proposed to process these vectors. Modified Tersoff potentials of this type have been reported for the binary systems "CH"[64,65], and "CN"[65]. Surprisingly, this promising approach has not been further developed for almost a decade. In 2007, however, Smith and coworkers have extended this approach to a true NN PES by abandoning the Tersoff functional form[66]. Instead, like in case of the Behler Parrinello scheme, which was proposed independently in the same year, the energy is constructed as a sum of atomic energy contributions. Similar to the original scheme, the atomic environments are described by a variable number of atomic chains, although chains of increased length are used to provide a better description of the structures. Again, the vectors describing these chains are used as input for NNs, whose architectures can be adjusted to the actual number of neighbors. The method has first been applied to reproduce tight-binding energies for silicon[66], and later also DFT energies have been used as reference for the same system[67].

4 Discussion

To date, neural network potentials have been constructed for a wide range of systems. Still, most applications have been reported for small molecules and molecule-surface interactions employing the frozen-surface approximation. A list of currently available potentials for such systems is compiled in Tables 1 and 2, respectively. These potentials are comparably easy to construct, because they are low-dimensional and in most cases a single feed-forward NN is sufficient to represent the full PES. Still, there is also an increasing number of high-dimensional NN PESs, with encouraging applications in the field of materials science. They are summarized in Tab. 3.

It is now time to return to the list of criteria for the "perfect potential" given in the introduction and to discuss, which points are fulfilled to which extent by current NN po-

Year	Ref.	System	Reference Method	Architecture
1996	Tafeit et al.[68]	tetrahydrobiopterin	Hartree Fock	2-3-1
1996	Tafeit et al.[68]	tetrahydrobiopterin	Hartree Fock	2-3-3-1
1996	Brown et al.[69]	$(HF)_2$	analytic PES	4-32-1
1996	Brown et al.[69]	HF-HCl complex	Hartree Fock	4-32-1
1997	No et al.[70]	$(H_2O)_2$	MP2	6-12-12-1
1998	Gassner et al.[40]	H_2O-Al^{3+}-H_2O	Hartree Fock	11-5-5-1
1998	Prudente et al.[71]	HCl^+	CI	1-3-4-1
1998	Prudente et al.[43]	H_3^+	—	3-15-1
2002	Cho et al.[72]	$(H_2O)_2$ in TIP4P	MP2	9-18-18-1
2003	Rocha Filho et al.[73]	H_3^+	—	3-12-3-1
2004	Bittencourt et al.[74]	OH	MRCI	1-3-1
2005	Raff et al.[51]	vinyl bromide	MP4	12-20-1
2005	Raff et al.[51]	Si_5 clusters	DFT (B3LYP)	9-45-1
2006	Agrawal et al.[75]	SiO_2 molecule	DFT (B3LYP)	3-40-1
2006	Manzhos et al.[76]	H_2O molecule	analytic PES	3-23-1
2006	Manzhos et al.[76]	HOOH	analytic PES	2 NNs
2006	Manzhos et al.[76]	H_2CO	analytic PES	2 NNs
2006	Manzhos et al.[77]	HOOH	analytic PES	6-90-1
2006	Manzhos et al.[77]	NOCl	analytic PES	3-10-1
2007	Houlding et al.[60]	$(HF)_2$	DFT (B3LYP)	4-3-1
2008	Darley et al.[61]	glycine	DFT (B3LYP)	several
2008	Darley et al.[61]	N-methylacetamide	DFT (B3LYP)	several
2008	Malshe et al.[78]	Si_5 clusters	DFT (B3LYP)	—
2008	Lee et al.[79]	HONO	MP4	6-41-1
2009	Malshe et al.[80]	vinyl bromide	MP4	15-140-1
2009	Le et al.[39]	BeH + H_2	MP2	6-60-1
2009	Pukrittayakamee et al.[81]	H+HBr	analytic PES	3-150-1
2009	Hung et al.[82]	HOOH	MP2	6-34-1

Table 1. List of neural network potentials for molecular systems reported in the literature.

tentials. Neural network potentials can be extremely *accurate*, and typically RMSEs of a few meV can be achieved for small molecules. For high-dimensional systems, the RMSEs need to be normalized per atom in order to make systems of different size comparable, and also for these NN PESs errors of a few meV per atom are usually obtained. Therefore, NN PESs are certainly among the most accurate potentials, if the criterion is the reproduction of reference ab initio energies. Still, of course, NN PESs cannot be more accurate than the underlying reference electronic structure method.

Another significant advantage of NN potentials is that they *can be improved systematically* and without the need to adjust the functional form. Whenever a situation is detected in which the potential is not sufficiently accurate, additional training points can be added to refine the potential without much effort.

Due to their "non-physical" functional form, NNs are equally *apt to construct PESs for very different systems* such as covalently bonded molecules, bulk metals and semiconductors, as well as for molecular complexes bound via weak van der Waals interactions.

Year	Ref.	System	Reference Method	NN architecture
1995	Blank et al.[83]	CO @ Ni(111)	empirical PES	3-15-1
1995	Blank et al.[83]	H_2 @ Si(100)-(2×1)	DFT (LDA)	12-8-1
2004	Lorenz et al.[41]	H_2 @ K(2×2)/Pd(100)	DFT (PW91)	8-24-18-1
2005	Behler et al.[84,85]	adiabatic O_2 @ Al(111)	DFT (RPBE)	11-40-40-1
2005	Behler et al.[84,42,85]	triplet O_2 @ Al(111)	DFT (RPBE)	11-40-40-1
2006	Lorenz et al.[86]	H_2 @ Pd(100)	empirical PES	—
2006	Lorenz et al.[86]	H_2 @ Pd(100)	empirical PES	8-50-50-1
2006	Lorenz et al.[86]	H_2 @ Pd(100)	empirical PES	8-50-50-1
2006	Lorenz et al.[86]	H_2 @ (2×2)S/Pd(100)	empirical PES	9-50-50-1
2006	Lorenz et al.[86]	H_2 @ (2×2)S/Pd(100)	empirical PES	9-20-20-1
2006	Lorenz et al.[86]	H_2 @ (2×2)S/Pd(100)	empirical PES	9-20-20-1
2007	Ludwig and Vlachos[87]	H_2 @ Pt(111)	empirical PES	7-25-25-1
2007	Ludwig and Vlachos[87]	H_2 @ Pt(111)	DFT (PW91)	7-50-50-1
2008	Behler et al.[85]	adiabatic O_2 @ Al(111)	DFT (PBE)	11-38-38-1
2008	Behler et al.[85]	triplet O_2 @ Al(111)	DFT (PBE)	11-40-10-1
2008	Latino et al.[88]	ethanol @ Au(111)	DFT (B3LYP)	6-8-1

Table 2. List of neural network potentials reported in the literature to describe molecule-surface interactions.

Year	Ref.	System	Reference Method	NN architecture
1999	Hobday et al.[64]	CH and carbon	experiment	$5N$-6-1
1999	Hobday et al.[65]	CN	experiment	$5N$-6-1
2006	Bholoa et al.[66]	silicon	tight binding	$9N$-11-11-1
2007	Behler and Parrinello[53]	silicon	DFT (LDA)	48-40-40-1
2008	Sanville et al.[67]	silicon	DFT (LDA)	$13N$-13-13-1
2010	Khaliullin et al.[57]	carbon	DFT (PBE)	48-25-25-1
2010	Eshet et al.[59]	sodium	DFT (PBE)	48-25-25-1
2011	Artrith et al.[62]	ZnO	DFT (PBE)	48-20-20-1

Table 3. List of high-dimensional neural network potentials reported in the literature.

Further, NN PESs provide the energy and forces as a function of the atomic positions and the nuclear charges, i.e., the chemical elements. Therefore, no bonds need to be specified and like electronic structure methods *NN potentials are reactive*, i.e., they allow to describe the making and breaking of bonds.

In Sec. 3 two schemes have been discussed, which are suitable to deal with *high-dimensional systems*. In principle, NNs can be applied to systems containing thousands of atoms, but a current limitation is the restriction to only a few chemical elements in these systems. The problems arising for systems containing more than three or four different elements are related to the symmetry functions, which are needed to describe the local chemical environments of the atoms. The complexity of the configuration space increases rapidly with the number of chemical species, and better approaches still need to be found for systems of arbitrary chemical composition. An advantage of NN PESs is that *they can be constructed without much human work*. Most parts of the optimization process includ-

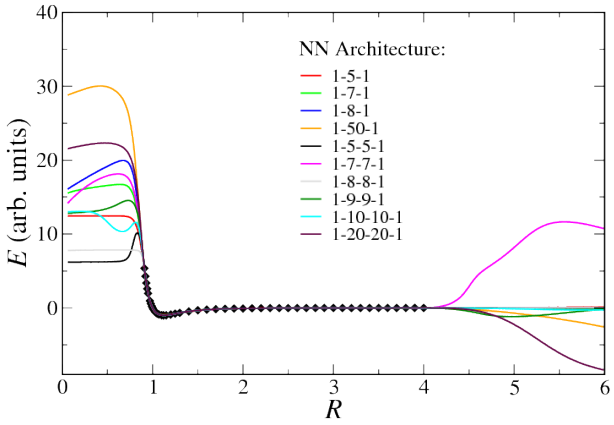


Figure 9. As a consequence of the high flexibility of neural networks, the shape of the potential energy surface can strongly vary outside the fitting range. In this example of a dimer potential, all fits are very similar in the interval spanned by the training points (black diamonds), but outside this interval the potentials can have an unphysical shape.

ing the search for relevant structures missing in the training set can in principle be carried out in a fully automatic way. Still, in terms of computer time some human intervention may reduce the effort significantly.

One of the major drawbacks of NN potentials is their *limited extrapolation capability*. If a NN has been trained, for instance, to reproduce the PES of a bulk metal, it is not directly applicable to metal surfaces unless surface structures have been included in the training set. Still, it is always possible to extend the range of validity of a NN potential by adding more training structures. In Fig. 9 the potential for some arbitrary diatomic molecule is shown, which has been trained using structures in the range $0.8 < R < 4.0$ employing several NN architectures. While in the range spanned by the training points all potentials reproduce the PES very accurately, it can be seen that the shape of the PESs outside this range shows physically wrong features. Such structures outside the training interval must not be visited in applications of the potential, otherwise wrong results will be obtained. Fortunately, such problematic cases are very easy to detect automatically. For each symmetry function (in the present example R would be a suitable symmetry function) the minimum and the maximum value present in the training set can be calculated and stored. Whenever the energy is requested for a structure with a symmetry function value outside this range, the NN program can issue a warning and stop the simulation. This procedure can also be applied to search systematically for structures missing in the training set to extend the range of validity of the potential.

Concerning the *efficiency*, NN PESs can be evaluated significantly faster than any electronic structure method, but due to their rather complicated functional form they are computationally more demanding than very basic potentials like e.g. classical force fields. To give an example, current high-dimensional NN codes can provide the energy and forces for about 100 to 200 atoms per second per compute core.

Due to their unbiased and very flexible functional form, the construction of a NN PES requires a large number of first-principles training points to ensure that the final PES has

the correct shape. This makes the *construction of NN PESs computationally more demanding* than the development of "physical potentials". Still, this effort pays off quickly, if extended applications like large-scale MD simulations are carried out using the NN potential. Assuming that about 20,000 reference calculations are needed, this corresponds to an *ab initio* MD simulation of about 20 ps, if a time step of 1 fs is used. Further, since the atomic energy contributions in high-dimensional NN schemes depend only on the local chemical environments of the atoms, NN PESs can be applied to much larger systems than have been used in the training set. Finally, *analytic derivatives* are easily accessible, since the NN has a well-defined functional form.

5 Conclusions

In this lecture, the current state of neural network potentials has been reviewed. NNs enable to construct numerically very accurate potentials for a wide range of systems with results very close to first principles methods. Still, due to their flexibility, they have to be constructed and validated with care, and a large number of reference calculations is required, which makes the construction of NN PESs rather costly compared to conventional empirical potentials. Still, NN PESs offer many advantages for systems, which are difficult do describe by other types of potentials, for instance if very different physical interactions or complicated bonding situations are present. Promising examples for future applications are the study of phase diagrams in materials science, but also complex chemical processes at interfaces or in solution.

Acknowledgments

Financial support by the DFG (Emmy Noether program) and the FCI is gratefully acknowledged.

References

1. W. Koch and M. C. Holthausen, *A Chemist's Guide to Density Functional Theory*, Wiley-VCH, Weinheim, 2001.
2. R. G. Parr and W. Yang, *Density-Functional Theory of Atoms and Molecules*, Oxford University Press, 1989.
3. D. Marx and J. Hutter, *Ab Initio Molecular Dynamics: Basic Theory and Advanced Methods*, Cambridge University Press, 2009.
4. R. Car, and M. Parrinello, *Unified Approach for Molecular Dynamics and Density-Functional Theory* Phys. Rev. Lett. **55**, 2471, 1985.
5. N. L. Allinger, Y. H. Yuh, and J.-H. Lii, *Molecular Mechanics. The MM3 force field for hydrocarbons* J. Am. Chem. Soc. **111**, 8551, 1989.
6. S. L. Mayo, B. D. Olafson, and W. A. Goddard III, *DREIDING: A Generic Force Field for Molecular Simulations* J. Phys. Chem. **94**, 8897, 1990.
7. A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. Goddard III, and W. M. Skiff, *UFF, a Full Periodic Table Force Field for Molecular Mechanics and Molecular Dynamics Simulations* J. Am. Chem. Soc. **114**, 10024, 1992.

8. B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, *CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations* J. Comp. Chem. **4**, 178, 1983.
9. W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, Jr., K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, *A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules* J. Am. Chem. Soc. **117**, 5179, 1995.
10. J. Tersoff, *New empirical model for the structural properties of silicon* Phys. Rev. Lett. **56**, 632, 1986.
11. J. Tersoff, *Empirical interatomic potential for silicon with improved elastic properties* Phys. Rev. B **38**, 9902, 1988.
12. M. S. Daw, and M. I. Baskes, *Embedded-atom method: Derivation and applications to impurities, surfaces and other defects in metals* Phys. Rev. B **29**, 6443, 1984.
13. M. I. Baskes, *Modified embedded-atom potentials for cubic materials and impurities* Phys. Rev. B **46**, 2727, 1992.
14. W. H. Press and S. A. Teukolsky and W. T. Vetterling and B. P. Flannery, *Numerical Recipes - The Art of Scientific Computing*, Cambridge University Press, Cambridge, 2007.
15. G. G. Maisuradze, D. L. Thompson, A. F. Wagner, and M. Minkoff, *Interpolating moving least-squares methods for fitting potential energy surfaces: Detailed analysis of one-dimensional applications* J. Chem. Phys. **119**, 10002, 2003.
16. Y. Guo, A. Kawano, D.L. Thompson, A. F. Wagner, and M. Minkoff, *Interpolating moving least-squares methods for fitting potential energy surfaces: Applications to classical dynamics calculations* J. Chem. Phys. **121**, 5091, 2004.
17. J. Ischtwan and M. A. Collins, *Molecular potential energy surfaces by interpolation* J. Chem. Phys. **100**, 8080, 1994.
18. M. J. T. Jordan, K. C. Thompson, and M. A. Collins, *Convergence of molecular potential energy surfaces by interpolation: Application to the OH+H_2 $\rightarrow$ H_2O+H reaction* J. Chem. Phys. **102**, 5647, 1995.
19. D. E. Makarov, and H. Metiu, *Fitting potential energy surfaces: A search in the function space by directed genetic programming* J. Chem. Phys. **108**, 590, 1998.
20. A. P. Bartok, M. C. Payne, R. Kondor, and G. Csanyi, *Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons* Phys. Rev. Lett. **104**, 136403, 2010.
21. J. Behler, *Neural network potential-energy surfaces in chemistry: a tool for large-scale simulations* Phys. Chem. Chem. Phys. **13**, 17930, 2011.
22. C. M. Handley, and P. L. A. Popelier, *Potential energy surfaces fitted by artificial neural networks* J. Phys. Chem. A **114**, 3371, 2010.
23. J. Behler, *Neural network potential-energy surfaces for atomistic simulations* Chem. Modelling **7**, 1, 2010.
24. W. McCulloch, and W. Pitts, *A logical calculus of the ideas immanent in nervous activity* Bull. Math. Biophysics **5**, 115, 1943.
25. F. Rosenblatt *The perceptron: A probabilistic model for information storage and organization in the brain* Psych. Rev. **65**, 386, 1958.
26. M. Minsky and S. Papert, *Perceptrons*, MIT Press, Cambridge, MA, 1969.

27. P. J. Werbos, *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences*, PhD Thesis, Harvard University, 1974.
28. W. A. Little, *The existence of persistent states in the brain* Math. Biosciences **19**, 101, 1974.
29. T. Kohonen, *An introduction to neural computing* Neural Networks **1**, 3, 1988.
30. C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, 1995.
31. S. Haykin, *Neural Networks: A Comprehensive Foundation*, MacMillan, 1994.
32. B. G. Sumpter, C. Getino, and D. W. Noid, *Theory and applications of neural computing in chemical science* Ann. Rev. Phys. Chem. **45**, 439, 1994.
33. J. Zupan, and J. Gasteiger, *Neural Networks: A new method for solving chemical problems or just a passing phase?* Anal. Chim. Acta **248**, 1, 1991.
34. M. T. Spining , J. A. Darsey , B. G. Sumpter, and D. W. Nold, *Opening Up the black box of artificial neural networks* J. Chem. Edu. **71**, 406, 1994.
35. J. Zupan and J. Gasteiger, *Neural Networks in Chemistry and Drug Design: An Introduction*, Wiley-VCH, Weinheim, 1999.
36. G. Cybenko, *Approximation by superpositions of a sigmoidal function* Math. Contr. Sign. Sys. **2**, 303, 1989.
37. K. Hornik, M. Stinchcombe, and H. White, *Multilayer feedforward Networks are universal approximators* Neural Networks **2**, 359, 1989.
38. C. Munoz-Caro, and A. Nino, *Neural modeling of torsional potential hypersurfaces in non-rigid molecules* Comp. Chem. **22**, 355, 1998.
39. H. M. Le, and L. M. Raff, *Molecular dynamics investigation of the bimolecular reaction BeH + $H_2 \rightarrow BeH_2$ + H on an ab initio potential-energy surface obtained using neural network methods with both potential and gradient accuracy determination* J. Phys. Chem. A **114**, 45, 2010.
40. H. Gassner, M. Probst, A. Lauenstein, and K. Hermansson, *Representation of intermolecular potential functions by neural networks* J. Phys. Chem. A **102**, 4596, 1998.
41. S. Lorenz, A. Groß, and M. Scheffler, *Representing high-dimensional potential-energy surfaces for reactions at surfaces by neural networks* Chem. Phys. Lett. **395**, 210, 2004.
42. J. Behler, S. Lorenz, and K. Reuter, *Representing molecule-surface interactions with symmetry-adapted neural networks* J. Chem. Phys. **127**, 014705, 2007.
43. F. V. Prudente, P. H. Acioli, J. J. S. Neto, *The fitting of potential energy surfaces using neural networks: Application to the study of vibrational levels of H_3^+* J. Chem. Phys. **109**, 8801, 1998.
44. D. E. Rumelhart, G. E. Hinton, and R. J. Williams, *Learning representations by back-propagating errors* Nature **323**, 533, 1986.
45. R. Fletcher and C. M. Reeves, *Function minimization by conjugate gradients* Comp. J. **7**, 149, 1964.
46. R. E. Kalman, *A new approach to linear filtering and prediction problems* J. Basic Eng. **82**, 35, 1960.
47. T. B. Blank, and S. D. Brown, *Adaptive, global, extended Kalman filters for training feed-forward neural networks* J. Chemometrics **8**, 391, 1994.
48. S. E. Fahlman, and C. Lebiere, *The cascade-correlation learning architectures* Adv. Neural Inf. Proc. Sys. **2**, 524, 1990.

49. K. A. Gernoth, J. W. Clark, J. S. Prater and H. Bohr, *Neural network models of nuclear systematics* Phys. Lett. B **300**, 1, 1993.
50. R. Dawes, D.L. Thompson, A. F. Wagner, and M. Minkoff, *Interpolating moving least-squares methods for fitting potential energy surfaces: A strategy for efficient automatic data point placement in high dimensions* J. Chem. Phys. **128**, 084107, 2008.
51. L. M. Raff, M. Malshe, M. Hagan, D. I. Doughan, M. G. Rockley, and R. Komanduri, *Ab initio potential-energy surfaces for complex, multichannel systems using modified novelty sampling and feedforward neural networks* J. Chem. Phys. **122**, 084104, 2005.
52. N. Artrith, and J. Behler, *Neural network potentials for metal surfaces: A prototype study for copper* submitted , , 2011.
53. J. Behler, and M. Parrinello, *Generalized neural-network representation of high-dimensional potential-energy surfaces* Phys. Rev. Lett. **98**, 146401, 2007.
54. J. Behler, *Atom-centered symmetry functions for constructing high-dimensional neural network potentials* J. Chem. Phys. **134**, 074106, 2011.
55. J. Behler, R. Martoňák, D. Donadio, and M. Parrinello *Metadynamics simulations of the high-pressure phases of silicon employing a high-dimensional neural network potential* Phys. Rev. Lett. **100**, 185501, 2008.
56. J. Behler, R. Martoňák, D. Donadio, and M. Parrinello, *Pressure-induced phase transitions in silicon studied by neural network-based metadynamics simulations* Phys. Stat. Sol. b **245**, 2618, 2008.
57. R. Z. Khaliullin, H. Eshet, T. D. Kühne, J. Behler, and M. Parrinello, *Graphite-diamond phase coexistence study employing a neural-network mapping of the ab initio potential energy surface* Phys. Rev. B **81**, 100103, 2010.
58. R. Z. Khaliullin, H. Eshet, T. D. Kühne, J. Behler, and M. Parrinello, *Nucleation mechanism for the direct graphite-to-diamond phase transition* Nature Materials **10**, 693, 2011.
59. H. Eshet, R. Z. Khaliullin, T. D. Kühne, J. Behler, and M. Parrinello, *Ab initio quality neural-network potential for sodium* Phys. Rev. B **81**, 184107, 2010.
60. S. Houlding, S. Y. Liem, and P. L. A. Popelier, *A polarizable high-rank quantum topological electrostatic potential developed using neural networks: Molecular dynamics simulations on the hydrogen fluoride dimer* Int. J. Quant. Chem. **107**, 2817, 2007.
61. M. G. Darley, C. M. Handley, P. L. A. Popelier, *Beyond point charges: Dynamic polarization from neural net predicted multipole moments* J. Chem. Theory Comp. **4**, 1435, 2008.
62. N. Artrith, T. Morawietz, and J. Behler, *High-dimensional neural-network potentials for multicomponent systems: Applications to zinc oxide* Phys. Rev. B **83**, 153101, 2011.
63. P. P. Ewald, *Die Berechnung optischer und elektrostatischer Gitterpotentiale* Ann. Phys. **64**, 253, 1921.
64. S. Hobday, R. Smith, and J. Belbruno, *Applications of neural networks to fitting interatomic potential functions* Modelling Simul. Mater. Sci. Eng. **7**, 397, 1999.
65. S. Hobday, R. Smith, and J. BelBruno, *Application of genetic algorithms and neural networks to interatomic potentials* Nucl. Instr. and Meth. Phys. Res. B **153**, 247, 1999.

66. A. Bholoa, S. D. Kenny, and R. Smith, *A new approach to potential fitting using neural networks* Nucl. Instr. Meth. Phys. Res. B **255**, 1, 2006.
67. E. Sanville, A. Bholoa, R. Smith, and S. D. Kenny, *Silicon potentials investigated using density functional theory fitted neural networks* J. Phys. Condens. Matter **20**, 285219, 2008.
68. E. Tafeit, W. Estelberger, R. Horejsi, R. Moeller, K. Oettl, K. Vrecko, and G. Reibnegger, *Neural networks as a tool for compact representation of ab initio molecular potential energy surfaces* J. Mol. Graphics **14**, 12, 1996.
69. D. F. R. Brown, M. N. Gibbs, and D. C. Clary, *Combining ab initio computations, neural networks, and diffusion Monte Carlo: An efficient method to treat weakly bound molecules* J. Chem. Phys. **105**, 7597, 1996.
70. K. T. No, B. H. Chang, S. Y. Kim, M. S. Jhon, and H. A. Scheraga, *Description of the potential energy surface of the water dimer with an artificial neural network* Chem. Phys. Lett. **271**, 152, 1997.
71. F. V. Prudente, and J. J. S. Neto, *The fitting of potential energy surfaces using neural networks. Application to the study of the photodissociation processes* Chem. Phys. Lett. **287**, 585, 1998.
72. K. W. Cho, K. T. No, and H. A. Scheraga, *A polarizable force field for water using an artificial neural network* J. Mol. Struct. **641**, 77, 2002.
73. T. M. Rocha Filho, Z. T. Oliveira, Jr., L. A. C. Malbouisson, R. Gargano, and J. J. Soares Neto, *The use of neural networks for fitting potential energy surfaces: A comparative case study for the H_3^+ molecule* Int. J. Quant. Chem. **95**, 281, 2003.
74. A. C. P. Bittencourt, F. V. Prudente, and J. D. M.Vianna, *The fitting of potential energy and transition moment functions using neural networks: transition probabilities in OH (A2sigma+ to X2pi)* Chem. Phys. **297**, 153, 2004.
75. P. M. Agrawal, L. M. Raff, M. T. Hagan,and R. Komanduri, *Molecular dynamics investigations of the dissociation of SiO_2 on an ab initio potential-energy surface obtained using neural network methods* J. Chem. Phys. **124**, 124306, 2006.
76. S. Manzhos, X. Wang, R. Dawes, and T. Carrington, Jr., *A nested molecule-independent neural network approach for high-quality potential fits* J. Phys. Chem. A **110**, 5295, 2006.
77. S. Manzhos, and T. Carrington, Jr., *Using neural networks to represent potential surfaces as sums of products* J. Chem. Phys. **125**, 194105, 2006.
78. M. Malshe, R. Narulkar, L. M. Raff, M. Hagan, S. Bukkapatnam, and R. Komanduri, *Parametrization of analytic interatomic potential functions using neural networks* J. Chem. Phys. **129**, 044111, 2008.
79. H. M. Lee, and L. M. Raff, *Cis $\rightarrow$ trans, trans $\rightarrow$ cis isomerizations and N-O bond dissociation of nitrous acid (HONO) on an ab initio potential surface obtained by novelty sampling and feed-forward neural network fitting* J. Chem. Phys. **128**, 194310, 2008.
80. M. Malshe, L. M. Raff, M. G. Rockley, and M. Hagan, *Theoretical investigation of the dissociation dynamics of vibrationally excited vinyl bromide on an ab initio potential-energy surface obtained using modified novelty sampling and feedforward neural networks. II. Numerical application of the method* J. Chem. Phys. **127**, 134105, 2007.

81. A. Pukrittayakamee, M. Malshe, M. Hagan, L. M. Raff, R. Narulkar, S. Bukkapatnum, and R. Komanduri, *Simultaneous fitting of a potential-energy surface and its corresponding force fields using feedforward neural networks* J. Chem. Phys. **130**, 134101, 2009.
82. H. M. Le, S. Huynh, and L. M. Raff, *Molecular dissociation of hydrogen peroxide (HOOH) on a neural network ab initio potential surface with a new configuration sampling method involving gradient fitting* J. Chem. Phys. **131**, 014107, 2009.
83. T. B. Blank, S. D. Brown, A. W. Calhoun, and D. J. Doren, *Neural network models of potential energy surfaces* J. Chem. Phys. **103**, 4129, 1995.
84. J. Behler, B. Delley, S. Lorenz, K. Reuter, and M. Scheffler, *Dissociation of O_2 at Al(111): The role of spin selection rules* Phys. Rev. Lett. **94**, 36104, 2005.
85. J. Behler, K. Reuter, and M. Scheffler, *Nonadiabatic effects in the dissociation of oxygen molecules at the Al(111) surface* Phys. Rev. B **77**, 115421, 2008.
86. S. Lorenz, M. Scheffler, and A. Groß, *Descriptions of surface chemical reactions using a neural network representation of the potential-energy surface* Phys. Rev. B **73**, 115431, 2006.
87. J. Ludwig, and D. G. Vlachos, *Ab initio molecular dynamics of hydrogen dissociation on metal surfaces using neural networks and novelty sampling* J. Chem. Phys. **127**, 154716–, 2007.
88. D. A. R. S. Latino, R. P. S. Fartaria, F. F. M. Freitas, J. Aires-de-Sousa, and F. M. S. S. Fernandes, *Mapping potential energy surfaces by neural networks: The ethanol/Au(111) interface* J. Electroanal. Chem. **624**, 109, 2008.

Large-Scale Molecular Dynamics Studies and Scale-Briding Models for Deformation and Failure of Materials

Alexander Hartmaier

Interdisciplinary Centre for Advanced Materials Simulation
Ruhr-Universität Bochum, Germany
E-mail: alexander.hartmaier@rub.de

Materials science faces a big challenge due to the different length and time scales that need to be considered. Materials are typically used within dimensions of micrometers to meters and for periods in time ranging from minutes to years. Yet, on a fundamental scale, material behavior is dominated by the electronic structure that is responsible for the interatomic bonding. Processes on the electronic or atomic level occur on length scales of Angstroms and on time scales of femtoseconds or below. Hence, models are needed that bridge the scales from the fundamental physical scales to the engineering scales on which materials are applied. In this contribution illustrative examples are given how such scale-bridging can be accomplished by either large-scale molecular dynamics methods that yield important information on critical deformation and failure mechanisms or by quantifying material specific parameters that can be directly used in macroscopic models.

1 Introduction

From a physics point-of-view mechanical behavior of materials is determined by interatomic bonds that decide for example whether a crack in a material will grow in a brittle manner or whether it will blunt by plastic deformation under a given loading situation. However, materials science teaches us that such local models are not appropriate to predict material behavior in a reliable way, because a crack moving in a brittle manner can be stopped by a grain boundary or a ductile phase in the material, such that the global, i.e. observable behavior might be quite different from the local behavior. Hence, it is important to understand the interplay of microstructure and material properties across different length scales to identify and understand the *critical* deformation and failure mechanisms that are taking place during testing and application of materials. Developing such microstructure-property relationships is an important task in material science and the classical models are described in many textbooks, see for example Ref. 1, 2.

Another important aspect of mechanical material behavior that must be kept in mind in particular when applying atomistic methods is that the mechanical energy stored in a material is directly proportional to its volume. Since the stored mechanical energy represents the driving force for all deformation and failure mechanisms this implies that in the small volumes typically investigated with atomistic methods not necessarily all mechanisms can take place in the same way they would occur in larger volumes, where more mechanical energy is available. This underlines the necessity of large-scale atomistic simulations when two or more competing mechanisms are studied. However, in cases when only one mechanisms is studied, simulations with only a small number of atoms can already provide useful information.

Applying large-scale atomistic simulations to understand and predict material behavior has its own challenges, because the number of degrees of freedom easily approaches hundreds of millions or even billions. Such large amounts of data cannot be stored in every time step of the simulation in order to enable a classical post-processing of the data. This means that during such large-scale simulation the analysis of the atomistic system has to be conducted on-the-fly, i.e. during the simulation such that only the derived quantities need to be stored. Only in this way the data management can be handled in an efficient way.

In the first part of this contribution we will describe how large-scale molecular dynamics simulation can be used to study plastic deformation of materials on a very fundamental level. Examples will illustrate how such simulations help us to understand the interplay between dislocation nucleation, dislocation propagation and work hardening during nanoindentation and which methods can be employed in such studies. In the second part, we will illustrate how fundamental data from smaller atomistic calculations provides the basis for scale-bridging models in material science that help us to understand material behavior and thus to make it predictable. It is noted here that this text does not attempt to give an exhaustive overview on the existing literature, but it rather selects some references that deem to be illustrative.

2 Large-Scale Molecular Dynamics Simulations

Plastic deformation of a crystal is caused by the motion of dislocations. In metals there is typically a high density of dislocations that will move at a certain level of applied stress, called the yield strength (for initial yielding) or flow stress (during plastic deformation). However, when deforming brittle materials or very small volumes there can be a lack of mobile dislocations such that dislocations need to be nucleated within the crystal before plastic deformation can occur. Nanoindentation is an example where such dislocation nucleation is experienced and it expresses itself in form of a so-called pop-in behavior[3–5]. While it is now widely accepted that the nucleation of the first dislocations occurs at the start of the pop-in event frequently observed in nanoindentation experiments, it is unclear how these initial dislocations multiply during the early stages of plastic deformation and produce pop-in displacements that are typically much larger than the magnitude of the Burgers vector. This uncertainty about the complex interplay between dislocation multiplication and strain hardening during nanoindentation makes a direct correlation between force-displacement curves and macroscopic material properties difficult. Recent nanoindentation experiments in single crystals of copper or aluminum revealed large deviations in the lattice rotation and an inhomogeneous distribution of the dislocation density in the plastic zone under the indenter tip[6,7].

Molecular dynamics (MD) simulations offer the possibility to study the origin of these phenomena on an atomistic scale. The only assumption that needs to be made is the interatomic potential that defines the stable lattice structure and how this lattice deforms. From this potential the forces acting on all atoms are calculated and consequently Newton's equation of motion is solved yielding the trajectories for all atoms. A detailed description of the MD method can be found for example in textbooks[8]. As discussed above such MD simulations provide fundamental insight into critical deformation and failure mechanisms if the studied volumes are sufficiently large. However, such large-scale MD simulations require sophisticated analysis routines in order to deal with the massive amount of gen-

erated data. As an example for an on-the-fly analysis of atomistic data, a skeletonization method to simplify defect structures in atomistic simulations enables the direct observation and quantitative analysis of dislocation nucleation and multiplication processes occurring in the bulk as well as at the surface[9]. An example for the application of this method is given in Fig. 1, where it is seen how the atomic structure of dislocation cores is simplified to a network of geometric lines. Building up on this work, an efficient approach is introduced to characterize the dislocation networks by quantifying Burgers vectors of dislocation segments, local plastic strains and lattice rotations on the timescale of picoseconds and below[10]. This data does not only reveal the evolution of dislocation structures, but it offers the possibility to quantify local dislocation density tensors calculated on an atomic level. By this analysis, the numerical results can be directly compared with experimental data despite of the huge differences in the length scales. This comparison provides useful insights into the active deformation mechanisms during plastic deformation. Currently models are being developed that build a bridge between the atomic scale and continuum descriptions[11].

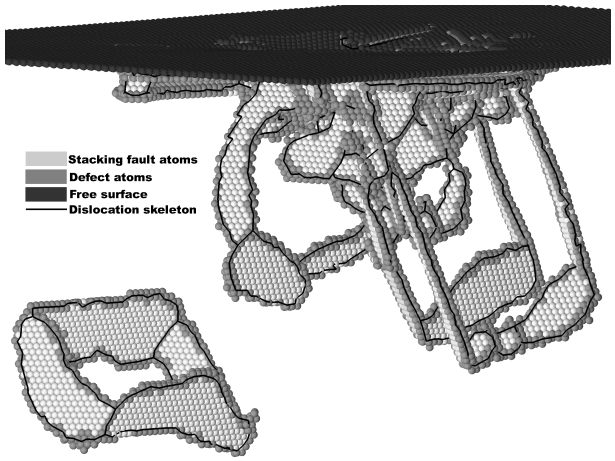


Figure 1. Atomistic defects and the derived dislocation skeleton resulting from a large-scale molecular dynamics simulation of nanoindentation simulation into a copper single crystal.

A further example that illustrates what kind of information on critical deformation and failure mechanisms can be gained from molecular dynamics simulations refers to glassy polymers[12]. Using a united atom model of amorphous polyethylene as generic model system for understanding failure mechanisms in bulk glassy polymers a detailed microscopic understanding of the mechanism of craze initiation has been obtained. To accomplish this molecular dynamics simulations of glassy polymer samples have been performed under different loading conditions. It was found that depending on the loading mode the samples failed by shearing or crazing. The standard models describing the global conditions for shearing or crazing are fulfilled by the numerical samples. A detailed microscopic analysis of internal stresses and non affine deformations within the material allow us to shed some light on the mechanism of craze initiation in the glassy polymer. Under the loading con-

ditions leading to shear failure the internal stresses in the material increase monotonously during the loading. This leads to a stable and homogeneous deformation, because parts that underwent large plastic deformations will only deform further at higher stress levels. However, during failure by crazing, the material's ability to strain harden is compromised, i.e. the internal stress level remains constant during the deformation. Such a constant flow stress causes instabilities and localization of the deformation, which is clearly observed in the numerical simulations, see Fig. 2.

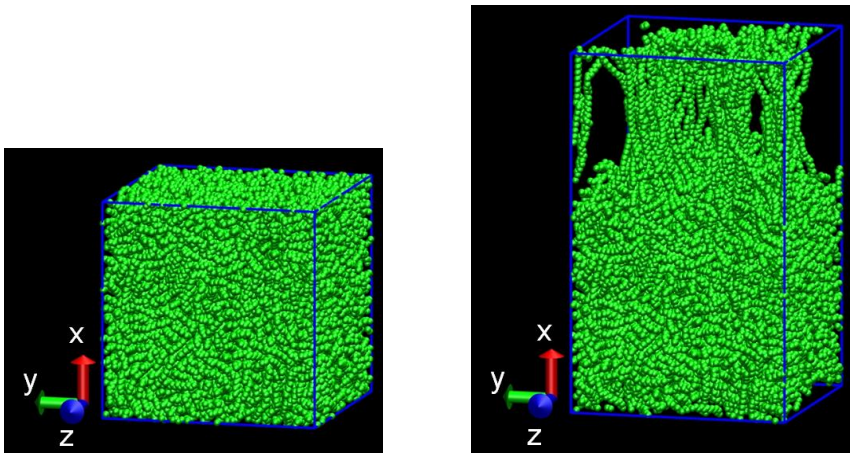


Figure 2. left: Initial atomistic configuration of glassy polymer; right: Configuration after applying tensile strain: The polymer fails by crazing.

3 Scale-Bridging Models

In this second part of the present contribution we assume that the critical deformation and failure mechanisms are known *a priori*. This is for example the case in plastic deformation of metals with a body-centered cubic (bcc) crystal structure, where the mobility of dislocations with parallel Burgers vector and line direction, i.e. so-called screw dislocations, is the lowest of all types of dislocations and hence limits plastic deformation of these metals. In a second example we will study the mechanical properties of grain boundaries in a polycrystal. Such interfaces are known to be weak links, because of the improper atomic bonds, and consequently they are prone to cause deformation and failure. In both examples the critical mechanisms, i.e. mobility of screw dislocations and strength of grain boundaries are quantified by atomistic methods and then this information is used as input for models operating on larger scales.

To develop atomistically informed crystal plasticity models for bcc metals MD studies are used to assess the mobility of screw dislocations, see for example Ref. 13–15 for bcc molybdenum and bcc tungsten or Ref. 16, 17 for bcc iron. It is known that the complicated core structure of screw dislocations in the bcc crystal is the origin of the complex flow behavior in such metals, summarized as non-Schmid behavior, where the mobility of the

dislocations is strongly affected by all components of the stress tensor (see Fig. 3 top). With the help of MD simulations the critical value of the shear stress that is necessary to move a screw dislocation can be calculated as a function of the total stress tensor acting on that dislocation. This quantity is then used as an ingredient for a crystal plasticity models that can be used to reliably describe the deformation of iron and steel on the macro scale with the help of the finite element method (see Fig. 3 bottom).

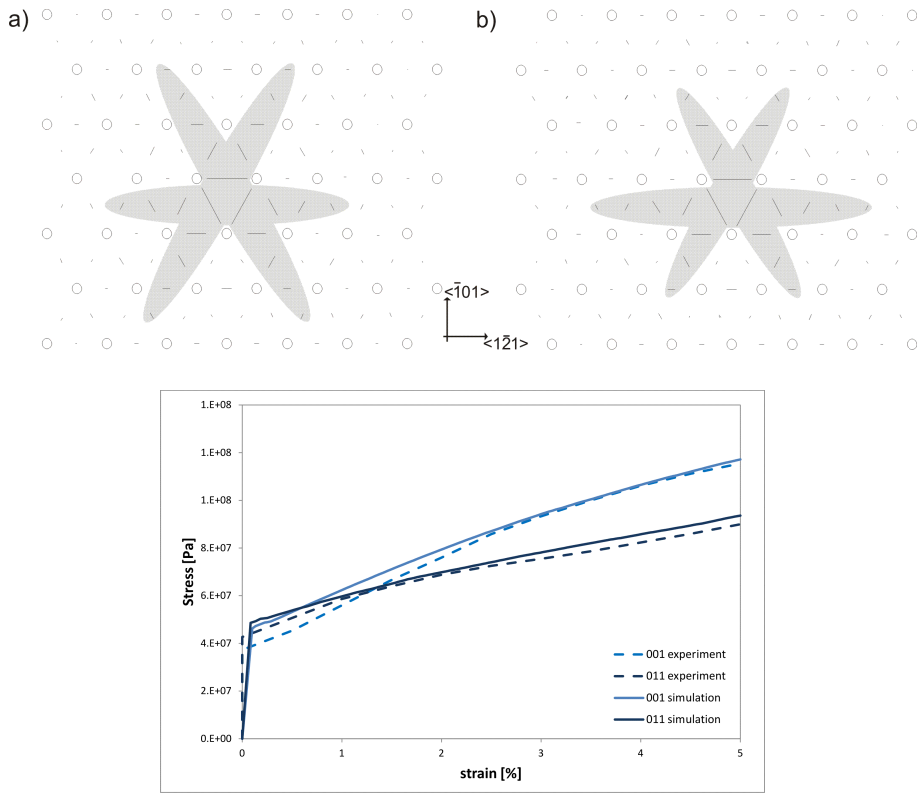


Figure 3. top: Complex atomic structure of a screw dislocation core under a) tensile and b) compressive stresses calculated by the MD method. bottom: Stress-strain curves for iron single crystals in different orientations resulting from a crystal plasticity model partly parameterized by atomistic simulations and fitted to experimental data taken from the literature[18].

As described above the atomic order is disturbed in crystal defects, like grain boundaries. Since the ratio of atoms sitting in the vicinity of grain boundaries to atoms sitting in undisturbed crystal regions gets larger for finer microstructures, the properties of nanostructured materials are widely considered to be controlled by the properties of their interfaces. Hence, internationally a huge amount of scientific and technological effort is devoted towards the investigation and description of the mechanical strength of grain boundaries. The bulk part of these activities has been directed towards modeling and understanding the role of grain boundaries during deformation and fracture of polycrystalline metals. To

accomplish this, electronic structure calculations in the density functional theory (DFT) framework were conducted to calculate grain boundary energies as a function of the grain boundary separation, i.e. a simplified model of grain boundary fracture, for different types of grain boundaries in pure aluminum as a model material[19]. It was found that the energy-separation curves for different types of grain boundaries are characterized very well by the so-called universal binding energy relation (UBER) that was hitherto only applied for interatomic bonds in the bulk[20] (see Fig. 4). This finding resulted in a general formulation for the work of separation of grain boundaries that will simplify further calculations of this quantity that serves as an important input parameter for continuum simulations of fracture processes. Currently we apply this method to quantify the shear strength of grain boundaries.

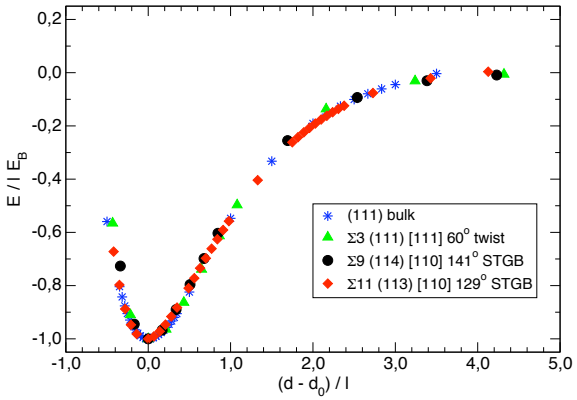


Figure 4. Energy displacement relationships of different types of grain boundaries in aluminum re-scaled according to the UBER relation.

4 Concluding Remarks

In this contribution examples are given that illustrate the role of atomistic methods in materials science. On the one hand, large scale atomistic simulations can be used to study deformation and failure of materials and to understand the critical mechanisms that determine material behavior on larger scales. On the other hand, if the critical mechanisms are known and can be described by physical models, atomistic methods can be used to quantify material specific parameters for these models that can then be used to describe macroscopic material behavior.

Acknowledgments

Financial support through ThyssenKrupp AG, Bayer MaterialScience AG, Salzgitter Mannesmann Forschung GmbH, Robert Bosch GmbH, Benteler Stahl/Rohr GmbH, Bayer Technology Services GmbH and the state of North-Rhine Westphalia as well as the EU in the framework of the ERDF is gratefully acknowledged.

References

1. Günther Gottstein. *Physical Foundations of Materials Science*. Springer-Verlag, Berlin/Heidelberg, 2004.
2. Rob Phillips. *Crystals, Defects and Microstructures*. Cambridge University Press, 2001.
3. S.G. Corcoran, R.J. Colton, E.T. Lilleodden, and W.W. Gerberich. Anomalous plastic deformation at surfaces: Nanoindentation of gold single crystals. *Physical Review B*, 55:R16057–R16060, 1997.
4. H.S. Leipner, D. Lorenz, A. Zeckzer, and P. Grau. Dislocation-related pop-in effect in gallium arsenide. *physica status solidi*, 183(2):R4–R6, 2001.
5. S. Suresh, T.-G. Nieh, and B.W. Choi. Nanoindentation of copper thin films on silicon substrates. *Scripta Materialia*, 41:951–957, 1999.
6. N. Zaafarani, D. Raabe, R. N. Singh, F. Roters, and S. Zaefferer. Three-dimensional investigation of the texture and microstructure below a nanoindent in a Cu single crystal using 3D EBSD and crystal plasticity finite element simulations. *Acta Materialia*, 54(7):1863–1876, 2006.
7. Eralp Demir, Dierk Raabe, Nader Zaafarani, and Stefan Zaefferer. Investigation of the indentation size effect through the measurement of the geometrically necessary dislocations beneath small indents of different depths using EBSD tomography. *Acta Materialia*, 57:559–569, 2009.
8. Daan Frenkel and Berend J. Smit. *Understanding Molecular Simulation*. Academic Press, San Diego, 2004.
9. C. Begau, A. Hartmaier, E. P. George, and G. M. Pharr. Atomistic processes of dislocation generation and plastic deformation during nanoindentation. *Acta Materialia*, 59(3):934–942, 2011.
10. C. Begau, J. Hua, and A. Hartmaier. A novel approach to study dislocation density tensors and lattice rotation patterns in atomistic simulations. *Journal of the Mechanics and Physics of Solids*, 60(4):711 – 722, 2012.
11. P. Engels, A. Ma, and A. Hartmaier. Continuum simulation of the evolution of dislocation densities during nanoindentation. *submitted for publication*, 2012.
12. D. Mahajan and A. Hartmaier. Conditions for craze initiation in glassy polymers revealed by molecular dynamics simulations. *in preparation*, 2012.
13. R. Gröger and V. Vitek. Breakdown of the Schmid law in bcc molybdenum related to the effect of shear stress perpendicular to the slip direction. *Materials Science Forum*, 482:123–126, 2005.
14. R. Groeger, A.G. Bailey, and V. Vitek. Plastic deformation of molybdenum and tungsten: I. Atomistic studies of the core structure and glide of $1/2\langle 111\rangle$ screw dislocations at 0K. *Acta Materialia*, 56:5401–5411, 2008.
15. R. Groeger, V. Racherla, J.L. Bassani, and V. Vitek. Plastic deformation of molybdenum and tungsten: II. Yield criterion for bcc metals involving the non-Schmid behavior of dislocations. *Acta Materialia*, 56:5412–5425, 2008.
16. A. Köster, A. Ma, and A. Hartmaier. Atomistically informed continuum model for body centered cubic iron. *MRS Online Proceedings Library*, 1296:mrsf10–1296–o06–6, 2011.
17. A. Köster, A. Ma, and A. Hartmaier. Atomistically informed crystal plasticity model for body centered cubic iron. *submitted for publication*, 2012.

18. A.S. Keh. Work hardening and deformation sub-structure in Iron single crystal deformed in tension at $298\,^{\circ}\mathrm{C}$. *Philosophical Magazine*, 12:9–30, 1964.
19. Rebecca Janisch, Naveed Ahmed, and Alexander Hartmaier. Ab initio tensile tests of Al bulk crystals and grain boundaries: universality of mechanical behavior. *Physical Review B*, 81:184108–1–6, 2010.
20. J. H. Rose, John Ferrante, and John R. Smith. Universal binding energy curves for metals and bimetallic interfaces. *Physical Review Letters*, 47(9):9, 1981.

Exploration of Multi-Dimensional Free Energy Landscapes in Molecular Dynamics

Mark E. Tuckerman[1,2]

[1] Department of Chemistry
New York University, New York, NY 10003, USA

[2] Courant Institute of Mathematical Sciences
New York University, New York, NY 10003, USA
E-mail: mark.tuckerman@nyu.edu

One of the computational grand challenge problems is the development of methodology capable of sampling conformational equilibria in systems with rough energy landscapes. If met, many important problems, most notably biomolecular structure prediction and the discovery of the polymorphs of organic molecular crystals could be significantly impacted. In this lecture, several new approaches for enhancing sampling and mapping the potential of mean force or free energy of systems with rough potential energy surfaces in terms of a small set of collective variables will be discussed. These include adiabatic dynamics, dynamical spatial warping, and large time-step, resonant-free molecular dynamics. First, we will show how temperature acceleration techniques combined with mass tensor dynamics can be used to predict multi-dimensional free energy surfaces in a small set of collective variables, and the approach will be shown to enhance sampling in a variety of simple biomolecular systems. A related approach will also be shown to enhance the sampling of the space of polymorphs of molecular crystals using the cell matrix as a set of collective variables. Finally, we will discuss the problem of resonance in multiple time-step molecular dynamics and how this problem limits the large time step. A resonant-free approach will then be introduced that permits outer time steps as large as 100 fs in all-atom simulations.

1 Introduction

The free energy difference associated with changes in conformation or thermodynamic state of a complex system is a key quantity in thermodynamics. Free energy differences are important for determining equilibrium constants, rates of processes, reversible work, and a variety of other thermodynamic variables. Molecular dynamics (MD) is a useful tool for calculating such free energy differences in systems of relevance to biology and material science. Generally MD simulation times are short compared with typical biological processes like protein folding, although several recent studies have been able to achieve such time scales[1–3]. Consequently, it is crucial to enhance sampling of configuration space over the course of a simulation. Many enhanced sampling algorithms require as input a set of collective variables (CVs), which are functions of the primitive atomic Cartesian coordinates of the system. These functions describe slow motions that are particularly important during a particular conformational change. These variables can also be used as the basis for coarse-graining a problem. Suppose there are n such variables, where n is small compared to the total number of degrees of freedom in a system. Let us denote the collective variables as $q_1(\mathbf{r}), ..., q_n(\mathbf{r})$, where $\mathbf{r}$ represents the full set of Cartesian coordinates with conjugate momenta $\mathbf{p}$. The free energy surface, which is also the rigorous basis for coarse

graining, is given by

$$A(s_1, ..., s_n) = -kT \ln \left(\int d\mathbf{r} e^{-\beta U(\mathbf{r})} \prod_{\alpha=1}^{n} \delta(q_\alpha(\mathbf{r}) - s_\alpha) \right) \quad (1)$$

where $U(\mathbf{r})$ is the interaction potential in the system, $\beta = 1/kT$, and the δ-functions restrict the configurational integrations to the intersection of the hypersurfaces defined by the n conditions $q_\alpha(\mathbf{r}) = s_\alpha$. This, then, yields the potential of mean force $A(s_1, ..., s_n)$ as a function of the coarse-grained variables $s_1, ..., s_n$. Note that the full canonical partition function $Q(N, V, T)$ of the system is given by

$$Q(N, V, T) \propto \int ds_1 \cdots ds_n e^{-\beta A(s_1, ..., s_n)} \quad (2)$$

Free Energy Perturbation[4], umbrella sampling[5,6] and thermodynamic integration[7–9] are two popular methods to map a free energy profile along one collective variable. For generation of multi-dimensional free energy surfaces (FESs) with respect to several collective variables, methods such as Adiabatic Free Energy Dynamics (AFED)[10–12], adaptive biasing force(ABF)[13,14], and metadynamics[15–17] (Conformational Flooding[18], Local Elevation[19]) have proved highly successful.

The AFED method, developed by Rosso *et al.*[10–12], is a specially designed dynamical scheme for the CVs that imposes an adiabatic decoupling between CVs and the remainder of the system. In addition, the CVs are maintained at a temperature sufficiently high that any barriers along the CV directions on the free energy surface can be easily crossed. This method performs well for simple geometrical collective variables such as distance and dihedral angles. However, as AFED requires a transformation to a coordinate system in which the CVs are explicit variables, the applicability of AFED is limited to simple geometrical CVs. An improved version of this approach was developed by Maragliano and Vanden-Eijnden (called Temperature-accelerate Molecular Dynamics or TAMD)[20] and Abrams and Tuckerman (called driven AFED or d-AFED)[21]. In the TAMD/d-AFED scheme, the CVs are harmonically coupled to a set of extended phase-space variables, and the adiabatic and high-temperature conditions are applied to these extended variables. In this way, variable transformations are completely avoided, thus allowing free energy surfaces in CVs of any mathematical form to be generated straightforwardly. In the first part of the lecture, these approaches and their applications will be discussed.

In the second part of the lecture, the problem of increasing the time step in molecular dynamics calculations will be discussed. Although molecular dynamics is a powerful tool in biomolecular simulations, the technique provides insufficient sampling to impact studies of the 200-300 residue proteins of greatest interest. One severe limitation of molecular dynamics is that the integrators, particularly multiple time-step integrators, are restricted by resonance phenomena to small time steps ($\Delta t < 8$ fs) much slower than the time scales of important structural and solvent rearrangements. The term "resonance" here describes the coupling between different time scales that causes the time step required for fast motion to limit that which can be used for slow motions. Thus, in this part of the lecture, a novel set of equations of motion and a reversible, resonance-free, integrator will be introduced that permit step sizes on the order of 100 fs to be used.

Finally, the last part of the lecture will describe a much more aggressive sampling approach will be discussed. This method is known as the reference potential spatial warping

algorithm or (REPSWA). REPSWA works by introducing a variable transformation in the classical partition function that reduces the volume of phase space associated with a priori known barrier regions while increasing that associated with attractive basins. In this way, the partition function is preserved so that enhanced sampling is achieved without the need for reweighting phase-space averages. Here, a new class of transformations, designed to overcome the barriers induced by intermolecular/nonbonded interactions, whose locations are *not* known *a priori*, is introduced. The new transformations are designed to work in synergy with transformations originally introduced for overcoming intramolecular barriers. The new transformation adapts to the fluctuating local environment and is able to handle barriers that arise "on the fly." Thus, the new method is referred to as dynamic contact REPSWA (DC-REPSWA). In addition, combining hybrid Monte Carlo (HMC) with DC-REPSWA allows more aggressive sampling to take place. The combined DC-REPSWA-HMC method and its variants are shown to substantially enhance conformational sampling in long molecular chains composed of interacting single beads and beads with branches. The latter topologies characterize the united residue and united side chain representation of protein structures.

2 Adiabatic Free Energy Molecular Dynamics and Temperature-Accelerated Molecular Dynamics

Consider a classical system of N particles with positions $\mathbf{r}_1, ..., \mathbf{r}_N \equiv \mathbf{r}$, momenta $\mathbf{p}_1, ..., \mathbf{p}_N \equiv \mathbf{p}$, and masses $m_1, ..., m_N$. The Hamiltonian is taken to be of the usual form.

$$H(\mathbf{r}_1, \cdots, \mathbf{r}_N, \mathbf{p}_1, \cdots, \mathbf{p}_n) = \sum_{i=1}^{N} \frac{\mathbf{p}_i^2}{2m_i} + U(\mathbf{r}_1, \cdots, \mathbf{r}_N), \tag{3}$$

where $U(\mathbf{r}_1, \cdots, \mathbf{r}_N) \equiv U(\mathbf{r})$ is the potential energy. Suppose we are able to identify a set of n collective variables $q_\alpha(\mathbf{r}), \alpha = 1, 2, \cdots, n$ that characterize some process of interest. As noted in the Introduction, the potential of mean force surface is given by Eq. 1:

$$e^{-\beta A(\mathbf{s})} = \int \mathrm{d}\mathbf{p}\mathrm{d}\mathbf{r} \exp\left\{-\beta\left[\sum_{i=1}^{N} \frac{\mathbf{p}_i^2}{2m_i} + U(\mathbf{r})\right]\right\} \prod_{\alpha=1}^{n} \delta\left(q_\alpha(\mathbf{r}) - s_\alpha\right), \tag{4}$$

where $\mathbf{s} = (s_1, s_2, \cdots, s_n)$ $A(\mathbf{s})$ is the free energy surface (FES) of the physical system when $q_\alpha(\mathbf{r}) = s_\alpha$. The constant k is the Boltzmann constant, T is the temperature of the physical system, and $\beta = 1/kT$.

The product of δ-functions imposes a condition on the configuration space that we sample only the intersection of the hypersurfaces represented by $q_\alpha(\mathbf{r}) = s_\alpha$. In principle, this would be handled via a set of constraints[8,9] with an appropriate unbiasing factor for corresponding constraints on the momentum space obtained by the additional condition $\dot{q}_\alpha(\mathbf{r}) = 0$, however, such a scheme is practical only for $n = 1$ (or, at great computational overhead, $n = 2$) because of the exponential dependence on n of the number of constraint values needed to generate the FES. Thus, in order to avoid this problem, consider replacing the δ-functions by the limit of a product of Gaussians:

$$\prod_{\alpha=1}^{n} \delta\left(q_\alpha(\mathbf{r}) - s_\alpha\right) = \left[\prod_{\alpha=1}^{n} \lim_{\kappa_\alpha \to \infty} \sqrt{\frac{\beta\kappa_\alpha}{2\pi}}\right] \exp\left\{-\sum_{\alpha=1}^{n} \frac{\beta}{2}\kappa_\alpha \left(q_\alpha(\mathbf{r}) - s_\alpha\right)^2\right\}, \tag{5}$$

Substituting this into Eq. 4, we obtain

$$e^{-\beta A(\mathbf{s})} = \left[\prod_{\alpha=1}^{n} \lim_{\kappa_\alpha \to \infty} \sqrt{\frac{\beta\kappa_\alpha}{2\pi}}\right] \times \int \mathrm{d}\mathbf{p}\mathrm{d}\mathbf{r} \exp\left\{-\beta\left[\sum_{i=1}^{N} \frac{\mathbf{p}_i^2}{2m_i} + U(\mathbf{r}) + \frac{1}{2}\sum_{\alpha=1}^{n} \kappa_\alpha \left(q_\alpha(\mathbf{r}) - s_\alpha\right)^2\right]\right\}, \quad (6)$$

The form of Eq. 6 suggests that we can map out the free energy surface by sampling the centers of the Gaussian functions $s_1, ..., s_n$ directly by introducing them as extended phase-space variables with conjugate momenta $p_{s_1}, ..., p_{s_n}$. In this way, we set up an extended dynamical system that can be simulated using molecular dynamics. The contribution from the Gaussians can be viewed as an additional harmonic potential that couples the physical system to the extended system. This approximation becomes accurate when the parameters κ_α become infinite. In general, due to the harmonic coupling, the CVs will follow the coordinates of extended variables. The whole system, including the real and extended components, is described by a new Hamiltonian:

$$\begin{aligned} H_{\mathrm{ex}}(\mathbf{r};\mathbf{s}) &= \sum_{\alpha=1}^{n} \frac{p_{s_\alpha}^2}{2m_\alpha} + \sum_{i=1}^{N} \frac{\mathbf{p}_i^2}{2m_i} + U(\mathbf{r}) + \sum_{\alpha=1}^{n} \frac{1}{2}\kappa_\alpha \left(q_\alpha(\mathbf{r}) - s_\alpha\right)^2. \\ &\equiv \sum_{\alpha=1}^{n} \frac{p_{s_\alpha}^2}{2m_\alpha} + \sum_{i=1}^{N} \frac{\mathbf{p}_i^2}{2m_i} + \tilde{V}(\mathbf{r};\mathbf{s}), \end{aligned} \quad (7)$$

where

$$\tilde{V}(\mathbf{r};\mathbf{s}) = U(\mathbf{r}) + \sum_{\alpha=1}^{n} \frac{1}{2}\kappa \left(q_\alpha(\mathbf{r}) - s_\alpha\right)^2. \quad (8)$$

Assuming we have a good choice of CVs that characterize the slow motions of the physical system, we can choose the masses of the extended variables m_α to be large enough such that other degrees of freedom have time to equilibrate while the values of CVs have only slightly changed. That is, we create an adiabatic decoupling between the physical and extended subsystems. Based on this assumption, it is straightforward to show that the extended system evolves under the potential of mean force:

$$V_{\mathrm{mf}}(\mathbf{s}) = -\frac{1}{\beta} \ln Z_1(\mathbf{s}), \quad (9)$$

where

$$Z_1(\mathbf{s}) = \int \mathrm{d}\mathbf{p}\mathrm{d}\mathbf{r} \exp\left\{-\beta\left[\sum_{i=1}^{N} \frac{\mathbf{p}_i^2}{2m_i} + \tilde{V}(\mathbf{r};\mathbf{s})\right]\right\}. \quad (10)$$

Thus, we have an effective Hamiltonian H_{eff} for the extended system, which is given by

$$H_{\mathrm{eff}}(s, p_s) = \sum_{\alpha=1}^{n} \frac{p_{s_\alpha}^2}{2m_\alpha} + V_{\mathrm{eff}}(\mathbf{s}). \quad (11)$$

In order to sample rare transitions among the CVs, the extended variables are coupled to a thermostat set to a high temperature $T_s > T$. This results in a set of equations of motion for the physical+extended system of the form

$$m_i \ddot{\mathbf{r}}_i = -\frac{\partial U}{\partial \mathbf{r}_i} - \sum_{\alpha=1}^{n} \kappa_\alpha \left(q_\alpha(\mathbf{r}) - s_\alpha\right) \frac{\partial q_\alpha}{\partial \mathbf{s}_i} + \text{heat bath}(T)$$

$$m_\alpha \ddot{s}_\alpha = \left(q_\alpha(\mathbf{r}) - s_\alpha\right) + \text{heat bath}(T_s) \tag{12}$$

where two heat bath couplings, which could be any correct canonical thermostat, e.g., Langevin, Nosé-Hoover chains[22], GGMTs[23],..., at temperatures T and T_s have been included. If m_α is large enough, the extended variables are adiabatically decoupled from the physical system, and we maintain different temperatures on the real and extended systems, then it can be rigorously proved (see Appendix) that the probability distribution of extended variables directly gives the free energy surface according to

$$P_{\text{adb}}(\mathbf{s}) \propto \int d^n p_s \; e^{-\beta_s H_{\text{eff}(s,p_s)}} \propto Z_1^{\frac{\beta_s}{\beta}}(\mathbf{s}),$$

$$A(\mathbf{s}, T) \simeq -\frac{1}{\beta} \ln Z_1(\mathbf{s}) + C = -\frac{1}{\beta_s} \ln P_{\text{adb}}(\mathbf{s}). \tag{13}$$

Here, $\beta_s = 1/kT_s$, and C is a an irrelevant constant. In fact, Z can be viewed as the true probability distribution of the extended variables at temperature T. This method is known as driven adiabatic free energy dynamics (d-AFED), also referred to as temperature-accelerated molecular dynamics (TAMD). As $\frac{T_s}{T}$ increases, P_{adb} will become more uniform, which means the probability of sampling rare events increases.

For certain applications, it might be possible to perform a variable transformation to a coordinate system in which the CVs $q_\alpha(\mathbf{r})$ are explicit variables. In this case, Eq. 4 becomes

$$e^{-\beta A(\mathbf{s})} = \int \mathrm{d}\mathbf{p}\mathrm{d}^{3N}q \exp\left\{-\beta\left[\sum_{i=1}^{N} \frac{\mathbf{p}_i^2}{2m_i} + \tilde{U}(q)\right]\right\} \prod_{\alpha=1}^{n} \delta\left(q_\alpha(\mathbf{r}) - s_\alpha\right), \tag{14}$$

where

$$\tilde{U}(q) = U(\mathbf{r}(q)) - kT \ln J(q) \tag{15}$$

with $J(q)$ being the Jacobian of the transformation. If we regard the $3N$ Cartesian momentum components as "conjugate" to $q_1, ..., q_{3N}$, then we can sample the free energy surface in much the same way as is done in the extended phase-space formulation. That is, we introduce a high temperature T_s associated with the first n coordinates $q_1, ..., q_n$ and let the associated masses $m_1, ..., m_n$ be large compared to the remaining $3N - n$ masses, and using these parameters, we perform a molecular dynamics calculation using the Hamiltonian

$$\tilde{H}(\mathbf{p}, q) = \sum_{i=1}^{N} \frac{\mathbf{p}_i^2}{2m_i} + \tilde{U}(q) \tag{16}$$

The adiabatic decoupling ensures the first n coordinates evolve slowly while the high temperature ensures that they are able to cross any barriers creating rare events. Under the high-temperature adiabatic conditions, the probability distribution generated takes the form

$$P_{\rm adb}(s_1,\cdots,s_n) = N\int d^n p\ \exp\left[-\beta_s\sum_{\alpha=1}^{n}\frac{p_\alpha^2}{2m_\alpha}\right][Z(s_1,\cdots,s_n,\beta)]^{T/T_{\rm s}} \qquad (17)$$

where

$$Z(s_1,\cdots,s_n,\beta) = \int d^{3N-n}p d^{3N}q$$

$$\times\exp\left\{-\beta\left[\sum_{\alpha=n+1}^{3N}\frac{p_\alpha^2}{2m_\alpha}+\tilde{U}(q_1,\cdots,q_{3N})\right]\right\}\prod_{\alpha=1}^{n}\delta(q_\alpha-s_\alpha) \qquad (18)$$

and N is an overall normalization factor. This direct formulation of adiabatic dynamics, although not as straightforward to implement, exhibits better convergence properties than the extended phase-space version due to its elimination of the harmonic coupling and the additional "noise" this approximation introduces into the phase-space sampling.

2.1 Predicting Polymorphism in Molecular Crystals

Structural diversity abounds in nature. Researchers in many disciplines are often faced with the considerable challenge of understanding this phenomenon in specific applications and its associated consequences. In chemistry, an area where structural diversity has profound implications is molecular crystals. Small organic molecules can crystallize into a variety of different forms, which gives rise to the phenomenon of *polymorphism*. While polymorphism is important in numerous problems involving molecular crystals, there are few in which the stakes are as high as they are in pharmaceutical applications[24]. Consider, for example, the case of the anti-AIDS drug Ritonavir (a protease inhibitor). When the drug was first launched in 1996, only one crystalline form was known, which was shown to be sufficiently water soluble for therapeutic applications. Subsequent to the launch, however, a second crystalline form, hitherto unknown, began to show up in the manufacturing process, and this new form, being much less water soluble, caused many lots to fail the dissolution test, thereby compromising their bioavailability. Unfortunately, this second form had already infiltrated market supplies, and hence, a massive and costly recall and reformulation was necessary before its rerelease in 2002. Polymorphism in the common heartburn medication Ranitidine hydrochloride lead to an expensive and protracted dispute over separate patents for two different crystalline forms when a generic drug manufacturer claimed that the synthetic procedure of the earlier of the two patents (which had expired) actually yielded the crystal form of the later patent, thereby rendering it invalid. From this and other examples[24], it is clear that *a priori* prediction and thermodynamic ranking of the different crystalline polymorphs of a given compound are important problems in which suitable computational techniques can play an important role.

Numerous theoretical methods have been developed for the crystal structure prediction[25]. However, very few of these are based on free energy sampling[26–28]. The theoretical challenge of exploring polymorphism in molecular crystals stems from the requirement of sampling a complex and rough energy landscape in order to obtain free energy differences

between the different polymorphs. Because of this, polymorphism prediction has been compared to the problem of exploring the conformational space of proteins[29]. Recently, we introduced an adaptation of the AFED[30] methodology for the exploration of crystalline polymorphism[27]. We call the approach Crystal-AFED.

Crystal-AFED is an adaptation for the isothermal-isobaric (NPT) ensemble of the adiabatic free energy dynamics approach. It employs the $\mathbf{h}$ matrix as a set of collective variables and seeks to map out the Gibbs free energy as a function of $\mathbf{h}$. Crystal-AFED can be applied with any NPT(flex) scheme provided that it correctly generates the isothermal-isobaric ensemble. In our case, the MTK equations are used[31]. As in the AFED scheme, the collective variables $\mathbf{h}$ are assigned a large mass W and a temperature $T_{\mathbf{h}}$ that is higher than the physical temperature T. This is tantamount to writing the MTK equations of motion as

$$\dot{\mathbf{r}}_i = \frac{\mathbf{p}_i}{m_i} + \frac{\mathbf{p_g}}{W'}\mathbf{r}_i$$

$$\dot{\mathbf{p}}_i = \mathbf{F}_i - \frac{\mathbf{p_g}}{W'}\mathbf{p}_i - \frac{1}{N_{\rm f}}\frac{\rm Tr[\mathbf{p_g}]}{\rm W'}\mathbf{p}_i + \text{heat bath}(T)$$

$$\dot{\mathbf{h}} = \frac{\mathbf{p_g}\mathbf{h}}{W'}$$

$$\dot{\mathbf{p}}_{\mathbf{g}} = \det(\mathbf{h})\left[\mathbf{P}^{(\rm int)} - P\mathbf{I}\right] + \frac{1}{N_{\rm f}}\sum_{i=1}^{N}\frac{\mathbf{p}_i^2}{m_i}\mathbf{I} + \text{heat bath}(T_{\mathbf{h}}) \tag{19}$$

where W' is a large mass and $T_{\mathbf{h}} >> T$, and we have generically denoted the thermostat coupling in the equations for $\mathbf{p}_i$ and $\mathbf{p_g}$. Within the Crystal-AFED approach, the Gibbs free energy $G(\mathbf{h}_s, T)$ is obtained from

$$G(\mathbf{h}_s, T) = -kT_{\mathbf{h}} \ln P_{\rm adb}(\mathbf{h}_s, T, T_{\mathbf{h}}) \tag{20}$$

where $P_{\rm adb}(\mathbf{h}_s, T, T_{\mathbf{h}})$ the adiabatic probability distribution for the cell matrix $\mathbf{h}$ to have the value $\mathbf{h}_s$ accumulated during a Crystal-AFED simulation and is given by

$$P_{adb}(\mathbf{h}_s, T, T_{\mathbf{h}}) = \frac{1}{V_0}\int d\mathbf{h}\ \left[\det(\mathbf{h})\right]^{1-d} e^{-\beta_s P \det(\mathbf{h})}\left[Q(N, \mathbf{h}, T)\right]^{T/T_{\rm h}} \delta(\mathbf{h} - \mathbf{h}_s) \tag{21}$$

Thermostatting of the cell matrix within Crystal-AFED requires some care. In contrast to a standard NPT(flex) MD simulation, in which the matrix $\mathbf{p_g}$ is coupled to a single NHC thermostat, Crystal-AFED benefits from a more robust temperature control mechanism. Since $\mathbf{h}$ is the key collective variable in Crystal-AFED, separate control of the temperature of the diagonal and off-diagonal elements of the $\mathbf{p_g}$ matrix enhances the fluctuations in the collective variable space and increases the efficiency of the approach. Thus, we employ a version of massive thermostatting to the cell matrix. In this case, each diagonal element of $\mathbf{p_g}$ is coupled to a separate thermostat, and additional separate thermostats are coupled to each of the three pairs of off-diagonal elements, $(p_{\mathbf{g},12}, p_{\mathbf{g},21})$, $(p_{\mathbf{g},13}, p_{\mathbf{g},31})$, and $(p_{\mathbf{g},23}, p_{\mathbf{g},32})$. To ensure efficient equipartitioning of the elements of the matrix $\mathbf{p_g}$ at the temperature $T_{\mathbf{h}}$, each of the diagonal elements has a target average kinetic energy $kT_{\mathbf{h}}/2$, and each of the three off-diagonal pairs has a target average kinetic energy of $kT_{\mathbf{h}}/2$ (that is, each off-diagonal element individually has an average kinetic energy of $kT_{\mathbf{h}}/4$ as explained in the Appendix). Fig. 1 shows the kinetic energy convergence behavior of $\mathbf{p_g}$ and

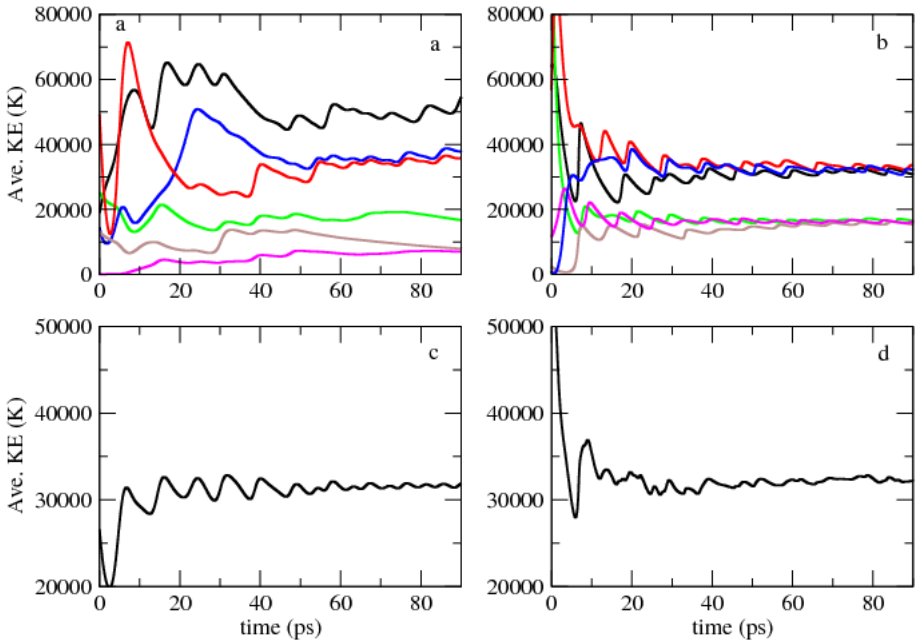


Figure 1. The convergence of the average kinetic energy (KE) of each element of $\mathbf{p_g}$ in a Crystal-AFED simulation. (a), (c): Global thermostat for the barostat. (b), (d): Massive thermostatting the barostat. Black: KE_{11}; Blue: KE_{22}; Red: KE_{33}; Green: $\mathrm{KE}_{12,21}$; Brown: $\mathrm{KE}_{13,31}$; Magenta: $\mathrm{KE}_{23,32}$. The convergence of the average kinetic energy of $\mathbf{p_g}$ in a Crystal-AFED simulation. The system temperature is maintained at 300 K. The atoms are thermostatted massively, and the barostat time scale is 20 ps.

its individual elements under the action of a single global thermostat on $\mathbf{p_g}$ as a whole and under the action of a "massive" thermostat. Under massive thermostatting, the kinetic energy of the elements of $\mathbf{p_g}$ quickly converges (b) while the global thermostat is unable to properly equipartition the kinetic energy among the elements of $\mathbf{p_g}$, although, as expected, the total average kinetic energy of the box reaches the desired target value (c). Therefore, the massive thermostatting strategy is highly recommended for Crystal-AFED simulations. Because of the two temperature scales, we find that the optimal thermostatting scheme for Crystal-AFED is the GGMT thermostat[23], which has been found to be particularly effective for AFED simulations. With the above protocol, Crystal-AFED appears to be an efficient algorithm for the exploration of crystalline polymorphism and for providing a thermodynamic ranking of the polymorphs based on the free energy hypersurface in $\mathbf{h}$, as has been demonstrated for the case of solid benzene[27].

In order to illustrate the performance of Crystal-AFED, we use the same benzene model as in Ref. 27 and carry out a Crystal-AFED simulation using $T_{\mathbf{h}} = 31000$ K and a barostat time scale $\tau = 8.5$ ps. The value of τ is used to determine the barostat mass parameter W' via $W' = (N_f/3 + 1)kT\tau^2$. In Crystal-AFED simulations, the real space summation requires some special attention. Because the cell matrix can occasionally undergo strong distortions, it is necessary to extend the real-space sum beyond the primary simulation cell and include interactions with the first few image cells. However, as this also increases

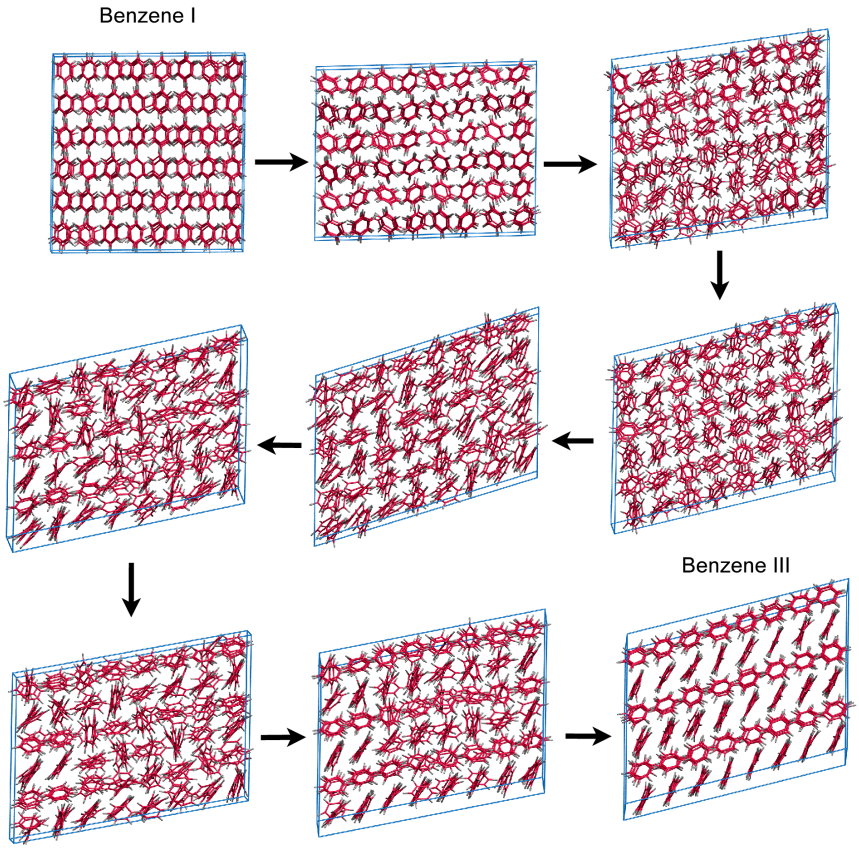


Figure 2. Polymorphic transition induced in Crystal-AFED. 192 molecules are simulated at 300 K with $T_{\mathbf{h}} = 31000$ K.

the cost of the real-space sum, we employed a dynamic replication scheme wherein if the face-to-face distances of the supercell becomes smaller than twice the real-space cutoff, extra replicas are added to the real-space sum. Fig. 2 shows a smooth phase transition generated using Crystal-AFED, in which an initial benzene I structure passes through several amorphous states and transforms into the benzene III structure. In Fig. 3, we show a longer trajectory segment of the cell lengths, angles, and molar volume under Crystal-AFED at $T = 100$ K in which the system visits four of the crystal structures accessible within the Gromos 96 force field[32]. At $T = 100K$, using six randomly initialized trajectories, we are able to visit all of the benzene polymorphs in a minimal time of just 470 ps. Longer trajectories such as that shown in Fig. 3 might become disordered, in which case, they are terminated, or remain in one structure for an extended period of time, which provides useful data for subsequent determination of relative free energies of the polymorphs. When a particular crystal form persists for a significant period of time in the trajectory, we attribute that form with a greater stability. If good sampling of different structures is achieved, then the amount of time spent in each structure can be used to obtain the relative probability

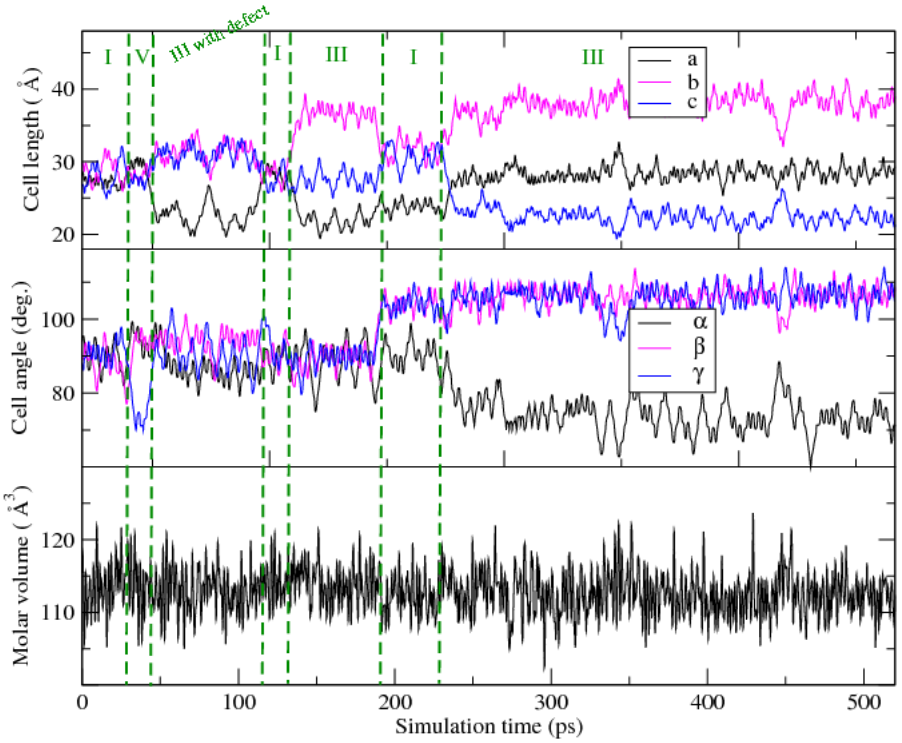


Figure 3. Crystal-AFED trajectories for benzene at 100 K showing a smooth transition from benzene I to benzene III.

of that structure and, therefore, the probability distribution of the different polymorphs. The relative Gibbs free energies of the polymorphs are obtained from this distribution. In Ref. 27, we reported these relative free energies as well as their space groups and unit-cell structures. Note that, as Fig. 3 suggests, there can be some variation in the cell lengths and angles associated with a given polymorph. From the trajectory, we were able to determine the free energy difference between different polymorphs of benzene, which is shown in Fig. 4. From Fig. 4, we see that the free energy and lattice energy measures are comparable and result in a consistent stability ordering except for phase II98. II98 is often found in a mixed stacking structure (see SI), which is essentially a phase III with line defects. The free energy analysis shows that mixed stacking structures have a considerable thermodynamic stability while the stability of pure benzene II (denoted II98) is low despite its low lattice energy. This result addresses a long-standing controversy: Crystal-AFED predicts that mixed stacking structures are thermodynamically more stable at the simulated conditions, and therefore, what is observed in experiment for phase II is actually such a mixed structure (II01 or III with a defect) rather than II98[33]. Indeed, a phase III with a stacking defect has the closest powder diffraction pattern compared to experiment[33].

Benzene IV is found to be stable (or metastable) only at pressures above 5 GPa, and in our simulations, IV was observed to change to I quickly under an ordinary anisotropic

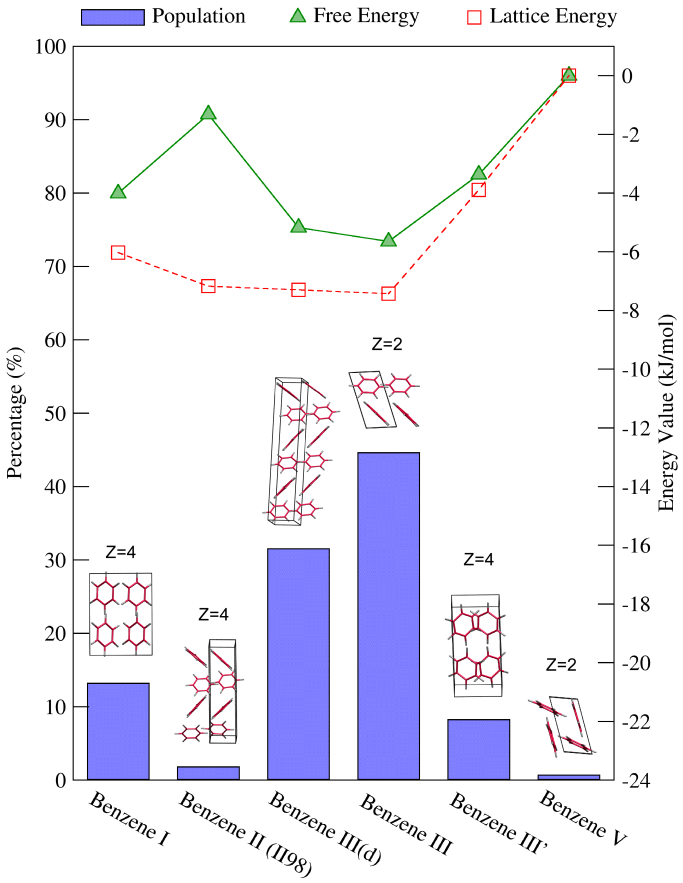


Figure 4. Probability distribution, unit cell structures, and corresponding free energies (with $T_{\mathbf{h}} = 31000$K) and lattice energies of the stable polymorphs of benzene at 100 K and 2 GPa obtained via Crystal-AFED. The unit cell for the mixed stacking structure (III with defect) is one representative structure among all those generated in the simulation. The lattice energy is the sum of the intermolecular energy and the PV contribution at 0 K and 2 GPa, which, for III with a defect, is averaged over several different mixed forms.

NPT simulation. Crystal-AFED trajectories visit the IV structure but remain there for times sufficiently short as to give them negligible contribution to the distribution ($\Delta G > 0$). A typical pathway observed in Crystal-AFED is from benzene I to V with IV appearing as an intermediate state.

Although the Gromos force field is not designed for condensed-phase systems, it gives a reasonably good prediction of benzene polymorphism at pressures in the range $0 \sim 4$ GPa[26]. Our overall conclusion is that at 2 GPa benzene III is the most stable form at 100 K while mixed stacking structures have a comparable stability with III. The latter could be

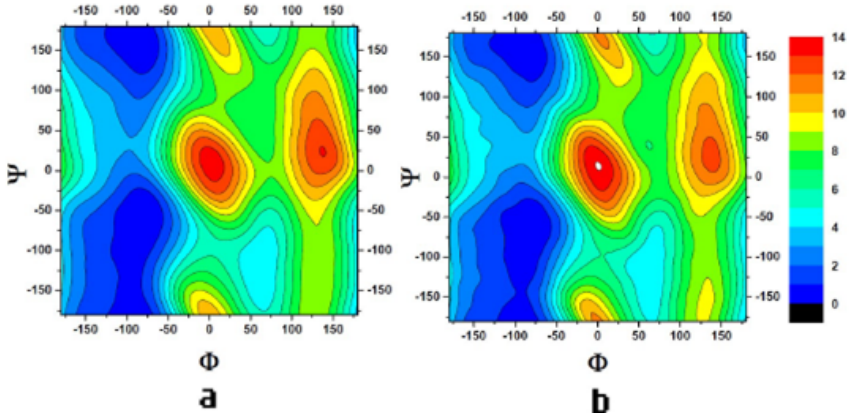


Figure 5. Ramachandran plot of the alanine dipeptide in solution comparing d-AFED (left) with metadynamics (right).

due to the fact that 2 GPa is near the phase transition region under this force field model. Phase I is the third most stable structure at 2 GPa, which is consistent with the phase digram proposed previously[26].

2.2 Other Examples

Fig. 5 shows the Ramachandran plot of the alanine dipeptide in aqueous solution (216 waters) using the CHARMM22 force field obtained using d-AFED with $T_s = 1000$ K. The simulation length is 5 ns, and a comparison with metadynamics[15] (also of length 5 ns) is presented. We see that the comparison is very good between the two method, however, d-AFED does somewhat better in the high free energy regions.

Fig. 6) shows a simulation of met-enkephalin in aqueous solution using d-AFED with $T_s = 600$ K. A run of length 200 ns is carried and compared to metadynamics[15] using 400 ns. The collective variables are the radius of gyration of the heavy atoms and the alpha-helical similarity. We see that the agreement is good between the methods, however, even with 200 ns, d-AFED yields a smoother surface than metadynamics with 400 ns.

3 Long Time-Step Molecular Dynamics

Hamiltonian's such as that in Eq. 7 contains many time scales including one created by the additional harmonic coupling term between the extended phase-space variables and the CVs. In principle, the equations could be integrated using multi time-step techniques[34,35], however, a system of this type will exhibit so-called resonance phenomena[36], which limits the largest time step that can be used.

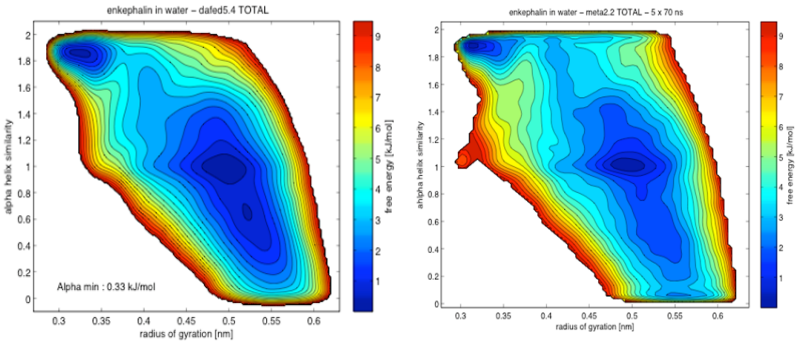


Figure 6. Free energy surface of meta-enkephalin in solution using d-AFED (left) and metadynamics (right).

In order to illustrate resonances, consider a single particle in one spatial dimension with unit mass subject to a harmonic potential of frequency $\omega^2 + \Omega^2$, where $\Omega << \omega$. The Hamiltonian takes the form

$$H = \frac{p^2}{2} + \frac{1}{2}\omega^2 x^2 + \frac{1}{2}\Omega^2 x^2 \tag{22}$$

An integrator for Hamilton's equations of motion can be derived from the Liouville operator

$$iL = p\frac{\partial}{\partial x} - \omega^2 x\frac{\partial}{\partial p} - \Omega^2 x\frac{\partial}{\partial p} \tag{23}$$

Consider separating this operator into two contributions:

$$iL = iL_{\text{fast}} + iL_{\text{slow}} \tag{24}$$

where

$$iL_{\text{fast}} = p\frac{\partial}{\partial x} - \omega^2 x\frac{\partial}{\partial p}$$

$$iL_{\text{slow}} = -\Omega^2 x\frac{\partial}{\partial p} \tag{25}$$

The propagator $\exp(iL\Delta t)$ for a discrete time step Δt, where Δt is chosen to be appropriate for the slow oscillatory motion, can be factorized using the Trotter theorem according to

$$e^{iL\Delta t} = e^{iL_{\text{slow}}\Delta t/2} e^{iL_{\text{fast}}\Delta t} e^{iL_{\text{slow}}\Delta t/2} \tag{26}$$

Applying this propagator on an initial condition $(x(0), p(0))$ to yield numerical solutions $(x(\Delta t), p(\Delta t))$, we can express the solution in the form of a matrix equation

$$\begin{pmatrix} x(\Delta t) \\ p(\Delta t) \end{pmatrix} = A(\omega, \Omega, \Delta t) \begin{pmatrix} x(0) \\ p(0) \end{pmatrix} \tag{27}$$

where $A(\omega, \Omega, \Delta t)$ is a 2×2 matrix of the form

$$A(\omega, \Omega, \Delta t) =$$

$$\begin{pmatrix} \cos(\omega\Delta t) - \frac{\Delta t\Omega^2}{2\omega}\sin(\omega\Delta t) & \frac{1}{\omega}\sin(\omega\Delta t) \\ \left(\frac{\Delta t^2\Omega^4}{4\omega} - \omega\right)\sin(\omega\Delta t) - \Delta t\Omega^2\cos(\omega\Delta t) & \cos(\omega\Delta t) - \frac{\Delta t\Omega^2}{2\omega}\sin(\omega\Delta t) \end{pmatrix} \quad (28)$$

Depending on how large Δt is, we find that $-2 < \mathrm{Tr}(A) < 2$ or $|\mathrm{Tr}(A)| \geq 2$. In the former case, the eigenvalues of A are complex conjugate pairs while in the latter, the eigenvalues of A are both real, which would lead to hyperbolic rather than oscillatory motion. The transition occurs at $|\mathrm{Tr}(A)| = 2$, in which case $\Delta t = n\pi/\omega$, where n is a positive integer. This fact tells us that the large time step cannot be chosen greater than π/ω, suggesting that the fast frequency places a fundamental limit on the time step we can choose for the *slow* motion!

Physical resonances are tantamount to the building up of energy in a particlar mode of motion. We can circumvent this problem by employing an approach in which we constrain the kinetic energy of the system to be a fixed constant. Such an approach is known as *isokinetic molecular dynamics*[37]. The kinetic energy constraint can be imposed using a Lagrange multiplier α. We begin with the equations of motion:

$$\dot{\mathbf{r}}_i = \frac{\mathbf{p}_i}{m_i}$$

$$\dot{\mathbf{p}}_i = \mathbf{F}_i - \alpha\mathbf{p}_i. \quad (29)$$

We now obtain a closed-form expression for the multiplier α. In order to do this, we first differentiate the constraint $\sum_{i=1}^{N} \mathbf{p}_i^2/2m_i = K$ once with respect to time, which yields

$$\sum_{i=1}^{N} \frac{\mathbf{p}_i}{m_i} \cdot \dot{\mathbf{p}}_i = 0. \quad (30)$$

Thus, substituting the second of Eqs. 29 into Eq. 30 gives

$$\sum_{i=1}^{N} \frac{\mathbf{p}_i}{m_i} \cdot \left[\mathbf{F}_i - \alpha\mathbf{p}_i\right], \quad (31)$$

which can be solved for α giving

$$\alpha = \frac{\sum_{i=1}^{N} \mathbf{F}_i \cdot \mathbf{p}_i/m_i}{\sum_{i=1}^{N} \mathbf{p}_i^2/m_i}. \quad (32)$$

When Eq. 32 is substituted into Eq. 29, the equations of motion for the isokinetic ensemble become

$$\dot{\mathbf{r}}_i = \frac{\mathbf{p}_i}{m_i}$$

$$\dot{\mathbf{p}}_i = \mathbf{F}_i - \left[\frac{\sum_{j=1}^{N} \mathbf{F}_j \cdot \mathbf{p}_j/m_j}{\sum_{j=1}^{N} \mathbf{p}_j^2/m_j}\right]\mathbf{p}_i. \quad (33)$$

Because Eqs. 33 were constructed to preserve the constraint, they manifestly *conserve* the kinetic energy; however, that the constraint is also a conservation law of the isokinetic equations of motion can also be verified by direct substitution. The isokinetic equations generate the following partition function:

$$\mathcal{Q}(N,V,T,K) = \frac{K_0}{N!h^{3N}} \int d^N\mathbf{p} \int_{D(V)} d^N\mathbf{r}\, \delta\left(\sum_{i=1}^{N} \frac{\mathbf{p}_i^2}{2m_i} - K\right) e^{-\beta U(\mathbf{r}_1,\ldots,\mathbf{r}_N)}, \quad (34)$$

where K is preset value of the kinetic energy, and K_0 is an arbitrary constant having units of energy. This partition function clearly shows that the configurational part of the partition function is the canonical factor $\exp(-\beta U(\mathbf{r}))$, which is usually all we are interested in.

The isokinetic equations of motion are resonance free. However, they are not sufficiently ergodic to be used on their own as a molecular dynamics sampling algorithm. If, however, we couple them to a standard Nosé-Hoover chain algorithm[22], we obtain an ergodic scheme that can be integrated using multiple time-step techniques[34,35]. We apply the coupling between isokinetic and Nosé-Hoover algorithms on *each individual Cartesian degree of freedom* in a scheme known as *massive thermostatting* so that the equations of motion become

$$\begin{aligned}
\dot{\mathrm{x}} &= \mathrm{v}\,; \quad \dot{\mathrm{v}} = \frac{F}{m} - \lambda \mathrm{v} \\
\dot{\bar{\eta}} &= -\sum_{j=1}^{L}\left[\frac{Q\mathrm{v}_{\eta_{2,j}}\mathrm{v}_{\eta_{1,j}}^2}{k_BT} - \sum_{i=2}^{M}\mathrm{v}_{\eta_{i,j}}\right] \\
\dot{\mathrm{v}}_{\eta_{1,j}} &= -\mathrm{v}_{\eta_{1,j}}\mathrm{v}_{\eta_{2,j}} - \lambda\mathrm{v}_{\eta_{1,j}} \qquad j = 1, L \\
\dot{\mathrm{v}}_{\eta_{i,j}} &= \frac{G_{i,j}}{Q} - \mathrm{v}_{\eta_{i,j}}\mathrm{v}_{\eta_{i+1,j}} \qquad j = 1, L; \quad i = 2, M-1 \\
\dot{\mathrm{v}}_{\eta_{M,j}} &= \frac{G_{M,j}}{Q} \qquad j = 1, L
\end{aligned}$$

where

$$F = -\frac{dU}{d\mathrm{x}}\,; \quad G_{i,j} = Q\mathrm{v}_{\eta_{i-1,j}}^2 - k_BT. \quad (35)$$

and $Q = k_BT\tau^2$, k_B is Boltzmann's constant, T is the temperature and τ is the time scale associated with the bath. The Lagrange multiplier, λ, is selected so that the equations of motion satisfy the constraint,

$$2K(\mathrm{v},\mathrm{v}_\eta) = \left\{ m\mathrm{v}^2 + \left(\frac{L}{L+1}\right)\sum_{j=1}^{L} Q\mathrm{v}_{\eta_{1,j}}^2 \right\} = Lk_BT, \quad (36)$$

which yields

$$\lambda = \frac{\mathrm{v}F - \left(\frac{L}{L+1}\right)\sum_{j=1}^{L} Q\mathrm{v}_{\eta_{1,j}}^2\mathrm{v}_{\eta_{1,j}}}{2K(\mathrm{v},\mathrm{v}_\eta)}\,. \quad (37)$$

Thus, the maximum kinetic energy that can be sustained by any one mode is Lk_BT which eliminate resonant artifacts in numerical solvers[38].

The equations of motion, Eqs. 35, can be integrated fairly straightforwardly using operator splitting techniques. Briefly, the equations of motion and the dynamics can be written in the Liouville operator formalism

$$iL = \dot{\mathrm{x}}\frac{\partial}{\partial x} + \dot{\mathrm{v}}\frac{\partial}{\partial \mathrm{v}} + \sum_{ij} \dot{\mathrm{v}}_{\eta_{ij}} \frac{\partial}{\partial \mathrm{v}_{\eta_{ij}}} \tag{38}$$

$$\Gamma(t) = \exp(iLt)\Gamma(0) = \prod_{k=1}^{P} \exp(iL\Delta t)\Gamma(0)$$

with $\Delta t = t/P$ defining a single time step of evolution, $\Gamma(\Delta t) = \exp(iL\Delta t)\Gamma(0)$. Next, the Liouville operator is decomposed

$$iL = iL_{\mathrm{x}} + \sum_{p=1}^{N_d} iL_{\mathrm{v},p} + iL_{NHC} \tag{39}$$

$$iL_{\mathrm{x}} = \mathrm{v}\frac{\partial}{\partial x}\ ; \quad iL_{\mathrm{v},p} = (F_p - \lambda_p \mathrm{v})\frac{\partial}{\partial \mathrm{v}} - \sum_j \lambda_p \mathrm{v}_{\eta_{1,j}} \frac{\partial}{\partial \mathrm{v}_{\eta_{1,j}}}$$

$$iL_{NHC} = \sum_{i=2}^{M}\sum_{j=1}^{L} \frac{G_{i,j}}{Q}\frac{\partial}{\partial \mathrm{v}_{\eta_{i,j}}} - \sum_{i=1}^{M-1}\sum_{j=1}^{L} \mathrm{v}_{\eta_{i,j}}\mathrm{v}_{\eta_{i+1,j}}\frac{\partial}{\partial \mathrm{v}_{\eta_{i,j}}} - \sum_{j=1}^{L} \lambda_{NHC}\mathrm{v}_{\eta_{1,j}}\frac{\partial}{\partial \mathrm{v}_{\eta_{1,j}}}$$

and the force has been split into N_d parts, $\sum_{p=1}^{N_d} F_p = F$, whose strength is assumed to decrease with p. Using the decomposition and the multiple time step (MTS) parameters, $\delta t = \Delta t/N_{MTS}$, $N_{MTS} = \prod_{p=1}^{N_d} n_p$, $n_{N_d} = 1$, $w_p = \prod_{k=1}^{p-1} n_k$, $w_1 = 1$, an accurate approximation to the true evolution can be written

$$\Gamma(\Delta t) \approx \left\{ e^{i\tilde{L}_{N_d}^{(t)}\delta t} \ldots \left\{ e^{i\tilde{L}_2^{(t)}\delta t} \left[e^{i\tilde{L}_1)\delta t} \right]^{n_1-2} e^{i\tilde{L}_2\delta t} \right\}^{n_2-2} \ldots e^{i\tilde{L}_{N_d}\delta t} \right\} \Gamma(0) \tag{40}$$

where

$$e^{i\tilde{L}_k\delta t} = e^{iL_{NHC}\frac{\delta t}{2}} e^{iL_{\mathrm{v},1}\frac{\delta t}{2}} e^{iL_{\mathrm{x}}\delta t} e^{\sum_{p=1}^{k} iL_{\mathrm{v},p}w_p\frac{\delta t}{2}} e^{iL_{NHC}\frac{\delta t}{2}} \tag{41}$$

and $\exp(i\tilde{L}_k^{(t)}\delta t)$ is the transpose of $\exp(i\tilde{L}_k\delta t)$. In this way, the weaker forces which are assigned larger p, are evaluated/applied less frequently but with larger weight, w_p, which equalizes them to the strong forces (e.g. $w_{p+1}/w_p = n_p$ is the ratio of the strength of the p^{th} and the (p+1)$^{\mathrm{st}}$ force). The number of evaluations of the p^{th} force is N_{MTS}/w_p where, again, N_{MTS} is the total number of small steps in the multiple time step procedure. The error in the scheme is $\mathcal{O}(\Delta t^3)$ for one full step and $\mathcal{O}(t\Delta t^2)$ for the trajectory. Analytical solutions for each of the factorized parts of the operators, $\exp(i\tilde{L}_k)$, can be obtained easily given iL_{NHC} is further decomposed following Ref. 35. The decomposed multipliers, $\{\lambda_p, \lambda_{NHC}\}$ are chosen to enforce Eq. 36 and appropriately sum to Eq. 37. Judicious choices of the force decomposition into strong intramolecular vibrations and weak short range forces and weaker long-range forces are discussed in detail elsewhere[39]. The multiple time step approach improves efficiency because for chemical systems the strongest

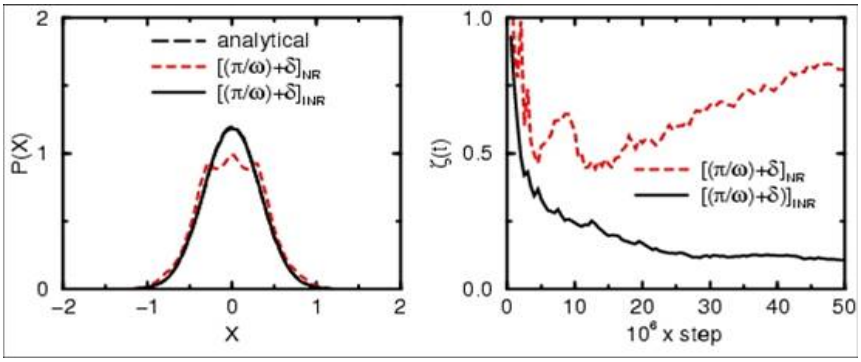


Figure 7. (a) Distribution function of the quartic oscillator $(9/2)x^2 + 0.025x^4$ for standard RESPA (NR) and the INR methods using $\Delta t = \pi/\omega$, which is the resonant time step, and $\delta t = \pi/(100\omega)$. (b) The error in the distribution function as a function of time as measured by $\zeta(t) = (1/N)\sum_{i=1}^{N} |P(x_i;t) - P_{\rm exact}(x_i)|$, where N is the number of bins in the histogram, $P(x_i;t)$ is the distribution at time t, and $P_{\rm exact}$ is the analytical distribution.

forces are least computational intensive to calculate. In total, the method will be referred to as the multiple time step, isokinetic, Nosé-Hoover chain (MTS-ISO-NHC) technique.

In order to demonstrate the efficacy of the new multiple time step technique, MTS-ISO-NHC, four problems with very different separations of time scale were selected for study. The quartic oscillator was chosen to demonstrate unequivocally that most basic resonance phenomena has been eliminated, the Lennard Jones fluid to show that the method enhances the sampling of simple solvent modes when long-range forces are rate limiting, water with flexible bonds and bends to show that a complex fluid with long-range forces and very high frequency vibrations can be tackled and finally a protein in *vacuo* to demonstrate that the method can handle large molecules with separations of time scales.

In Fig. 7, the converge of the probability distribution function of the quartic oscillator as a function of time step is given under the new equations of motion and multiple time step integrator, MTS-ISO-NHC, in comparison to standard methods, NHC. While standard methods become unstable, MTS-ISO-NHC yields the correct answer with extremely long time steps compared to the period of oscillation.

Third, a very challenging problem with extremely large separations of time scales is flexible water at room temperature and pressure. Due to resonance, standard methodology could not be employed for time steps larger than 1fs. However, the MTS-ISO-NHC generates the correct radial distribution function using a 100fs time step (see Fig. 8) without the degradation of mass diffusion (see Fig. 9) that would be caused by the introduction of an overdamped bath. Here, bonds and bends are treated using a 0.5 fs time step, and short-range forces within 5 Å are treated using a 3 fs time step. Long-range forces are assigned a cutoff of 12 Å and also include reciprocal space sums in an Ewald summation.

Last, a protein (HIV-protease) is studied in *vacuo* in order to demonstrate that the technique can handle large molecules without masking inefficiencies by including solvent. Again, a 100fs time step was capable of providing excellent results for both short and

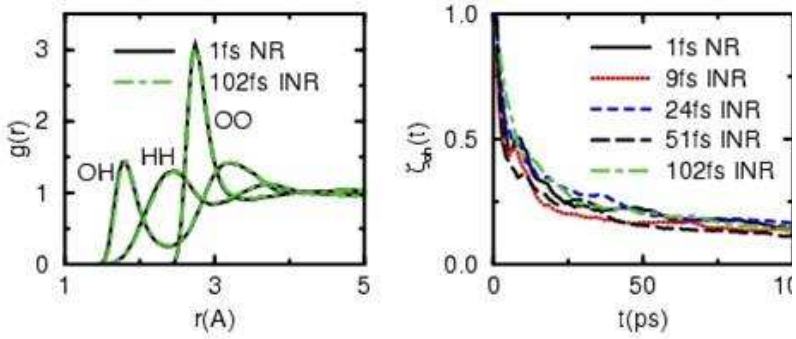


Figure 8. (a) The radial distribution functions of a flexible liquid water model (flex-TIP3P) using standard RESPA (NR) and the INR methods. (b) Error in the converged radial distribution functions. Time steps in the legend is the outer time step.

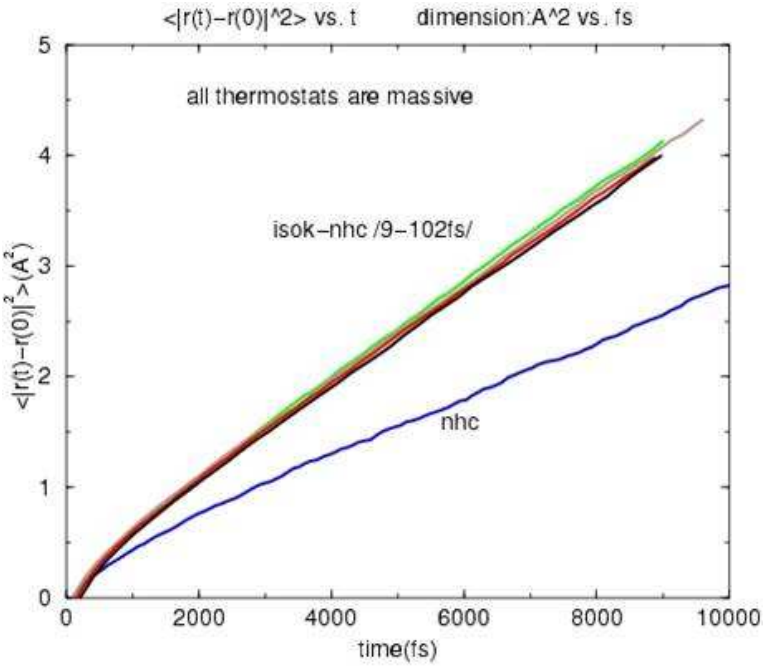


Figure 9. Diffusion constants of the flexible water model computed using the different methods. Although the methods are not designed to give the correct diffusion constant, the figure shows that diffusion is not arrested due to the use of INR.

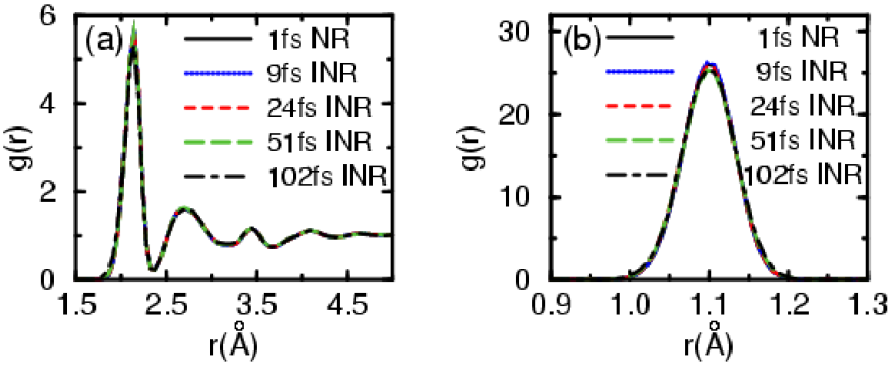


Figure 10. (a) The C-H radial distribution of the HIV protease in vacuo computed using standard RESPA (NR) and the INR methods. (b) The intramolecular part of the C-H distribution. Time steps in the legend are the outer time steps.

long-range distributions as shown in Fig. 10. Here, bonds and bends are treated using a 0.5 fs time step, and short-range forces within 5 Å are treated using a 3 fs time step.

Appendix

In this appendix, we prove that the AFED/d-AFED schemes generate the correct free energy surface. We consider The time evolution of the AFED system is generated by the Liouville operator. In order to keep the discussion general, we write this operator as

$$iL = \sum_{\alpha=1}^{3N} \left[\frac{p_\alpha}{m'_\alpha} \frac{\partial}{\partial q_\alpha} + F_\alpha(q) \frac{\partial}{\partial p_\alpha} \right] + iL_{\mathrm{therm},1}(T_q) + iL_{\mathrm{therm},2}(T), \tag{42}$$

where $F_\alpha(q) = -\partial \tilde{V}/\partial q_\alpha$ and $iL_{\mathrm{therm},1}(T_q)$ and $iL_{\mathrm{therm},2}(T)$ are the Liouville operators for the two thermostats. If x_t denotes the full phase space vector, including all variables related to the thermostats, then the time evolution of the system is formally given by

$$\mathrm{x}_t = e^{iLt}\mathrm{x}_0. \tag{43}$$

The key to analyzing this unusual dynamics is to factorize the propagator $\exp(iLt)$ in a way consistent with the adiabatic decoupling. To this end, we define the following combinations of terms in Eq. 42:

$$\begin{aligned} iL_{\mathrm{ref},1} &= \sum_{\alpha=1}^{n} \frac{p_\alpha}{m'_\alpha} \frac{\partial}{\partial q_\alpha} + iL_{\mathrm{therm},1}(T_q) \\ iL_{\mathrm{ref},2} &= \sum_{\alpha=n+1}^{3N} \frac{p_\alpha}{m'_\alpha} \frac{\partial}{\partial q_\alpha} + iL_{\mathrm{therm},2}(T) \\ iL_2 &= iL_{\mathrm{ref},2} + \sum_{\alpha=1}^{3N} F_\alpha(q) \frac{\partial}{\partial p_\alpha}. \end{aligned} \tag{44}$$

We next express the total Liouville operator as

$$iL = iL_{\mathrm{ref},1} + iL_2. \tag{45}$$

Let Δt be a time interval characteristic of the motion of the hot, heavy, and slow-moving reaction coordinates $q_1, ..., q_n$. Then, a Trotter decomposition of the propagator appropriate for the adiabatically decoupled motion is

$$e^{iL\Delta t} = e^{iL_2\Delta t/2} e^{iL_{\mathrm{ref},1}\Delta t} e^{iL_2\Delta t/2} + \mathcal{O}\left(\Delta t^3\right). \tag{46}$$

Note that the operator $\exp(iL_2\Delta t/2)$ has terms that vary on a time scale much faster than Δt and must be further decomposed. Using the ideas underlying multiple time-scale integration, we write this operator using the Trotter theorem as

$$\exp\left(iL_2\frac{\Delta t}{2}\right) = \lim_{M\to\infty}\left[\exp\left(\frac{\Delta t}{4M}\sum_{\alpha=1}^{3N} F_\alpha \frac{\partial}{\partial q_\alpha}\right)\right.$$
$$\left.\times \exp\left(iL_{\mathrm{ref},2}\frac{\Delta t}{2M}\right)\exp\left(\frac{\Delta t}{4M}\sum_{\alpha=1}^{3N} F_\alpha \frac{\partial}{\partial q_\alpha}\right)\right]^M. \tag{47}$$

It proves useful to decompose the phase space vector as $\mathrm{x} = (X, Y, P_X, P_Y, \Gamma_X, \Gamma_Y)$, where X denotes the full set of reaction coordinates, P_X, their momenta, Y, the remaining $3N - n$ coordinates, P_Y, their momenta, and Γ_X and Γ_Y, the thermostat variables associated with the temperatures T_q and T, respectively. Thus, when Eq. 47 is substituted into Eq. 46 and the resulting operator is taken to act on the initial phase space vector x_0, the result for heavy, slow reaction coordinates is

$$X_\alpha(\Delta t) = X_{\alpha,\mathrm{ref}}[X(0), \dot{X}(\Delta t/2), \Gamma_X(0); \Delta t]$$

$$\dot{X}_\alpha(\Delta t) = \dot{X}_{\alpha,\mathrm{ref}}[X(0), \dot{X}(\Delta t/2), \Gamma_X(0); \Delta t]$$
$$+\left(\frac{\Delta t}{2m'_\alpha}\right)\frac{2}{\Delta t}\int_{\Delta t/2}^{\Delta t} dt\; F_\alpha[X(\Delta t), Y_{\mathrm{adb}}(Y(\Delta t/2), \dot{Y}(\Delta t/2), \Gamma_Y(\Delta t/2), X(\Delta t); t)]$$

$$\dot{X}_\alpha(\Delta t/2) = \dot{X}_\alpha(0)$$
$$+\left(\frac{\Delta t}{2m'_\alpha}\right)\frac{2}{\Delta t}\int_{0}^{\Delta t/2} dt\; F_\alpha[X(0), Y_{\mathrm{adb}}(Y(0), \dot{Y}(0), \Gamma_Y(0), X(0); t)]$$

$$Y_\gamma(\Delta t/2) = Y_{\gamma,\mathrm{adb}}[Y(0), \dot{Y}(0), \Gamma_Y(0), X(0); \Delta t/2]$$
$$\dot{Y}_\gamma(\Delta t/2) = \dot{Y}_{\gamma,\mathrm{adb}}[Y(0), \dot{Y}(0), \Gamma_Y(0), X(0); \Delta t/2]$$
$$Y_\gamma(\Delta t) = Y_{\gamma,\mathrm{adb}}[Y(\Delta t/2), \dot{Y}(\Delta t/2), \Gamma_Y(\Delta t/2), X(\Delta t); \Delta t]$$
$$\dot{Y}_\gamma(\Delta t) = \dot{Y}_{\gamma,\mathrm{adb}}[Y(\Delta t/2), \dot{Y}(\Delta t/2), \Gamma_Y(\Delta t/2), X(\Delta t); \Delta t]. \tag{48}$$

In Eq. 48, $X_{\alpha,\mathrm{ref}}[X(0), \dot{X}(0), \Gamma_X(0); \Delta t]$ represents the evolution of X_α ($\alpha = 1, ..., n$) up to time Δt under the action of the reference-system operator $\exp(iL_{\mathrm{ref},1}\Delta t)$ starting from the initial conditions $X(0)$, $\dot{X}(0)$, $\Gamma_X(0)$, with an analogous meaning for $\dot{X}_\alpha[X(0), \dot{X}(0), \Gamma_X(0); \Delta t]$. $Y_{\gamma,\mathrm{adb}}[Y(0), \dot{Y}(0), \Gamma_Y(0), X(0); \Delta t/2]$ denotes the exact

evolution of Y_γ ($\gamma = 1, ..., 3N-n$) up to time $\Delta t/2$ under the first action of the operator $\exp(iL_2\Delta t/2)$ given in the form of Eq. 47 starting from initial conditions $Y(0)$, $\dot{Y}(0)$, $\Gamma_Y(0)$, $X(0)$ with an analogous meaning for $\dot{Y}_{\gamma,\mathrm{adb}}[Y(0), \dot{Y}(0), \Gamma_Y(0), X(0); \Delta t/2]$. The functions in the last two lines of Eq. 48 are similarly defined for the second action of $\exp(iL_2\Delta t/2)$. Although we do not have closed-form expressions for these functions in general, we do not need them for the present analysis. The important terms in Eq. 48 are the time integrals of the forces on the slow reaction coordinates. These time integrals result from the action of the operator $\exp(iL_2\Delta t/2)$ on the reaction coordinates which, for finite M, leads to a sum of force terms at different times $\Delta t/M$. This sum is in the form of a trapezoidal rule for a numerical integration in time. Thus, when the limit $M \to \infty$ is taken, these sums become continuous time integrals.

Physically, Eq. 48 tells us that the force driving the slow reaction coordinates is a time average over the motion of the $3N-n$ adiabatically decoupled fast variables. If the n masses assigned to the reaction coordinates are very large, the remaining variables will follow the slow reaction coordinates approximately instantaneously and sample large regions of their phase space at roughly fixed values of the reaction coordinates. In this limit, the time integrals in Eq. 48 can be replaced by configuration-space integrals, assuming that the motion of the fast variables is ergodic:

$$\begin{aligned}\frac{2}{\Delta t}\int_\tau^{\tau+\Delta t/2} dt\ F_\alpha[X, Y_{\mathrm{adb}}(Y(\tau), \dot{Y}(\tau), \Gamma_Y(\tau), X; t)] \\ = \frac{\int dY\ F_\alpha(X,Y) e^{-\beta\tilde{V}(X,Y)}}{\int dY\ e^{-\beta\tilde{V}(X,Y)}} \\ = \frac{\partial}{\partial q_\alpha}\frac{1}{\beta}\ln Z_Y(q_1, ..., q_n; \beta). \end{aligned} \tag{49}$$

Here

$$Z_Y(q_1, ..., q_n; \beta) = Z_Y(X; \beta) = \int dY\ e^{-\beta\tilde{V}(X,Y)} \tag{50}$$

is the configurational partition function at fixed values of the reaction coordinates $X = (q_1, ..., q_n)$. Eq. 49 defines an effective potential, the potential of mean force, on which the reaction coordinates move. Thus, we can define an effective Hamiltonian for the reaction coordinates as

$$H_{\mathrm{eff}}(X, P_X) = \sum_{\alpha=1}^{n}\frac{p_\alpha^2}{2m'_\alpha} - \frac{1}{\beta}\ln Z_Y(q_1, ..., q_n; \beta). \tag{51}$$

Since we assume the dynamics to be adiabatically decoupled, thermostats applied to this Hamiltonian yield the canonical distribution of $H_{\mathrm{eff}}(X, P_X)$ at temperature T_q:

$$\begin{aligned}P_{\mathrm{adb}}(X) = C_n\left[\int d^n p\ \exp\left\{-\beta_q\sum_{\alpha=1}^{n}\frac{p_\alpha^2}{2m'_\alpha}\right\}\right] \\ \times \exp\left\{-\beta_q\left(-\frac{1}{\beta}\ln Z_Y(q_1, ..., q_n)\right)\right\}. \end{aligned} \tag{52}$$

From Eq. 52, we see that

$$P_{\mathrm{adb}}(X) \propto [Z_Y(q_1, ..., q_n)]^{\beta_q/\beta}. \tag{53}$$

Since $Z_Y(q_1, ..., q_n)$ is the potential of mean force for the reaction coordinates, the free energy hypersurface $A(q_1, ..., q_n)$ is, by definition,

$$A(q_1, ..., q_n) = -\frac{1}{\beta} \ln Z_Y(q_1, ..., q_n), \tag{54}$$

but from Eq. 53, it follows that

$$A(q_1, ..., q_n) = -\frac{1}{\beta_q} \ln P_{\mathrm{adb}}(q_1, ..., q_n) + \mathrm{const}, \tag{55}$$

which is the true free energy profile. The constant in the second term comes from factors dropped in Eq. 52 and is irrelevant to the overall free energy hypersurface.

References

1. D. E. Shaw, M. M. Deneroff, R. O. Dror, J. S. Kuskin, R. H. Larson, J. K. Salmon, C. Young, B. Batson, K. J. Bowers, J. C. Chao, M. P. Eastwood, J. Gagliardo, J. p. Grossman, C. Richard Ho, D. J. Ieradi, I. Kolossvary, J. L. Klepeis, T. Layman, C. McLeavey, M. A. Moraes, R. Mueller, E. C. Priest, Y. Shan, J. Spengler, M. Theobald, B. Towles, and S. C. Wang, *Anton: A special-purpose machine for molecular dynamics simulation*, in: Proceedings of the 34th Annual International Symposium on Computer Architecture (ISCA 07), New York, NY, 2007.
2. D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. P. Eastwood, J. A. Bank, J. M. Jumper, J. K. Salmon, Y. Shan, and W. Wriggers, Science, **330**, 341, 2010.
3. J. L. Kleipeis, K. Lindorff-Larsen, R. O. Dror, and D. E. Shaw, Curr. Opin. Struct. Biol., **19**, 1, 2009.
4. R. W. Zwanzig, J. Chem. Phys., **22**, 1420, 1954.
5. G. M. Torrie and J. P. Valleau, Chem. Phys. Lett., **28**, 578, 1974.
6. G. M. Torrie and J. P. Valleau, J. Comput. Chem., **23**, 187, 1977.
7. J. G. Kirkwood, J. Chem. Phys., **3**, 300, 1935.
8. E. A. Carter, G. Ciccotti, J. T. Hynes, and R. Kapral, Chem. Phys. Lett., **156**, 472, 1989.
9. M. Sprik and G. Ciccotti, J. Chem. Phys., **109**, 7737, 1998.
10. L. Rosso and M. E. Tuckerman, Mol. Simul., **28**, 91, 2002.
11. L. Rosso, P. Minary, Z. Zhu, and M. E. Tuckerman, J. Chem. Phys., **116**, 4389, 2002.
12. L. Rosso, J. B. Abrams, and M. E. Tuckerman, J. Phys. Chem. B, **109**, 4162, 2005.
13. E. Darve and A. Pohorille, J. Chem. Phys., **115**, 9169, 2001.
14. E. Darve, D. Rodríguez-Gómez, and A. Pohorille, J. Chem. Phys., **128**, 144120, 2008.
15. A. Laio and M. Parrinello, Proc. Natl. Acad. Sci. U.S.A., **99**, 12562, 2002.
16. M. Bonomi and M. Parrinello, Phys. Rev. Lett., **104**, 190601, 2010.
17. G. A. Tribello, M. Ceriotti, and M. Parrinello, Proc. Natl. Acad. Sci. U.S.A., **107**, 17509, 2010.

18. H. Grubmüller, *Predicting slow structural transitions in macromolecular systems: Conformational flooding*, Phys. Rev. E, **52**, no. 3, 2893, 1995.
19. T. Huber, A. E. Torda, and W. F. van Gunsteren, *Local elevation: A method for improving the searching properties of molecular dynamics simulation*, J. Comput.-Aided Mol. Des., **8**, 695–708, 1994.
20. L. Maragliano and E. Vanden-Eijnden, Chem. Phys. Lett., **426**, 168, 2006.
21. J. B. Abrams and M. E. Tuckerman, J. Phys. Chem. B, **112**, 15742, 2008.
22. G. J. Martyna, M. L. Klein, and M. Tuckerman, *Nosé–Hoover chains: The canonical ensemble via continuous dynamics*, J. Chem. Phys., **97**, 2635, 1992.
23. Yi Liu and Mark E. Tuckerman, J. Chem. Phys., **112**, 1685, 2000.
24. J. Bernstein, *Polymorphism in Molecular Crystals*, Oxford University Press, Great Clarendon Street, Oxford OX2 6DP, 2002.
25. S. M. Woodley and Catlow R., *Crystal structure prediction from first principles*, NAT. MATER., **7**, 937, 2008.
26. P. Raiteri, R. Martoňák, and M. Parrinello, *Exploring Polymorphism: The Case of Benzene*, Angew. Chem. Int. Ed., **44**, 3769, 2005.
27. T. Q. Yu and M. E. Tuckerman, *Temperature accelerated approach for rapidly exploring crystalline polymorphism based on free energy*, Phys. Rev. Lett., (submitted).
28. T. Q. Yu and M. E. Tuckerman, *Constrained molecular dynamics in the isothermal-isobaric ensemble and its adaptation for adiabatic free energy dynamics*, Euro. Phys. J., **200**, 183, 2011.
29. J. D. Dunitz and H. A. Scheraga, *Exercises in prognostication: Crystal structures and protein folding*, Proc. Natl. Acad. Sci., **101**, 14309, 2004.
30. L. Rosso, P. Minary, Z. Zhu, and M. E. Tuckerman, *On the use of the adiabatic molecular dynamics technique in the calculation of free energy profiles*, J. Chem. Phys., **116**, 4389, 2002.
31. G. J. Martyna, D. J. Tobias, and M. L. Klein, *Constant-pressure molecular dynamics algorithms*, J. Chem. Phys., **101**, 4177, 1994.
32. D. van der Spoel, A. R. Buuren, D. P. Tieleman, and H. J. C. Berendsen, *Molecular dynamics simulations of peptides from BPTI: A closer look at amidearomatic interactions*, J. Biomol. NMR., **8**, 229, 1996.
33. Y. Yonetani and K. Yokoi, *Solid structures of benzene at high pressures: molecular dynamics study*, Mol. Phys., **99**, 1743, 2001.
34. M. E. Tuckerman, B. J. Berne, and G. J. Martyna, J. Chem. Phys., **97**, 1990, 1992.
35. G. J. Martyna, M. E. Tuckerman, D. J. Tobias, and M. L. Klein, Mol. Phys., **87**, 1117, 1996.
36. T. Schlick, M. Mandzuik, R. D. Skeel, and K. Srinivas, *Nonlinear resonance artifacts in molecular dynamics simulations*, J. Comput. Phys., **140**, 1, 1998.
37. P. Minary, G. J. Martyna, and M. E. Tuckerman, *Algorithms and novel applications based on the isokinetic ensemble. I. Biophysical and path integral molecular dynamics*, J. Chem. Phys., **118**, 2510, 2003.
38. P. Minary, G. J. Martyna, and M. E. Tuckerman, *Long time molecular dynamics for enhanced conformational sampling in biomolecular systems*, Phys. Rev. Lett., **93**, 150201, 2004.
39. J. A. Morrone, R. Zhu, and B. J. Berne, *Molecular Dynamics with Multiple Time Scales: How to Avoid Pitfalls*, J. Chem. Theor. Comput., **6**, 1798, 2010.

Methods on TDDFT-Based Nonadiabatic Dynamics with Applications

Ivano Tavernelli

Laboratory of Computational Chemistry and Biochemistry
Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland
E-mail: ivano.tavernelli@epfl.ch

1 Introduction

Ab initio molecular dynamics has led to a large number of theoretical predictions for molecules and solids. In the well-established mixed quantum-classical formulation, and thanks to highly efficient electronic structure methods like Kohn-Sham Density Functional Theory (DFT)[1], simulations on molecular systems with up to thousands of atoms are nowadays feasible. Initially restricted to a single adiabatic state (Born-Oppenheimer dynamics), molecular dynamics was recently extended to the nonadiabatic regime[2–4] becoming an important tool for the study of photophysical and photochemical processes.

Among the most commonly used nonadiabatic molecular dynamics schemes are Ehrenfest dynamics and Tully's trajectory surface hopping[5] (TSH). In the first case, the nuclear dynamics is replaced by a single point-like trajectory evolving in the mean-field potential derived from the time-evolution of the electronic wavefunction. Differently, in TSH the nuclear wavepacket is represented by a swarm of *independent* classical trajectories while the nonadiabatic couplings (NACs) induce hops between different electronic states that occur according to a stochastic algorithm. The classical approximation for the nuclei breaks down when interferences[6], wavepacket bifurcation[7], (de)coherence or tunneling effects occur during the dynamics. A trajectory-based solution of the quantum dynamics able to describe theses phenomena was introduced by Wyatt and co-workers. The so-called quantum trajectory method (QTM) describes the time evolution of the nuclear wavefunction by means of the quantum hydrodynamics (Bohmian) equations of motion[8]. Trajectory-based Bohmian dynamics differs from the classical TSH approach for the action of an additional potential, called the quantum potential, which is responsible for all quantum nuclear effects neglected in TSH.

As an alternative to trajectory-based approaches, quantum dynamics methods use an exact treatment of both electronic and nuclear wavefunctions (see for example Ref. 9). However, the applicability of these methods is hampered by their high computational costs, which limit the number of accessible nuclear degrees of freedom. This usually requires fitting of the relevant electronic potential energy surfaces (PESs) prior to propagation.

In the effort to extend the applicability of nonadiabatic MD to larger systems of physical, chemical and biological interest and relevance, we will also discuss the possibility of bridging different time and length scales, while keeping a valid description of the photoactive components. This is done through a hierarchical scheme, which combines the ultrafast dynamics of photoexcited electrons (purely QM tier) with the nuclear dynamics of the excited molecule (mixed-quantum classical tier) and with the reorganization of the environment (fully classical, MM, level).

After a short description of the most relevant approaches used to describe the nonadiabatic dynamics of molecular systems in the *unconstrained* phase space (Ehrenfest and TSH dynamics), I will discuss their implementation within the framework of DFT/TDDFT, which allow for an efficient *on-the-fly* calculation of all required electronic structure properties such as potential energy surfaces (PESs), nuclear forces, and nonadiabatic couplings. The coupling of the dynamics with the photochemical "inert" environment within the QM/MM scheme is also discussed together with a number of applications in different scientific domains.

For space reasons, in this lecture notes I will skip the derivation of the main mixed quantum-classical solutions. The interested reader can find more information in literature.

2 Mixed Quantum-Classical Nonadiabatic Molecular Dynamics: A TDDFT-Based Prospective

The starting point of our derivation is the time-dependent Schrödinger equation

$$\hat{H}_{mol}\Psi(\boldsymbol{r},\boldsymbol{R},t) = i\hbar\frac{\partial}{\partial t}\Psi(\boldsymbol{r},\boldsymbol{R},t) \tag{1}$$

where $\boldsymbol{R} = (\boldsymbol{R}_1, \boldsymbol{R}_2, \ldots, \boldsymbol{R}_{N_n})$ is the collective vector of the nuclear positions in $\mathbb{R}^{3N_n}$ and $\boldsymbol{r} = (\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_{N_{el}})$ the one for the electrons. In Eq. 1, $\hat{H}_{mol}$ is the molecular Hamiltonian

$$\begin{aligned}\hat{H}_{mol}(\boldsymbol{r},\boldsymbol{R}) =& -\sum_{\gamma}\frac{\hbar^2}{2M_\gamma}\nabla_\gamma^2 - \sum_i\frac{\hbar^2}{2m_e}\nabla_i^2 + \sum_{i<j}\frac{e^2}{|\boldsymbol{r}_i-\boldsymbol{r}_j|} - \sum_{\gamma,i}\frac{e^2Z_\gamma}{|\boldsymbol{R}_\gamma-\boldsymbol{r}_i|} \\ &+ \sum_{\gamma<\zeta}\frac{e^2Z_\gamma Z_\zeta}{|\boldsymbol{R}_\gamma-\boldsymbol{R}_\zeta|} \\ =& -\sum_{\gamma}\frac{\hbar^2}{2M_\gamma}\nabla_\gamma^2 + \sum_{\gamma<\zeta}\frac{e^2Z_\gamma Z_\zeta}{|\boldsymbol{R}_\gamma-\boldsymbol{R}_\zeta|} + \hat{\mathcal{H}}_{el}(\boldsymbol{r},\boldsymbol{R})\end{aligned} \tag{2}$$

and $\Psi(\boldsymbol{r},\boldsymbol{R},t)$ the total wavefunction for electrons and nuclei.

In this lecture, I will derive the equation of motion for the nuclear and electronic degrees of freedom using a trajectory-based approach. In this framework, the electrons are described at a quantum mechanical level, while the nuclear wavepacket is discretized into an ensemble of points in the phase space and then propagated along classical (or quantum Bohmian) trajectories that, as we will see, will keep some flavor of the underlying quantum dynamics (in particular nonadiabatic effects).

The first step in the derivation of the equation of motions for the combined electron-nuclear dynamics is the definition of a suited representation of the total system wavefunction. Depending on the particular choice of this *Ansatz* we can obtain different (approximated) solutions of the initial molecular Schrödinger equation (Eq. 1). In the following we will restrict to two main representations of the total molecular wavefunction that will give rise to two main trajectory-based nonadiabatic molecular dynamics solutions: mean field Ehrenfest dynamics and surface hopping dynamics.

$$\Phi(\boldsymbol{r},t)\Omega(\boldsymbol{R},t)e^{\left[\frac{i}{\hbar}\int_{t_0}^{t}E_{el}(t')dt'\right]} \xleftarrow{\text{Ehrenfest}} \Psi(\boldsymbol{r},\boldsymbol{R},t) \xrightarrow{\text{Born-Huang}} \sum_j^{\infty}\Phi_j(\boldsymbol{r};\boldsymbol{R})\Omega_j(\boldsymbol{R},t)$$

In both approaches the nuclei are described as classical trajectories (a single one in the mean field, Ehrenfest solution) and therefore these methods belong to the class of the **mixed quantum-classical** solutions of Eq. 1.

2.1 Ehrenfest Dynamics

In Ehrenfest dynamics we make use of a single-configuration *Ansatz* for the total wavefunction

$$\Psi(\boldsymbol{r},\boldsymbol{R},t) = \Phi(\boldsymbol{r},t)\Omega(\boldsymbol{R},t)\exp\left[\frac{i}{\hbar}\int_{t_0}^{t} E_{el}(t')dt'\right] \tag{3}$$

where $\Phi(\boldsymbol{r},t)$ describes the (time-dependent) electronic wavefunction and $\Omega(\boldsymbol{R},t)$ the (time-dependent) nuclear wavefunction.

The exponential part of Eq. 3 is called the "phase term"

$$E_{el}(t) = \int\!\!\int d\boldsymbol{r}\, d\boldsymbol{R}\, \Phi^*(\boldsymbol{r},t)\Omega^*(\boldsymbol{R},t)\hat{\mathcal{H}}_{el}(\boldsymbol{r},\boldsymbol{R})\Phi(\boldsymbol{r},t)\Omega(\boldsymbol{R},t) \tag{4}$$

and represents the average value of the electronic Hamiltonian, $\langle\hat{\mathcal{H}}_{el}(\boldsymbol{r},\boldsymbol{R})\rangle$, at time t.

The TDDFT-based Ehrenfest MD scheme. As mention in the introduction, I will not go through the derivation of the equation of motions but I will limit myself to the presentation of the working equations and the discussion of their meaning, as well as their advantages, and disadvantages. For a derivation see for instance[10,11].

The Ehrenfest MD scheme requires the simultaneous solution of the coupled differential equations for the electronic and nuclear dynamics

$$i\hbar\frac{\partial\Phi(\boldsymbol{r};\boldsymbol{R},t)}{\partial t} = \hat{\mathcal{H}}_{el}(\boldsymbol{r};\boldsymbol{R})\Phi(\boldsymbol{r};\boldsymbol{R},t) \tag{5}$$

$$M_\gamma\ddot{\boldsymbol{R}}_\gamma = -\nabla_\gamma\langle\hat{\mathcal{H}}_{el}(\boldsymbol{r};\boldsymbol{R})\rangle\,, \tag{6}$$

where γ is a label for the nuclei and '$;\boldsymbol{R}$' denotes the parametric dependence of the electronic Schrödinger equation from the atomic positions. This is accomplished by means of a two-step Runge-Kutta algorithm as described in Refs. 12, 13. All quantities required in Eqs. 5 and 6 are therefore computed *on-the-fly* at the instantaneous potential generated by the nuclear positions ($\boldsymbol{R}_\gamma(t)$) and electronic wavefunctions $\Phi(\boldsymbol{r};\boldsymbol{R},t)$.

Until this point, the formulation was kept as general as possible and did not depend from a particular choice of the electronic structural method used to solve the time-dependent Schrödinger equation for the electrons (Eq. 5). In many cases, it is however convenient to *map* the set of Eqs. 5 and 6 into a DFT/TDDFT-based formulation, which allows for an efficient calculation of the molecular electronic structure and properties for systems composed by several hundreds of atoms at a moderate computational cost.

Using the Runge-Gross theorem[14], we can derive the equation of motion for the electron density from the variational principle applied to the action functional[a]

[a]This form of the action functional violates causality because a potential of the form

$$v_{xc}(\boldsymbol{r},t) = \frac{\delta\mathcal{A}_{xc}[\rho]}{\delta\rho(\boldsymbol{r},t)} \tag{7}$$

has a functional derivative $\frac{\delta v_{xc}(\boldsymbol{r},t)}{\delta\rho(\boldsymbol{r}',t')}$ which is symmetric in time, and therefore does not guarantee that information only travels forward in time. Different solutions to the causality problem exists[15,16], however they do not affect our derivations.

$$\mathcal{A}[\rho] = \langle \Psi[\rho] | i\hbar \frac{\partial}{\partial t} - \hat{T} - \hat{H}_{ee} | \Psi[\rho] \rangle \,, \tag{8}$$

where $\hat{H}_{ee}$ is the exact electron-electron interaction and $\Psi[\rho](t)$ is the time-dependent wavefunction associated to the time-dependent density $\rho(\boldsymbol{r}, t)$. In the Kohn-Sham formulation of DFT, the action becomes a functional of the one-electron KS orbitals,

$$\begin{aligned} \mathcal{A}[\rho] = & \sum_i \int_{t_0}^{t_1} dt \langle \phi_i(t) | i\hbar \frac{\partial}{\partial t} + \frac{1}{2}\nabla^2 | \phi_i(t) \rangle - H[\rho(\boldsymbol{r}, t)] - \mathcal{A}_{xc}[\rho(\boldsymbol{r}, t)] \\ & - \int d\boldsymbol{r} \int_{t_0}^{t_1} dt \, v_{ext}(\boldsymbol{r}, t) \rho(\boldsymbol{r}, t) \end{aligned} \tag{9}$$

where $H[\rho(\boldsymbol{r}, t)]$ is the Hartree energy functional. In the DFT and TDDFT context, the variable $\boldsymbol{r}$ refers to a position in the Euclidean space ($\boldsymbol{r} \in \mathbb{R}^3$) and should not be confused with the collective electron coordinate used for the many-electron wavefunctions.

Applying the variational principle to Eq. 9 subject to the constraint $\rho(\boldsymbol{r}, t) = \sum_k |\phi_k(\boldsymbol{r}, t)|^2$ results in the time-dependent Kohn-Sham equations (TDKS)

$$i\hbar \frac{\partial}{\partial t} \phi_k(\boldsymbol{r}, t) = -\frac{1}{2m_e} \nabla^2 \phi_k(\boldsymbol{r}, t) + v_{\text{eff}}[\rho, \Phi_0](\boldsymbol{r}, t) \, \phi_k(\boldsymbol{r}, t) \,, \quad k = 1, \ldots, N_{el} \,. \tag{10}$$

where

$$v_{\text{eff}}[\rho, \Phi_0](\boldsymbol{r}, t) = v_{ext}(\boldsymbol{R}, t) + v_H(\boldsymbol{r}, t) + \frac{\delta \mathcal{A}_{xc}[\rho, \Phi_0](\boldsymbol{r}, t)}{\delta \rho(\boldsymbol{r}, t)} \,, \tag{11}$$

$v_{ext}(\boldsymbol{R}, t)$ is the external potential, and $v_H(\boldsymbol{r}, t)$ is the Hartree potential. The effective potential $v_{\text{eff}}[\rho, \Phi_0]$ also depends on the initial wavefunction at time t_0 ($\Phi_0(\boldsymbol{r}, t)$) or, equivalently, the corresponding density $\rho(\boldsymbol{r}, t_0)$. However, to simplify the notation, in the following I will remove the dependence from the the initial value conditions.

The simplest approximation to the time-dependent exchange-correlation action functional $\mathcal{A}_{xc}[\rho(\boldsymbol{r}, t)]$ is the so-called *adiabatic approximation*, AA,

$$\mathcal{A}_{xc}[\rho] = \int d\boldsymbol{r} \int_{t_0}^{t_1} dt \, \rho(\boldsymbol{r}, t) \epsilon_{xc}[\rho(\boldsymbol{r})]|_{\rho(\boldsymbol{r}) \leftarrow \rho(\boldsymbol{r}, t)} \tag{12}$$

where ϵ_{xc} is the DFT (ground state) exchange and correlation energy density functional. This approximation is sufficient for most of the applications. However, examples of Ehrenfest-type MD simulations beyond the AA are already present in the literature[17,18].

An alternative solution to the propagation of the electronic wavefunction within DFT/TDDFT is to introduce the representation of the time-dependent KS orbitals in a linear combination of *static* KS orbitals, $\{\phi_k^{opt}(\boldsymbol{r})\}$, obtained from the diagonalization of the KS Hamiltonian at time t with effective potential $v_{\text{eff}}[\rho]|_{\rho(\boldsymbol{r}) \leftarrow \rho(\boldsymbol{r}, t)}$

$$\phi_k(\boldsymbol{r}, t) = \sum_k^{\infty} c_k(t) \, \phi_k^{opt}(\boldsymbol{r}) \,. \tag{13}$$

Inserting this expansion into Eq. 10 produces a set of differential equations for the coefficients $c_k(t)$ that, when squared, can be interpreted as KS orbital occupations. While this

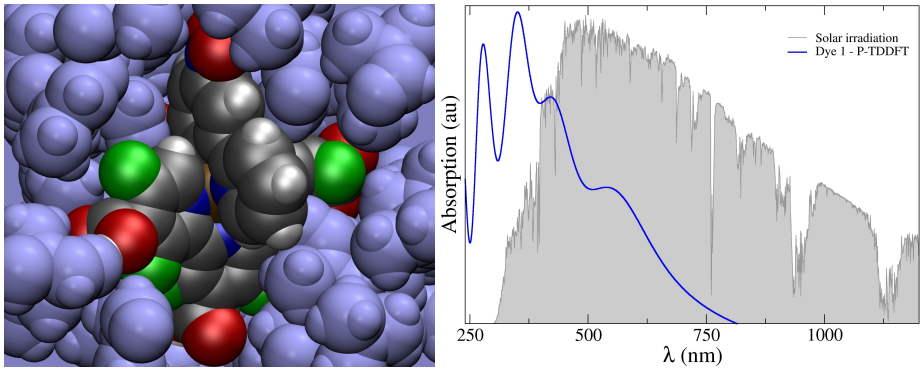


Figure 1. TDDFT spectra of a Ruthenium-based dye in Dimethylformamide (DMF) solution computed using the propagation of the time-dependent KS orbitals according to Eq. 10. The solvent is treated at classical level within a QM/MM setup. After an initial equilibration at 300K, a perturbation is applied to the KS orbitals and the time-evolution of the dipole moment is recorded. The Fourier transform of this signal provides the full absorption spectra, whose energy resolution depends on the total length of the propagation. Left panel: QM/MM setup: atoms in the QM part are represented with colored vdW spheres (gray: carbon, white: hydrogen, blue: nitrogen, red: oxygen, green: fluorine), whereas the MM atoms are shown in light blue color. Right panel: Computed TDDFT/MM spectra. The solar irradiation spectra at the Earth surface (AM 1.5) is shown in gray color.

offers some additional information about the nature of the propagated state, the diagonalization of the KS Hamiltonian requires additional computational costs that can be avoided using the straightforward propagation of the KS orbitals in Eq. 10.

The mapping of the nuclear dynamics (Eq. 6) into the DFT formalism is more straightforward and only requires the calculation of the forces $-\nabla_\gamma \langle \hat{\mathcal{H}}_{el}(\boldsymbol{r};\boldsymbol{R})\rangle$ as a functional of the time-dependent density $\rho(\boldsymbol{r},t)$. Replacing the expectation value of the electronic Hamiltonian with the DFT energy evaluated with the exchange-correlation potential $v_{xc}[\rho]$ the gradient with respect to the nuclear coordinates can be performed analytically as in the case of the Born-Oppenheimer[10] and Car-Parrinello[10,19] MD schemes.

Limitations and use of Ehrenfest MD. In Ehrenfest MD the nuclei evolve in time according to forces computed as the gradient of average energy, $\langle \hat{\mathcal{H}}_{el}(\boldsymbol{r},\boldsymbol{R})\rangle$, using the instantaneous many-electron wavefunction $\Phi(\boldsymbol{r};\boldsymbol{R},t)$. Both quantum Hamiltonian and electronic wavefunction depend parametrically on $\boldsymbol{R}$. The *mean-field* character of this dynamics is evident if we use the following expansion of the wavefunction $\Phi(\boldsymbol{r};\boldsymbol{R},t)$ in the adiabatic (static) base, $\{\phi_k(\boldsymbol{r};\boldsymbol{R})\}$, obtained from the solutions of the electronic time-independent Schrödinger equation

$$\Phi(\boldsymbol{r};\boldsymbol{R},t) = \sum_{k=0}^{\infty} \tilde{c}_k(t)\phi_k(\boldsymbol{r};\boldsymbol{R}) \tag{14}$$

(do not confuse this equation with Eq. 13, which deals with one-electron KS orbitals). The square of the time-dependent coefficients $\tilde{c}_k(t)$ describes the "occupation probability" of the different states that contribute to the nuclear forces. The validity of the "mean-field" approximation is restricted to the case in which the classical trajectories corresponding to the different states do not differ too much[20]. In fact, after leaving a region of strong

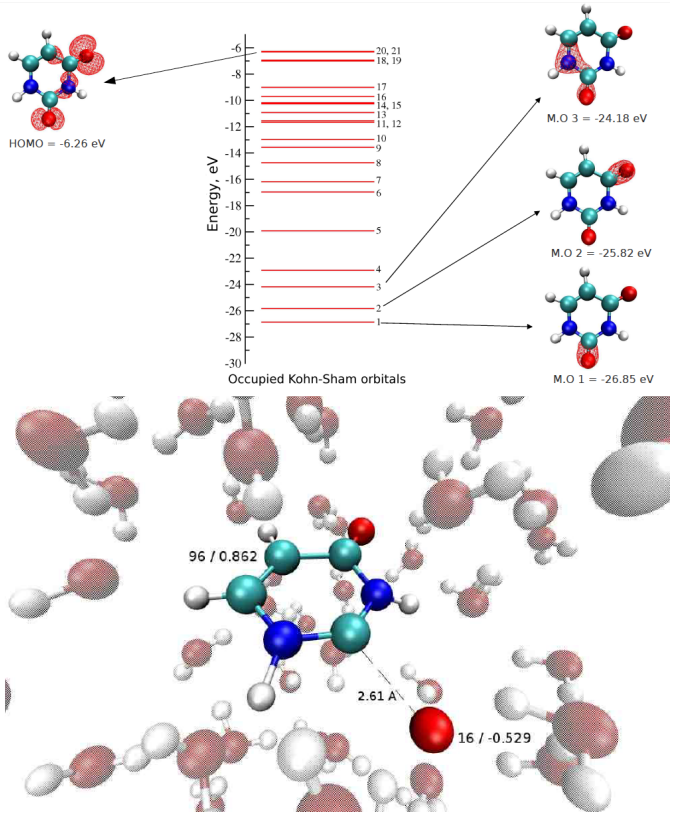


Figure 2. Double ionization dynamics of the uracyl molecule. After removal of two electrons from an occupied KS orbitals the excess positive charge of the system induced a so-called *Coulomb explosion* with the formation of different dissociation fragments depending from the nature of the ionized orbital (upper panel). This process is mimicking the effect of the collision of ultrafast and highly charged ions used in radiotherapy with biological molecules. The lower panel shows a frame obtained during the dynamics of Uracil^{2+} in water. All calculations are performed using the TDDFT-based Ehrenfest dynamics using a very small time step (less than 0.5 *as*) due to the ultrafast electron-nuclear dynamics.

nonadiabatic coupling (mixing), the system keeps evolving on an average potential without collapsing to one of the adiabatic states. In the case trajectories on different states will have different evolutions in the configuration space, the average Ehrenfest trajectory could lose its physical meaning (even though it could still provide acceptable expectation values for quantum observables).

For this reason, the use of Ehrenfest dynamics in molecular computational physics and chemistry should be restricted to the cases in which the atomic rearrangements along the different reaction channels associated to the electronic states involved in the mean-field dynamics are not too different. Examples are the calculation of absorption spectra and dielectric functions of systems in gas phase and solution[21,22], for which the electron dynamics of the perturbed electronic structure relaxes on a time scale that is much shorter than the one of the nuclei (in many cases one can use frozen atomic positions). Other ap-

plications include the investigation of ultrafast processes triggered by intense laser fields that produces ionized states followed by fast electronic rearrangement of the electronic structure like photoionization and Coulomb explosion[23–26] (when only dissociation channels are populated).

Another severe deficiency of Ehrenfest MD is the violation of the detailed balance (microscopic reversibility), which is an essential property of the kinetics of physical and chemical processes.

2.2 Born-Oppenheimer MD and its Nonadiabatic Extensions

The Born-Oppenheimer MD equations can be derived starting from the Born-Huang representation of the molecular wavefunction

$$\Psi(\boldsymbol{r},\boldsymbol{R},t)=\sum_{j=0}^{\infty}\Omega_j(\boldsymbol{R},t)\Phi_j(\boldsymbol{r};\boldsymbol{R})\,. \tag{15}$$

In this equation, $\{\Phi_j(\boldsymbol{r};\boldsymbol{R})\}$ describes a complete set of orthonormal electronic wavefunctions solution of the time-independent Schrödinger equation

$$\hat{\mathcal{H}}_{el}(\boldsymbol{r};\boldsymbol{R})\Phi_j(\boldsymbol{r};\boldsymbol{R})=E_j^{el}(\boldsymbol{R})\Phi_j(\boldsymbol{r};\boldsymbol{R}) \tag{16}$$

with $\langle\Phi_j|\Phi_i\rangle=\delta_{ij}$. Note that only the nuclear wavefunctions depend explicitly on time, while $\hat{\mathcal{H}}_{el}(\boldsymbol{r};\boldsymbol{R})$ and $\Phi_j(\boldsymbol{r};\boldsymbol{R})$ only depend on t through the implicit time-dependence of $\boldsymbol{R}(t)$.

Inserting Eq. 15 into the time-dependent Schrödinger equation (Eq. 1) we obtain (after multiplying by $\Phi_k^*(\boldsymbol{r};\boldsymbol{R})$ from the left-hand-side and integrating over $d\boldsymbol{r}$)

$$i\hbar\frac{\partial}{\partial t}\Omega_k(\boldsymbol{R},t)=\left[-\sum_\gamma\frac{\hbar^2}{2M_\gamma}\nabla_\gamma^2+E_{el,k}(\boldsymbol{R})\right]\Omega_k(\boldsymbol{R},t)+\sum_j\mathcal{F}_{kj}\Omega_j(\boldsymbol{R},t) \tag{17}$$

The quantities $\mathcal{F}_{kj}(\boldsymbol{R})$

$$\begin{aligned}\mathcal{F}_{kj}(\boldsymbol{R})=&\int d\boldsymbol{r}\;\Phi_k^*(\boldsymbol{r};\boldsymbol{R})\left[\sum_\gamma\frac{\hbar^2}{2M_\gamma}\nabla_\gamma^2\right]\Phi_j(\boldsymbol{r};\boldsymbol{R})\\&+\sum_\gamma\frac{1}{M_\gamma}\left\{\int d\boldsymbol{r}\;\Phi_k^*(\boldsymbol{r};\boldsymbol{R})\left[-i\hbar\nabla_\gamma\right]\Phi_j(\boldsymbol{r};\boldsymbol{R})\right\}\left[-i\hbar\nabla_\gamma\right]\end{aligned} \tag{18}$$

are the *nonadiabatic couplings*, where the first contribution originates from the nuclear kinetic operator and a second from the momentum operator. In the most general case, the non-diagonal elements $\mathcal{F}_{kj}(\boldsymbol{R})$ are non-zero and induce a coupling between different electronic states due to the motion of the nuclei. In fact, the last term in Eq. 17 brings amplitude ($\mathcal{F}_{kj}\Omega_j(\boldsymbol{R},t)$) from the "electronic state" with energy $E_j^{el}(\boldsymbol{R})$, to the actual state k, with energy $E_k^{el}(\boldsymbol{R})$. This interpretation of the nuclear wavefunction dynamics (Eq. 17) is at the basis of the *surface hopping* description of nonadiabatic dynamics.

The adiabatic solution. In the *adiabatic approximation* only the diagonal terms, $\mathcal{F}_{kk}$, are retained

$$\mathcal{F}_{kk} = \int \Phi_k^*(\boldsymbol{r};\boldsymbol{R}) \left[\sum_\gamma \frac{\hbar^2}{2M_\gamma}\nabla_\gamma^2\right] \Phi_k(\boldsymbol{r};\boldsymbol{R}) d\boldsymbol{r}\,, \tag{19}$$

which only induce a shift of the electronic potential energy surfaces $E_k^{el}(\boldsymbol{R})$ felt by the nuclear wavefunctions (the second term of Eq. 18 is zero for $k = j$, when $\Phi_.(\boldsymbol{r};\boldsymbol{R})$ are real).

In this approximation, the nuclei move in the potential of a single electronic state, the potential energy surface (PES) $E_k^{el}(\boldsymbol{R})$, and the electronic (Eq. 11) and nuclear (Eq. 17) Schrödinger equations become completely decoupled. The term $\mathcal{F}_{kk}$ is called Born-Oppenheimer diagonal correction and, depending on the nuclear mass, induces an isotope-dependence of the total energy, $E_k^{el} + \mathcal{F}_{kk}$. However, this term is usually small and is neglected in the so-called Born-Oppenheimer approximation.

At this point, I introduce the polar representation of the nuclear wavefunction $\Omega_k(\boldsymbol{R},t)$

$$\Omega_k(\boldsymbol{R},t) = A_k(\boldsymbol{R},t)\exp\left[\frac{i}{\hbar}S_k(\boldsymbol{R},t)\right] \tag{20}$$

with real amplitudes, $A_k(\boldsymbol{R},t)$, and phases, $S_k(\boldsymbol{R},t)/\hbar$. Inserting this equation into Eq. 17 (with $\mathcal{F}_{kj} = 0$) and separating the real and the imaginary parts, we obtain

$$\frac{\partial S_k}{\partial t} = \frac{\hbar^2}{2}\sum_\gamma M_\gamma^{-1}\frac{\nabla_\gamma^2 A_k}{A_k} - \frac{1}{2}\sum_\gamma M_\gamma^{-1}\big(\nabla_\gamma S_k\big)^2 - E_k^{el}(\boldsymbol{R}) \tag{21}$$

$$\frac{\partial A_k}{\partial t} = -\sum_\gamma M_\gamma^{-1}\nabla_\gamma A_k \nabla_\gamma S_k - \frac{1}{2}\sum_\gamma M_\gamma^{-1} A_k \nabla_\gamma^2 S_k \tag{22}$$

where the dependences of the fields S and A are omitted for clarity.

Taking the classical limit $\hbar \to 0$ [b] in both Eqs. 21 and 22, we obtain a Hamilton-Jacobi equation for the action function $S(\boldsymbol{R},t)$

$$\frac{\partial S_k}{\partial t} = -\frac{1}{2}\sum_\gamma M_\gamma^{-1}\big(\nabla_\gamma S_k\big)^2 - E_k(\boldsymbol{R})\,, \tag{23}$$

which correspond to a classical point-particle time evolution of the nuclei, and a continuity equation for the propagation of the amplitude on the adiabatic state of interest, $d/dt(\int d\boldsymbol{R}|\Omega_k(\boldsymbol{R},t)|^2) = 0$. Comparing this result with the one obtained for the Ehrenfest dynamics, we observe that in this case the potential acting on the nuclei is derived from a static expectation value of the electronic Hamiltonian computed for the time-independent state $\Phi_k(\boldsymbol{r};\boldsymbol{R})$, solution of the Schrödinger equation (Eq. 11).

Using the relation, $\nabla_\gamma S_k = \boldsymbol{P}_{k,\gamma}$, we obtain a Newton-like equation of motion for the 'classical' nuclei

$$M_\gamma \ddot{\boldsymbol{R}}_\gamma = -\nabla_\gamma E_k^{el}(\boldsymbol{R})\,. \tag{24}$$

[b] The classical limit proposed in this derivation is sometimes called the "canonical condition" for enforcing classical behavior. It is mainly a *mathematical* procedure with limited physical content. Alternative formulations with their physical implications can be found in different Refs. 27, 28.

In summary, the BO MD can be described by the following system of coupled equations

$$\hat{\mathcal{H}}_{el}(\boldsymbol{r};\boldsymbol{R})\Phi_k(\boldsymbol{r};\boldsymbol{R}) = E_k^{el}(\boldsymbol{R})\Phi_k(\boldsymbol{r};\boldsymbol{R}) \tag{25}$$

$$M_\gamma \ddot{\boldsymbol{R}}_\gamma = -\nabla_\gamma E_k^{el}(\boldsymbol{R}) = -\underset{min\Phi_k}{\nabla_\gamma} \langle\Phi_k|\hat{\mathcal{H}}_{el}|\Phi_k\rangle \tag{26}$$

where only the second one describes an explicit time evolution. The electronic energies and the forces acting on the nuclei are computed *statically* solving Eq. 25 *on-the-fly* at each new set of nuclear positions sampled along the trajectory $\boldsymbol{R}(t)$. Contrary to what obtained in Ehrenfest dynamics, in BO MD there is no explicit time-dependence of the electronic degrees of freedom. It is important to further stress that, due to the assumption that $\mathcal{F}_{kj} = 0$, the BO MD always evolves on a single electronic PES, even in the case where the system approaches regions of strong coupling between electronic and nuclear degrees of freedoms. In practice, the state of interest is mostly the ground state for which the adiabatic separation from all other states (excited states) holds in most (nonmetallic) cases.

The combination of BO MD with DFT for the on-the-fly calculation of the electronic structure properties (energies and forces) at each MD step is straightforward and can be found in many textbooks (see for instance[10]). Using the Hohenberg-Kohn theorem one first maps the electronic structure problem from the wavefunction space into the density space and then, within the Kohn-Sham formulation of DFT, the electronic ground state energy functional, $E_0[\rho(\boldsymbol{r};\boldsymbol{R})]$ and its gradients are computed.

2.3 Trajectory-Based Nonadiabatic Dynamics

The adiabatic approximation breaks down when electronic states approach in energy, which especially occurs when the dynamics is initiated in one of the excited states of the system. This is the usual situation encountered in *pump-probe* experiments, where an initial pulse is exciting the system while a second one its monitoring is time-dependent relaxation towards the ground state (or a stable excited state).

The starting point is the time-dependent Schrödinger equation for the molecular system (Eq. 17) that we rewrite as

$$\begin{aligned} i\hbar\frac{\partial\Omega_j(\boldsymbol{R},t)}{\partial t} = &-\sum_\gamma \frac{\hbar^2}{2M_\gamma}\nabla_\gamma^2\Omega_j(\boldsymbol{R},t) + E_j^{el}(\boldsymbol{R})\Omega_j(\boldsymbol{R},t) \\ &+\sum_{\gamma i}\frac{\hbar^2}{2M_\gamma}D_{ji}^\gamma(\boldsymbol{R})\Omega_i(\boldsymbol{R},t) - \sum_{\gamma,i\neq j}\frac{\hbar^2}{M_\gamma}\boldsymbol{d}_{ji}^\gamma(\boldsymbol{R})\nabla_\gamma\Omega_i(\boldsymbol{R},t) \end{aligned} \tag{27}$$

where

$$\boldsymbol{d}_{ji}^\gamma(\boldsymbol{R}) = \int \left\{\Phi_j^*(\boldsymbol{r};\boldsymbol{R})\left[\nabla_\gamma\Phi_i(\boldsymbol{r};\boldsymbol{R})\right]\right\} d\boldsymbol{r} \tag{28}$$

are the first order coupling elements, and

$$D_{ji}^\gamma(\boldsymbol{R}) = \int \left\{\Phi_j^*(\boldsymbol{r};\boldsymbol{R})\left[\nabla_\gamma^2\Phi_i(\boldsymbol{r};\boldsymbol{R})\right]\right\} d\boldsymbol{r} \tag{29}$$

are the second order coupling elements.

J. C. Tully proposed an approximate solution of the coupled electron-nuclear Schrödinger equation, which is known as Trajectory Surface Hopping (TSH) dynamics[29]. Within this approach, the nuclear wavepacket propagation is described by the time evolution of an ensemble of classical trajectories evolving *independently* on adiabatic potential energy surfaces, E_k^{el}. This *independent trajectory approximation* (ITA) implies that all nuclear quantum correlation effects are neglected. The transfer of amplitude (or better, trajectories) between different PESs is taken in charge by a stochastic surface hopping procedure, which requires the evaluation of the first order coupling elements, $\boldsymbol{d}_{ji}^{\gamma}(\boldsymbol{R})$ in Eq. 28. In practice, the nuclear wavepacket $\Omega_j(\boldsymbol{R},t)$ in the expansion in Eq. 15 is replaced by the complex-valued time-dependent amplitude $C_j^{\alpha}(t)$, which apportions trajectories (labelled by α) among electronic states according to the correct quantum probability, so that

$$|\Omega_j(\boldsymbol{R},t)|^2 \sim \frac{1}{M}\sum_{\{\alpha\}}\int_{t=0}^{\infty} dt' \, |C_j^{\alpha}(t')|^2 \delta(\boldsymbol{R}-\boldsymbol{R}^{\alpha}(t'))\delta(t-t') \,, \tag{30}$$

once a sufficient number of trajectories has been sampled. This relation holds due to the ITA assumption, while the $\boldsymbol{R}$ dependence of the $C_j^{\alpha}(t)$ coefficients is determined by the initial conditions, $\boldsymbol{R}(t=0)$, and Tully's equations of motion for the nuclei. The time-dependent differential equation for the amplitudes $C_j^{\alpha}(t)$ is obtained by replacing

$$\Psi^{\alpha}(\boldsymbol{r},\boldsymbol{R},t) = \sum_{j}^{\infty} C_j^{\alpha}(t)\Phi_j(\boldsymbol{r};\boldsymbol{R}) \tag{31}$$

in the time-dependent Schrödinger equation and reads (in the Schrödinger representation)

$$i\hbar\dot{C}_j^{\alpha}(t) = \sum_i C_i^{\alpha}(t)(H_{ji} - i\hbar\dot{\boldsymbol{R}}^{\alpha}\cdot\boldsymbol{d}_{ji}^{\alpha}) \tag{32}$$

where the label α indicates that the corresponding quantities are evaluated for a specific trajectory that contributes to the final ensemble. Because of the adiabatic representation of the electronic wavefunctions, the matrix elements H_{ji} are diagonal $H_{ji} = \delta_{ji}E_j^{el}(\boldsymbol{R})$, where $E_j^{el}(\boldsymbol{R})$ are the eigenvalues of Eq. 11. All matrix elements in Eq. 32 are computed using an *ab initio* electronic structure calculation or, as in the present case, DFT for the ground state and TDDFT for the excited states.

In Tully's dynamics, the classical trajectories evolve adiabatically according to Born-Oppenheimer dynamics until a *hop* between two potential energy surfaces (H_{ii} and H_{jj}) occurs with a probability given by a Monte Carlo-type procedure. In the "fewest switches" algorithm, the transition probability from state i to state j in the time interval $[t, t+dt]$ is

$$g_{ij}^{\alpha}(t,t+dt) \approx 2\int_t^{t+dt} d\tau \frac{\mathit{Im}[C_j^{\alpha}(\tau)C_i^{\alpha *}H_{ji}(\tau)] - \mathit{Re}[C_j^{\alpha}(\tau)C_i^{\alpha *}(\tau)\Xi_{ji}^{\alpha}(\tau)]}{C_i^{\alpha}(\tau)C_i^{\alpha *}(\tau)}, \tag{33}$$

where $\Xi_{ji}^{\alpha}(\tau) = \dot{\boldsymbol{R}}^{\alpha}\cdot\boldsymbol{d}_{ji}^{\alpha}(\tau)$, and a hop occurs if and only if

$$\sum_{k\le j-1} g_{ik}^{\alpha} < \zeta < \sum_{k\le j} g_{ik}^{\alpha} \,, \tag{34}$$

where ζ is generated randomly in the interval $[0,1]$. In practice, a swarm of trajectories is propagated independently starting from different initial conditions, and the final statistical distribution of all these trajectories is assumed to reproduce the correct time evolution of

the nuclear wavepacket. It is important to stress that, at present, no formal justification of Tully's algorithm has been formulated.

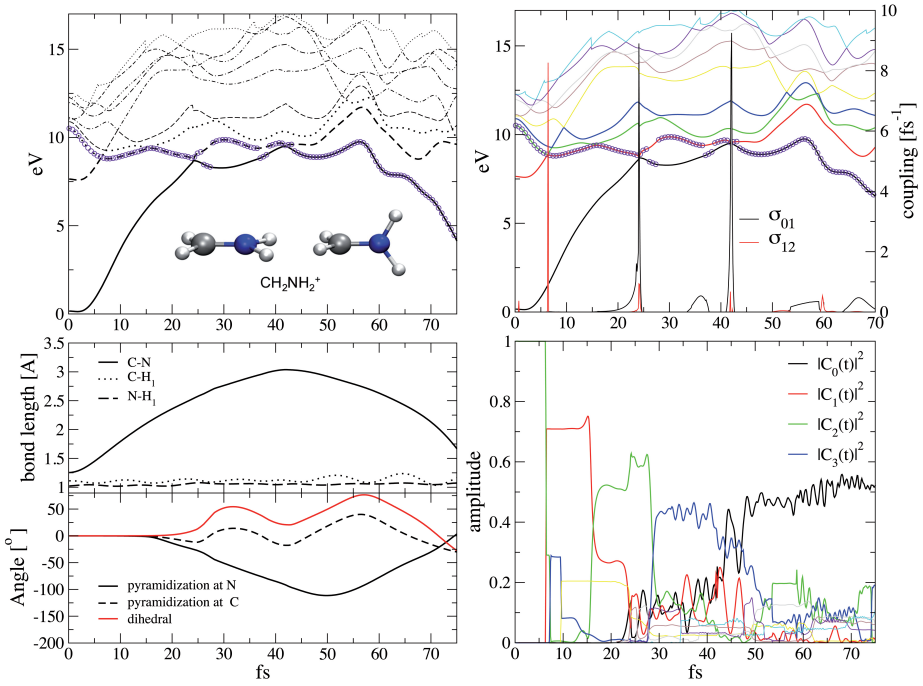


Figure 3. TDDFT-based TSH dynamics of photoexcited protonated formaldimine ($CH_2NH_2^+$), a model for the isomerization of the visual pigment retinal[30]. Left upper panel: Time series for the first eight excited state energies computed for a single trajectory initiated on S_2. The energy profiles are labeled (with different line styles) according to their increasing energy values and therefore they do not necessary follow the electronic character of the states. The state that drives the dynamics is marked with circles. Left lower panel: Time series for the different bond lengths, pyramidization angles and dihedral angle computed along the same trajectory. Right upper panel: Time series of the potential energy corresponding to the first 8 singlet states plotted together with the coupling strengths σ_{01} and σ_{12}. Lower panel: corresponding state populations, $|C_i|^2$, computed using the amplitudes defined in Eq. 31. The color code refers to the different energy curves in the upper panel.

The TSH algorithm: advantages and pitfalls. In TSH dynamics a series of independent trajectories are computed starting from a previously equilibrated population of initial configurations sampled at a given temperature, T, on a chosen PES, j (better would be the sampling of the corresponding Wigner distribution, which, in general, is however more difficult to compute[31]). All trajectories are classical in the sense that only classical forces are computed from as gradient of the selected PES. Together with the nuclear coordinates propagated with Eq. 24, the quantum amplitudes are also evolved in time using Eq. 32. In a region of coupling between two PES j and i the classical trajectories can eventually hop from one surface to another according to the probability g_{ij}. After a surface hop, the excess (deficient) energy due to the transition is redistributed into (extracted from) the motion along the direction of the nonadiabatic coupling vectors[32,33]. In this way, energy conser-

vation is guarantee along the entire trajectory. *Frustrated hops* occur when the quantum particles have insufficient kinetic energy to compensate the potential energy loss in upward transitions.

This method has the advantage to be simple to implement in any existing DFT/TDDFT code that offers the possibility to efficiently compute PESs, classical forces, and nonadiabatic couplings. These quantities can either be precomputed or can be evaluated *on-the-fly* along the growing trajectory. However, the first approach is only suited in the case of relatively small systems made of only few atoms for which the multidimensional $3N_n - 6$ potential energy hypersurfaces (or a restricted portion of them) can be computed for all relevant states. The *on-the-fly* approach reduces dramatically the number on electronic structure calculations to the number of integration steps of the classical nuclear dynamics and can therefore be used for the simulation of larger systems (made of up to thousands of atoms).

Due to the classical nature of the dynamics, TSH cannot describe nuclear tunneling and nuclear dephasing processes. The only nuclear quantum effects reproducible with TSH dynamics are those related to wavepacket splittings induced by avoided crossings and conical intersections (regions of strong nonadiabatic couplings). Due to the ITA and the local nature of the transitions (in space and time), quantum coherence and decoherence effects between states are difficult to capture in TSH dynamics. However, variations of TSH algorithm to circumvent these limitations are available[34].

Compared to Ehrenfest MD, the trajectories in TSH evolve on adiabatic PESs (except for the instantaneous transitions), which simplifies the physical interpretation of the results and the comparison with the experiments. In addition, TSH MD in the FSSH implementation obeys detailed balance approximately, with deviations that tend to vanish in the limits of small adiabatic splitting and the limit of large nonadiabatic couplings[35]. The use of the ITA together with the representation of the nuclear wavepacket amplitude by the density of trajectories offers a simple (even though approximated) solution to the problem of computing the high-dimensional derivatives $\nabla_\gamma \Omega_i(\boldsymbol{R}, t)$ in Eq. 27.

DFT/TDDFT-based TSH. Several implementations of on-the-fly TSH MD are nowadays available in different software packages and they mainly differ in the way the electronic structure calculations are preformed. Among the DFT-based TSH MD implementations, the method by Doltsinis and Marx[2] is based on the restricted open-shell formulation of the first singlet DFT excited state and is available in the software package CPMD[36]. More recently, Prezhdo et al. have developed a TSH MD scheme based on the time-dependent propagation of the KS orbitals, which are also used to approximate the many-electron Slater-type wavefunctions for the calculation of the nonadiabatic couplings[3]. This method is simple and efficient, but the description of excited states by means of excited KS Slater determinants is not rigorous and therefore the approximation done in the representation of the PESs and their couplings cannot be controlled and systematically improved. Finally, Tavernelli and coworkers have derived a TDDFT-based TSH MD scheme in which all ingredients required for the propagation of the trajectories (Eq. 24) and of the amplitudes (Eq. 32) are rigorously derived from TDDFT in the linear response formulation[4,37,30,38,39]. These also include the nonadiabatic coupling vectors computed for a pair of excited states and the coupling with external (time-dependent) fields[40–42]. For more information see the Sec. 3.

2.4 External Fields

The coupling of nonadiabatic MD with an external time-dependent electric field is given by the interaction Hamiltonian (with no spin-magnetic field contributions)

$$\hat{H}_{int} = \sum_{i=1}^{N_{el}} \left[-\frac{e}{2m_e c}(\hat{\boldsymbol{p}}_i \cdot \hat{\boldsymbol{A}}(\boldsymbol{r}_i, t) + \hat{\boldsymbol{A}}(\boldsymbol{r}_i, t) \cdot \hat{\boldsymbol{p}}_i) + \frac{e^2}{2m_e c^2} \hat{\boldsymbol{A}}(\boldsymbol{r}_i, t)\hat{\boldsymbol{A}}(\boldsymbol{r}_i, t) \right] , \quad (35)$$

where the vector potential $\boldsymbol{A}(\boldsymbol{r}_i, t)$ is related to the actual electric field by $\boldsymbol{E} = -\frac{1}{c}\frac{\partial \boldsymbol{A}}{\partial t}$. The summation in Eq. 35 is over all electrons, while the interaction with the nuclei is treated at the fully classical level and is therefore not included in the following derivations. $\hat{H}_{int}$ can be directly added to the Hamiltonian that governs the Ehrenfest dynamics (Eqs. 5 and 6)[21,22].

In TSH nonadiabatic MD[40,41], one needs to evaluate the radiation field coupling matrix elements

$$\langle \hat{H}_{int} \rangle_{ji} = i\omega_{ji} \frac{\boldsymbol{A}_0}{c} \cdot \boldsymbol{\mu}_{ji} e^{-i\omega t} \quad (36)$$

where $\boldsymbol{A}_0 = A_0 \boldsymbol{\epsilon}^{\lambda}$ and

$$\boldsymbol{\mu}_{ji} = -e\langle \Phi_j | \sum_{i=1}^{N_{el}} \hat{\boldsymbol{r}}_i | \Phi_i \rangle \quad (37)$$

is the the transition dipole vector, and $\omega_{ji} = (E_j - E_i)/\hbar$.

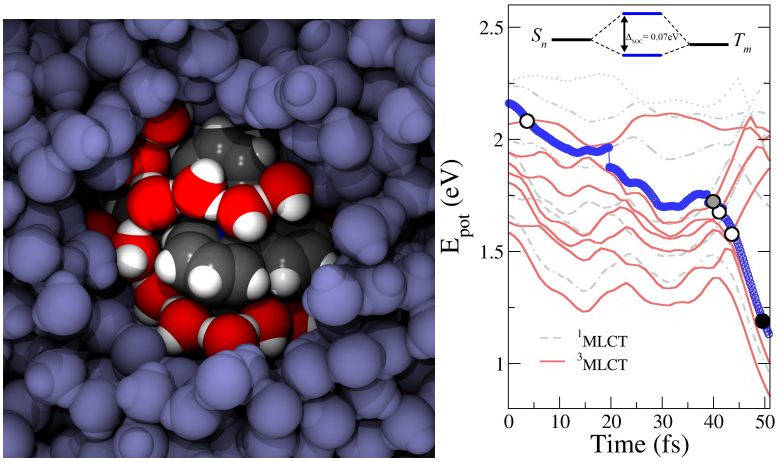


Figure 4. TDDFT-based TSH dynamics of Ruthenium tris-2,2'-bipyridine in water (left panel). The dynamics is initiated in the 5^{th} excited singlet state (right panel), which has a metal-to-ligand charge transfer (MLCT) character (all singlet states are shown with gray lines). Within the first $50fs$ we observe several intersystem crossings with triplet states (in red). The intensity of the spin orbit couplings (SOCs) is indicated with a color code: small (white circles), intermediate (gray circles), and strong (black circles). As observed experimentally, the system can undergo an ultrafast singlet to triple transition mediated by the solvent dynamics[43,44,42]. This example emphasizes the importance of including the calculation of SOC within TDDFT-based TSH. The inset in the right panel shows the splitting between a singlet and a triplet state computed with ZORA.

In particular, in the presence of an external radiation field, the differential equations for the TSH coefficients (Eq. 32) become

$$i\hbar\frac{dC_j^\alpha(t)}{dt} = \sum_i C_i(t)(H_{ji} - i\hbar\dot{\boldsymbol{R}}^\alpha \cdot \boldsymbol{d}_{ji}^\alpha(\boldsymbol{R}) + i\omega_{ji}\frac{\boldsymbol{A}_0}{c}\boldsymbol{\epsilon}^\lambda \cdot \boldsymbol{\mu}_{ji}^\alpha e^{-i\omega t}) \,. \tag{38}$$

In addition, a classical electrostatic interaction term of the form

$$E^{nucl}(\boldsymbol{R}^\alpha) = -\sum_\gamma Z_\gamma \boldsymbol{R}_\gamma^\alpha \cdot \boldsymbol{E}(t) \tag{39}$$

is used to couple the external field to the nuclear dynamics.

When the system of interest is coupled to its environment by means of a QM/MM setup, the electrostatic interaction between the QM subsystem (treated at TDDFT level) and the MM subsystem (classically described through an empirical Hamiltonian) is described by[42]

$$E_{QM/MM}^{el} = \sum_\gamma \int v_{sC}^\gamma(|\boldsymbol{R}_\gamma - \boldsymbol{r}|)(\rho_0(\boldsymbol{r}) + \delta\rho_i(\boldsymbol{r},t))d\boldsymbol{r} \tag{40}$$

where $\delta\rho_i(\boldsymbol{r},t)$ is the electron density perturbation induced by the transition into the i^{th} excited state. In Eq. 40, the potential $v_{sC}(|\boldsymbol{R}_\gamma - \boldsymbol{r}|)$ is a screened Coulomb potential generated by the atom at $\boldsymbol{R}_\gamma$ and modified at short range in order to avoid spurious over-polarization effects[46]

$$v_{sC}^\gamma(|\boldsymbol{R}_\gamma - \boldsymbol{r}|) = q_\gamma \frac{r_C^4 - (|\boldsymbol{R}_\gamma - \boldsymbol{r}|)^4}{r_C^5 - (|\boldsymbol{R}_\gamma - \boldsymbol{r}|)^5} \,, \tag{41}$$

where q_γ are the classical force-field charges of atom γ and r_C its covalent radius. In a QM/MM setup we therefore need, in addition to the calculation of the TDDFT energies, forces, and nonadiabatic couplings, an explicit evaluation of the TDDFT perturbed densities[42] (see Appendix).

3 TDDFT Quantities for Nonadiabatic Dynamics

In this chapter I will give a brief description of main electronic structure quantities used in nonadiabatic dynamics and their formulation within linear response TDDFT (LR-TDDFT).

3.1 Excited State Energies and Nuclear Forces from LR-TDDFT

The linear response formulation of TDDFT (LR-TDDFT) has become the method of choice for the calculation of excited state PESs in different nonadiabatic MD schemes.

In LR-TDDFT, excited state energies are computed from the poles of the many-body density response function. In matrix form, these energies are solution of the so-called Casida's equation (note that since the indices i and j are used to label the KS orbitals, in the following I will use n and m – instead of i and j as previously done – to index electronic states)

$$\begin{bmatrix} \mathbb{A} & \mathbb{B} \\ \mathbb{B}^* & \mathbb{A}^* \end{bmatrix} \begin{bmatrix} \boldsymbol{X}_n \\ \boldsymbol{Y}_n \end{bmatrix} = \omega_n \begin{bmatrix} \mathbb{I} & 0 \\ 0 & -\mathbb{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{X}_n \\ \boldsymbol{Y}_n \end{bmatrix} \tag{42}$$

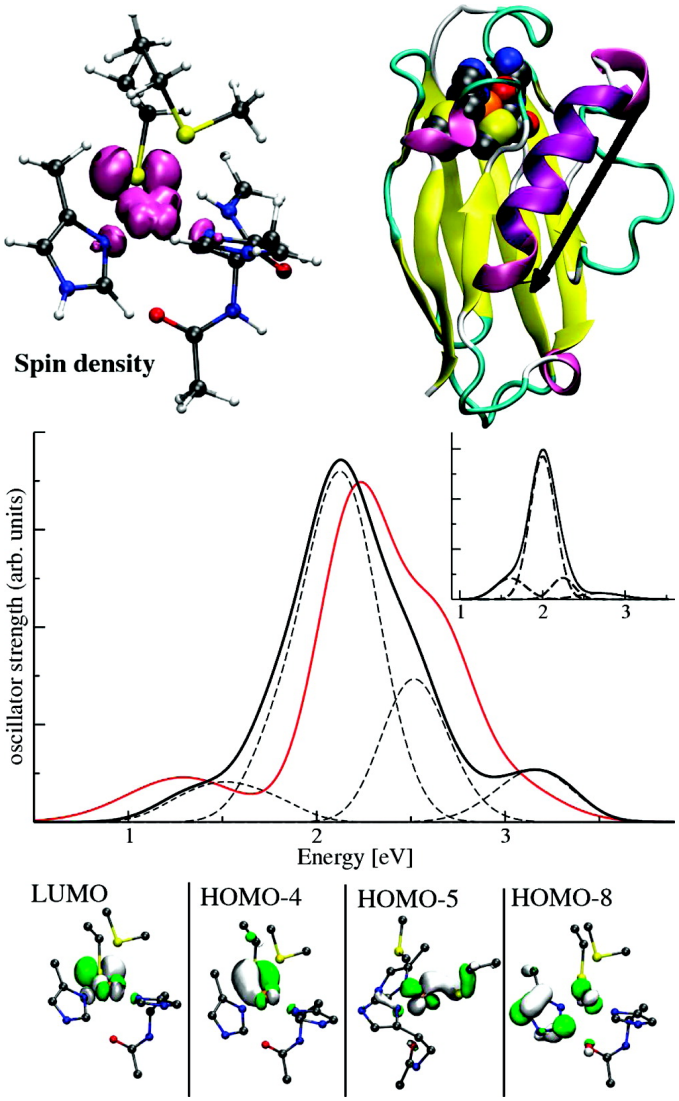


Figure 5. TDDFT/MM study of the absorption spectra of Azurin[45]. Upper panel: Distribution of the unpaired spin density at the binding site of azurin computed with DFT (PBE functional). Left: The electron spin density is drawn in mauve; MM atoms are not drawn for clarity. Right: Structure of azurin. The atoms of the copper binding site are drawn in spheres. The cartoon representation of the protein indicates the secondary structure elements. The black arrow specifies the direction of the electrostatic dipole produced by the β-helix. Middle panel: Absorption spectrum of Cu(II) azurin. The black line is the computed LR-TDDFT spectrum. A Gaussian decomposition of the spectrum, corresponding to tentative band assignments, is given as dashed lines. Red line: LR-TDDFT spectrum neglecting the electrostatic coupling to the MM part of the system. Right inset: Experimental spectrum. Bottom panels: Kohn-Sham wave functions for the α-spin states dominantly involved in the electronic excitations (HOMO-4 to LUMO: lower energy band; HOMO-5 to LUMO: central bands; HOMO-8 to LUMO: higher energy band).

where

$$A_{ij\sigma,kl\tau} = \delta_{\sigma,\tau}\delta_{i,k}\delta_{j,l}\frac{\varepsilon_{k\tau} - \varepsilon_{l\tau}}{f_{k\tau} - f_{l\tau}} - K_{ij\sigma,kl\tau}(\omega)\,, \tag{43}$$

$$B_{ij\sigma,kl\tau} = -K_{ij\sigma,lk\tau}(\omega)\,, \tag{44}$$

$$K_{ij\sigma,kl\tau}(\omega) = \int d\boldsymbol{r}d\boldsymbol{r}'\frac{\phi^*_{i\sigma}(\boldsymbol{r})\phi_{j\sigma}(\boldsymbol{r})\,\phi_{k\tau}(\boldsymbol{r}')\phi^*_{l\tau}(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|} + \int d(t-t')\,e^{i\omega(t-t')}$$
$$\int d\boldsymbol{r}d\boldsymbol{r}'\,\phi^*_{i\sigma}(\boldsymbol{r})\phi_{j\sigma}(\boldsymbol{r})\frac{\delta^2\mathcal{A}_{xc}[\rho]}{\delta\rho_\sigma(\boldsymbol{r},t)\,\delta\rho_\tau(\boldsymbol{r}',t')}\phi_{k\tau}(\boldsymbol{r}')\phi^*_{l\tau}(\boldsymbol{r}')\,. \tag{45}$$

In Eq. 42 we assume that the matrices $\mathbb{A}$ and $\mathbb{B}$ are frequency independent (adiabatic approximation for the TDDFT kernel, f_{xc}).

Other forms of the LR-TDDFT equations exist like, for instance, the one proposed by Gross[47] and the one based on Sternheimer's time-dependent perturbation theory[48–50].

The calculation of analytic nuclear gradients (forces) within the LR-TDDFT formalism is essential to all mixed quantum-classical MD schemes. Among the different approaches developed for the calculation of analytical derivatives, the Lagrangian method[51] is of particular interest because of its compact form. However, the derivation of LR-TDDFT is technically involved and since it does not bring any new physical insights, I simply refer the reader to the reach literature on the subject[10,50,52,53].

3.2 The Auxiliary Many-Electron Wavefunction

It may be useful at this point to investigate the possibility to further simplify the calculation of matrix elements within LR-TDDFT by means of the definition of a set of "auxiliary" multideterminantal many electron wavefunctions based on Kohn-Sham (KS) orbitals. This route was first explored by Casida[54] to solve the assignment problem of the LR-TDDFT excited state transitions and then further developed by Tavernelli et al.[4,37,30,38] in relation to the calculation of matrix elements in the linear and second order response regimes[39].

In Ref. 38, we showed that defining the ground state many electron wavefunction as a Slater determinant of all occupied Kohn-Sham orbitals $\{\phi_i\}_{i=1}^N$

$$\langle \boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_3, \ldots, \boldsymbol{r}_N|\tilde{\Psi}_0\rangle = \frac{1}{\sqrt{N}}det|\phi_i(\boldsymbol{r}_1)\phi_1(\boldsymbol{r}_2)\phi_2(\boldsymbol{r}_3), \ldots, \phi_N(\boldsymbol{r}_N)| \tag{46}$$

and the excited state wavefunction corresponding to the excitation energy ω_n as

$$\langle \boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_3, \ldots, \boldsymbol{r}_N|\tilde{\Psi}_n\rangle = \sum_{ia\sigma}\sqrt{\frac{\varepsilon_a - \varepsilon_i}{\omega_n}}(\boldsymbol{Z}_n)_{ia\sigma}\langle \boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_3, \ldots, \boldsymbol{r}_N|\tilde{\Psi}_0\rangle \tag{47}$$

we obtain for any one-body operator of the form $\hat{\mathcal{O}} = \sum_{ij\sigma} o_{ij\sigma}\hat{a}^\dagger_{j\sigma}\hat{a}_{i\sigma}$ the correct linear response expression for the matrix element $\langle\Psi_0|\hat{\mathcal{O}}|\Psi_n\rangle$. In Eq. 47 the index i runs over all occupied and a over the unoccupied (virtual) KS orbitals, $\boldsymbol{Z}_n = (\mathbb{A}-\mathbb{B})^{-1/2}(\boldsymbol{X}_n + \boldsymbol{Y}_n)$, and $\hat{a}^\dagger_{i\sigma}$ and $\hat{a}_{i\sigma}$ are the creation and the annihilation operators for the KS orbital $\phi_{i\sigma}(\boldsymbol{r})$, respectively. This theory was then successfully extended to the case of the calculation of matrix elements between two excited state wavefunctions, $\langle\Psi_n|\hat{\mathcal{O}}|\Psi_m\rangle$ as will be shown in the next chapter on the calculation of nonadiabatic coupling vectors.

It is important to stress the fact that both auxiliary functions introduced in Eqs. 46 and 47 have only a physical meaning when used within LR-TDDFT for the calculation of matrix elements of the type $\langle\tilde{\Psi}_0|\hat{\mathcal{O}}|\tilde{\Psi}_n\rangle$ and eventually $\langle\tilde{\Psi}_n|\hat{\mathcal{O}}|\tilde{\Psi}_m\rangle$ (see Ref. 39). The use of these representations for the many-electron ground and excited states wavefunctions in other contexts is not justified. In particular, $|\tilde{\Psi}_0\rangle$ has little to do with the ground state wavefunction of the system (and as an approximation, it is even worse that the Hartree-Fock Slater determinant).

3.3 The Nonadiabatic Coupling Vectors in LR-TDDFT

Traditionally, the computation of the nonadiabatic coupling vectors (NACVs) is carried out using wavefunction-based *ab initio* quantum chemistry approaches (MRCISD, CCSD), which, however, are not well suited for applications in the condensed phase and become computationally unaffordable when large molecular systems are considered.

3.3.1 The Couplings with the Ground State

We start from the definition of the NACV between the ground state and the n^{th} excited state for a molecular system characterized by nuclear coordinates $\boldsymbol{R}$ in the configuration space ($\mathbb{R}^{3N_n}$)

$$\boldsymbol{d}_{0n,\mu} = -\frac{\langle\Psi_0(\boldsymbol{R})|\nabla_\mu\hat{\mathcal{H}}_{el}|\Psi_n(\boldsymbol{R})\rangle}{\epsilon_0(\boldsymbol{R}) - \epsilon_n(\boldsymbol{R})} \tag{48}$$

where μ is an atomic label, $\hat{\mathcal{H}}_{el}$ is the molecular Hamiltonian, and $\nabla_\mu\hat{\mathcal{H}}_{el} = \partial\hat{\mathcal{H}}_{el}/\partial\boldsymbol{R}_\mu$.

Applying the results of Sec. 3.2 on the evaluation of matrix elements of the form $\langle\Psi_0|\hat{\mathcal{O}}|\Psi_n\rangle$ in LR-TDDFT to the NACV gives directly the desired expression

$$\boldsymbol{d}_{0n,\mu} = \sum_{ij\sigma}^{(f_{i\sigma}-f_{j\sigma})>0} \frac{1}{\sqrt{\omega_I}} h^\mu_{ij\sigma}(\mathbb{S}^{-1/2}\boldsymbol{Z}_n)_{ij\sigma} \tag{49}$$

where $h^\mu_{ij\sigma} = \int d\boldsymbol{r}\,\partial_\mu\hat{\mathcal{H}}_{el}(\boldsymbol{R})\,\phi^*_{i\sigma}(\boldsymbol{r})\phi_{j\sigma}(\boldsymbol{r})$, $\mathbb{S}^{-1/2} = -\mathbb{C}(\mathbb{A}-\mathbb{B})^{-1}\mathbb{C}$, and $C_{ij\sigma,kl\tau} = (\delta_{\sigma,\tau}\delta_{i,k}\delta_{j,l})/(f_{k\tau} - f_{l\tau})$.

This formula for the NACVs within LR-TDDFT was derived several times in the literature using slightly different formalisms. The first derivation was by Chernyak and Mukamel[55] using a classical Liouville dynamics for the single-electron density matrix. Later, Tavernelli et al.[4,37] and Hu et al. [56] arrived to the same result (Eq. 49) using the most widely used formulation based on Casida's LR-TDDFT equations[54].

Concerning the numerical implementation of Eq. 48 several approaches have also been proposed that differ mainly in the choice of the basis set and in the way the implicit dependence of the pseudopotentials on the nuclear positions is treated. Due to the technical nature of this subject, we will not go through the numerical details but better refer to the literature[4,37,56,57].

3.3.2 The Finite Difference Formulation of the Nonadiabatic Couplings

In the TSH dynamics, the nonadiabatic coupling terms appears as a scalar product of the NACVs with the particle velocities, $\sigma_{0n} = \boldsymbol{d}_{0n} \cdot \dot{\boldsymbol{R}}$, in the time evolution of the amplitudes associated to the different states. Starting from an equivalent definition of the NACV,

$$\langle \Psi_0(\boldsymbol{r}; \boldsymbol{R}(t))|\nabla_\mu|\Psi_n(\boldsymbol{r}; \boldsymbol{R}(t))\rangle \tag{50}$$

we therefore obtain[4,38]

$$\begin{aligned}\sigma_{0n}|_{t+\delta t/2} &= \sum_\mu \langle \Psi_0(\boldsymbol{r}; \boldsymbol{R}(t))|\nabla_\mu|\Psi_n(\boldsymbol{r}; \boldsymbol{R}(t))\rangle \dot{\boldsymbol{R}}_\mu = \langle \Psi_0(\boldsymbol{r}; \boldsymbol{R}(t))|\frac{\partial}{\partial t}|\Psi_n(\boldsymbol{r}; \boldsymbol{R}(t))\rangle \\ &\simeq \frac{1}{2\delta t}[\langle \Psi_0(\boldsymbol{r}; \boldsymbol{R}(t))|\Psi_n(\boldsymbol{r}; \boldsymbol{R}(t+\delta t))\rangle - \langle \Psi_0(\boldsymbol{r}; \boldsymbol{R}(t+\delta t))|\Psi_n(\boldsymbol{r}; \boldsymbol{R}(t))\rangle]\,,\end{aligned} \tag{51}$$

where $|\Psi_0\rangle$ and $|\Psi_n\rangle$ are evaluated at subsequent time t and $t + \delta t$ using the auxiliary wavefunctions defined in Eqs. 46 and 46, respectively. Since $\sigma_{0n}|_{t+\delta t/2}$ is the only quantity needed in the evaluation of the surface hopping probability in the time interval $[t, t + \delta t]$ between the electronic states $|\tilde{\Psi}_0\rangle$ and $|\tilde{\Psi}_n\rangle$, it is numerically more efficient to use Eq. 51 instead of the cumbersome evaluation of $\boldsymbol{d}_{0n,\mu}$ followed by the multiplication with the particle velocities, $\boldsymbol{d}_{0n,\mu} \cdot \dot{\boldsymbol{R}}_\mu$. However, at a "surface hop", the calculation of the full NACVs is required for the redistribution of the potential energy difference in order to guarantee energy conservation.

3.3.3 Nonadiabatic Couplings between Excited States

In the excited state nonadiabatic dynamics of molecular systems we also need to compute NACVs between pairs of excited states. These are beyond the reach of linear response theory and therefore cannot be evaluated using Eq. 49. A second order response theory for the evaluation of matrix elements of the form $\langle \Psi_n|\hat{\mathcal{O}}|\Psi_m\rangle$ within TDDFT was first derived by Mukamel and co-workers[58–60] using an approximate mapping of the original electronic problem into a boson system sharing the same response properties.

The use of the "auxiliary" electronic wavefunctions introduced in Sec. 3.2 offers a valid alternative to this approach and produces second order matrix elements that include contributions from the de-excitation of the correlated ground state[39], which are neglected in the derivation given in Ref. 61. In fact, the two approaches coincide in the TDA up to terms of third order in $\boldsymbol{Z}_m$[39].

For the calculation of the NACVs between a pair of excited states (PESs), E_n^{el} and E_m^{el} within the TDDFT based TSH dynamics of Sec. 2.3, we therefore use the expression

$$\begin{aligned}\langle \Psi^n|\hat{\mathcal{O}}|\Psi^m\rangle &= \sum_{ia}\sum_{jb} c_{ia}^{n\dagger}\, c_{jb}^{m} \langle \tilde{\Psi}_{ia}^{n}|\hat{\mathcal{O}}|\tilde{\Psi}_{jb}^{m}\rangle \\ &= \sum_{iab} c_{ia}^{n\dagger}\, c_{ib}^{m} \langle \psi_a|\hat{\mathcal{O}}|\psi_b\rangle - \sum_{aij} c_{ia}^{n\dagger}\, c_{ja}^{m} \langle \phi_i|\hat{\mathcal{O}}|\phi_j\rangle\,,\end{aligned} \tag{52}$$

with $\hat{\mathcal{O}}$ replaced by $\nabla_\mu \hat{\mathcal{H}}_{mol}$ and $c_{ia}^n = \sqrt{(\varepsilon_a - \varepsilon_i)/\omega_n}(\boldsymbol{Z}_n)_{ia}$. The quality of these matrix elements was assessed in Ref. 39.

Acknowledgments

I thank Basile Curchod for his contributions and his critical reading of the lecture notes. Many thanks also to Ursula Röthlisberger for her support and the entire LCBC group for the help with the applications.

Appendix

The LR-TDDFT Response Density

Within Casida's formulation, the density response $\delta\rho_n(\boldsymbol{r},t)$ can be expanded in the auxiliary many-electron wavefunctions of Eq. 47

$$\rho_n(\boldsymbol{r},t) = \langle\tilde{\Psi}_n|\sum_{\kappa=1}^{N}\delta(\boldsymbol{r}-\boldsymbol{r}_\kappa)|\tilde{\Psi}_n\rangle = \rho_0(\boldsymbol{r}) + \delta\rho_n(\boldsymbol{r},t) \tag{53}$$

which in first order becomes

$$\begin{aligned}\delta\rho_n(\boldsymbol{r},t) &= \sum_{ia} c_{ia}^n \langle\tilde{\Psi}_0|\sum_{\kappa=1}^{N}\delta(\boldsymbol{r}-\boldsymbol{r}_\kappa)|\hat{a}_a^\dagger\hat{a}_i\tilde{\Psi}_0\rangle\, e^{-i\Omega_I t} + \text{c.c.} \\ &= \sum_{ia} c_{ia}^n \rho'(\boldsymbol{r}) e^{-i\omega_n t} + \text{c.c.} \\ &= \sum_{pq} f_p\tilde{f}_q c_{pq}^n \rho'(\boldsymbol{r}) e^{-i\omega_n t} + \text{c.c.}\end{aligned} \tag{54}$$

where N is the number of electrons and $\tilde{f}_p = (1-f_p)$.

Using the LR-TDDFT equation for the transition density $\rho'(\boldsymbol{r}) = \sum_{pq} f_p\tilde{f}_q(\boldsymbol{X}_n + \boldsymbol{Y}_n)_{pq}\phi_p(\boldsymbol{r})\psi_q(\boldsymbol{r})$ and the definition of the coefficients c_{pq}^n [4,37,62] we finally get, in agreement with Ref. 63, 62,

$$\delta\rho_n(\boldsymbol{r}) = \sum_{pq}\Delta P_{pq}^n \xi_p^*(\boldsymbol{r})\xi_q(\boldsymbol{r}) \tag{55}$$

where

$$\begin{aligned}\Delta P_{pq}^n = &- f_p f_q\left(\sum_a X_{pa}^{n\dagger}X_{qa}^n + \sum_a Y_{qa}^{n\dagger}Y_{pa}^n\right) \\ &+ \tilde{f}_p\tilde{f}_q\left(\sum_i X_{iq}^{n\dagger}X_{ip}^n + \sum_i Y_{ip}^{n\dagger}Y_{iq}^n\right).\end{aligned} \tag{56}$$

Within the Sternheimer framework, the response density matrix elements in TDA are given by[50]

$$\Delta P_{pq}^n = \sum_i x_{qi}^n (x^n)_{pi}^* + \sum_{rij} x_{rj}^n c_{pj}^{\{0\}}(c_{qi}^{\{0\}})^*(x_{ri}^n)^*\,. \tag{57}$$

For a detailed account of the implementation see Refs. 50, 37.

References

1. W. Kohn and L. J. Sham, *Self-consistent equations including exchange and correlation effects*, Phys. Rev., **140**, 1133, 1965.
2. N. L. Doltsinis and D. Marx, *Nonadiabatic Car-Parrinello Molecular Dynamics*, Phys. Rev. Lett., **88**, 166402, 2002.
3. C. F. Craig, W. R. Duncan, and O. V. Prezhdo, *Trajectory surface hopping in the time-dependent Kohn-Sham approach for electron-nuclear dynamics*, Phys. Rev. Lett., **95**, 163001, 2005.
4. E. Tapavicza, I. Tavernelli, and U. Rothlisberger, *Trajectory surface hopping within linear response time-dependent density-functional theory*, Phys. Rev. Lett., **98**, 023001, 2007.
5. J. C. Tully, *Molecular dynamics with electronic transitions*, J. Chem. Phys., **93**, 1061–1071, 1990.
6. Illia Horenko, Christian Salzmann, Burkhard Schmidt, and Christof Schutte, *Quantum-classical Liouville approach to molecular dynamics: Surface hopping Gaussian phase-space packets*, J. Chem. Phys., **117**, no. 24, 11075–11088, 2002.
7. Yasuki Arasaki, Kazuo Takatsuka, Kwanghsi Wang, and Vincent McKoy, *Pump-Probe Photoionization Study of the Passage and Bifurcation of a Quantum Wave Packet Across an Avoided Crossing*, Phys. Rev. Lett., **90**, 248303–, 2003.
8. Courtney L. Lopreore and Robert E. Wyatt, *Quantum Wave Packet Dynamics with Trajectories*, Phys. Rev. Lett., **82**, no. 26, 5190, 1999.
9. H. D. Meyer, U. Manthe, and L. S. Cederbaum, *The multi-configurational time-dependent hartree approach*, Chem. Phys. Lett., **165**, 73–78, 1990.
10. D. Marx and J. Hutter, *Ab Initio Molecular Dynamics: Basic Theory and Advanced Methods*, Cambridge University Press, 2009.
11. B.F.E. Curchod, U. Rothlisberger, and I. Tavernelli, In preparation, 2012.
12. J. Theilhaber, *Ab initio simulations of sodium using time-dependent density-functional theory*, Physical Review B, **46**, no. 20, 12990, 1992.
13. I. Tavernelli, U.F. Röhrig, and U. Rothlisberger, *Molecular dynamics in electronically excited states using time-dependent density functional theory*, Molecular Physics, **103**, no. 6-8, 963–981, 2005.
14. E. Runge and E. K. U. Gross, *Density-functional theory for time-dependent systems*, Phys. Rev. Lett., **52**, 997–1000, 1984.
15. R. van Leeuwen, *Causality and symmetry in time-dependent density-functional theory*, Physical review letters, **80**, no. 6, 1280–1283, 1998.
16. G. Vignale, *Real-time resolution of the causality paradox of time-dependent density-functional theory*, Physical Review A, **77**, no. 6, 062511, 2008.
17. Yair Kurzweil and Roi Baer, *Time-dependent exchange-correlation current density functionals with memory*, The Journal of Chemical Physics, **121**, no. 18, 8731–8741, 2004.
18. H. O. Wijewardane and C. A. Ullrich, *Real-Time Electron Dynamics with Exact-Exchange Time-Dependent Density-Functional Theory*, Phys. Rev. Lett., **100**, 056404, Feb 2008.
19. R. Car and M. Parrinello, *Unified Approach for Molecular Dynamics and Density-Functional Theory*, Phys. Rev. Lett., **55**, 2471, 1985.

20. G.D. Billing, Int. Rev. Phys. Chem., **13**, 309, 1994.
21. K. Yabana and G. F. Bertsch, *Time-dependent local-density approximation in real time*, Phys. Rev. B, **54**, 4484–4487, Aug 1996.
22. I. Tavernelli, *Electronic density response of liquid water using time-dependent density functional theory*, Phys. Rev. B, **73**, 094204, 2006.
23. I. Tavernelli, M.P. Gaigeot, R. Vuilleumier, C. Stia, M. A. H. Penhoat, and M. F. Politis, *Time-dependent density functional theory molecular dynamics simulations of liquid water radiolysis*, ChemPhysChem, **9**, 2099, 2008.
24. Ivano Tavernelli, Marie-Pierre Gaigeot, Rodolphe Vuilleumier, Carlos Stia, Marie-Anne HervÃ© du Penhoat, and Marie-FranÃ§oise Politis, *Time-Dependent Density Functional Theory Molecular Dynamics Simulations of Liquid Water Radiolysis*, Chem. Phys. Chem., **9**, no. 14, 2099–2103, 2008.
25. M. P. Gaigeot, P. Lopez-Tarifa, F. Martin, M. Alcami, R. Vuilleumier, I. Tavernelli, M. A. Hervédu Penhoat, and M. F. Politis, *Theoretical investigation of the ultrafast dissociation of ionised biomolecules immersed in water: Direct and indirect effects*, Mutation Research/Reviews in Mutation Research, **704**, no. 1-3, 45–53, 2010.
26. P. Lopez-Tarifa, M.-A. Herve du Penhoat, R. Vuilleumier, M.-P. Gaigeot, I. Tavernelli, A. Le Padellec, J.-P. Champeaux, M. Alcami, P. Moretto-Capelle, F. Martin, and M.-F. Politis, *Ultrafast Nonadiabatic Fragmentation Dynamics of Doubly Charged Uracil in a Gas Phase*, Phys. Rev. Lett., **107**, 023202, Jul 2011.
27. G.E. Bowman, *On the classical limit in Bohms theory*, Foundations of Physics, **35**, no. 4, 605–625, 2005.
28. Peter R. Holland, *The Quantum Theory of Motion - An Account of the de Broglie-Bohm Causal Interpretation of Quantum Mechanics*, Cambridge University Press, 1993.
29. R. K. Preston and J. C. Tully, J. Chem. Phys., **54**, 4297, 1971.
30. I. Tavernelli, E. Tapavicza, and U. Rothlisberger, *Non-adiabatic dynamics using time-dependent density functional theory: Assessing the coupling strengths*, J. Mol. Struc. (Theochem), **914**, 22–29, 2009.
31. DC Marinica, MP Gaigeot, and D Borgis, *Generating approximate Wigner distributions using Gaussian phase packets propagation in imaginary time*, CHEMICAL PHYSICS LETTERS, **423**, no. 4-6, 390–394, JUN 1 2006.
32. Philip Pechukas, *Time-Dependent Semiclassical Scattering Theory. I. Potential Scattering*, Phys. Rev., **181**, 166–174, May 1969.
33. DF Coker and L. Xiao, *Methods for molecular dynamics with nonadiabatic transitions*, The Journal of chemical physics, **102**, 496, 1995.
34. J.C. Tully, *Nonadiabatic molecular dynamics*, International Journal of Quantum Chemistry, **40**, no. S25, 299–309, 1991.
35. JR Schmidt, P.V. Parandekar, and J.C. Tully, *Mixed quantum-classical equilibrium: Surface hopping*, The Journal of chemical physics, **129**, 044104, 2008.
36. *CPMD*, Copyright IBM Corp 1990-2001, Copyright MPI für Festkörperforschung Stuttgart, 1997-2001, `http://www.cpmd.org`.
37. I. Tavernelli, E. Tapavicza, and U. Rothlisberger, *Nonadiabatic coupling vectors within linear response time-dependent density functional theory*, J. Chem. Phys., **130**, 124107, 2009.
38. I. Tavernelli, B. F. E. Curchod, and U. Rothlisberger, *On nonadiabatic coupling vectors in time-dependent density functional theory*, J. Chem. Phys., **131**, 196101, 2009.

39. I. Tavernelli, B. F. E. Curchod, A. Laktionov, and U. Rothlisberger, *Nonadiabatic coupling vectors for excited states within time-dependent density functional theory in the Tamm-Dancoff approximation and beyond*, J. Chem. Phys., **133**, 194104, 2010.
40. I. Tavernelli, B. F. E. Curchod, and U. Rothlisberger, *Mixed quantum-classical dynamics with time-dependent external fields: A time-dependent density-functional-theory approach*, Phys. Rev. A, **81**, 052508, 2010.
41. B.F.E. Curchod, T.J. Penfold, U. Rothlisberger, and I. Tavernelli, *Local control theory in trajectory-based nonadiabatic dynamics*, Physical Review A, **84**, no. 4, 042507, 2011.
42. I. Tavernelli, B.F.E. Curchod, and U. Rothlisberger, *Nonadiabatic Molecular Dynamics with Solvent Effects: a LR-TDDFT QM/MM Study of Ruthenium (II) Tris (bipyridine) in Water*, Chemical Physics, **391**, 101, 2011.
43. Marc-Etienne Moret, Ivano Tavernelli, and Ursula Rothlisberger, *Combined QM/MM and classical molecular dynamics study of [Ru(bpy)3]2+ in water.*, J. Phys. Chem. B, **113**, no. 22, 7737–7744, Jun 2009.
44. Marc-Etienne Moret, Ivano Tavernelli, Majed Chergui, and Ursula Rothlisberger, *Electron localization dynamics in the triplet excited state of [Ru(bpy)3]2+ in aqueous solution.*, Chem. Eur. J., **16**, no. 20, 5889–5894, May 2010.
45. M. Cascella, M. L. Cuendet, I. Tavernelli, and U. Rothlisberger, *Optical Spectra of Cu(II)-Azurin by Hybrid TDDFT-Molecular Dynamics Simulations*, J. Phys. Chem. B, **111**, no. 34, 10239–10247, 2007.
46. Alessandro Laio, Joost VandeVondele, and Ursula Rothlisberger, *A Hamiltonian electrostatic coupling scheme for hybrid Car–Parrinello molecular dynamics simulations*, J. Chem. Phys., **116**, no. 16, 6941–6947, 2002.
47. M. Petersilka, U. J. Gossmann, and E. K. U. Gross, *Excitation Energies from Time-Dependent Density-Functional Theory*, Phys. Rev. Lett., **76**, 1212–1215, 1996.
48. R. Sternheimer, *On Nuclear Quadrupole Moments*, Phys. Rev., **84**, no. 2, 244–253, 1951.
49. F. Furche and R. Ahlrichs, *Adiabatic time-dependent density functional methods for excited state properties*, J. Chem. Phys., **117**, 7433–7447, 2002.
50. J. Hutter, *Excited state nuclear forces from the TammâDancoff approximation to time-dependent density functional theory within the plane wave basis set framework*, J. Chem. Phys., **118**, 3928–3934, 2003.
51. T. Helgaker and P. Jorgensen, *Configuration-interaction energy derivatives in a fully variational formulation*, Theor. Chim. Acta, **75**, 111–127, 1989.
52. P. Pulay, *Analytical derivative methods in quantum chemistry*, Advances in Chemical Physics, **69**, 241–286, 1987.
53. Peter Deglmann, Filipp Furche, and Reinhart Ahlrichs, *An efficient implementation of second analytical derivatives for density functional methods*, Chemical Physics Letters, **362**, no. 5-6, 511 – 518, 2002.
54. M. E. Casida, *Time-dependent density-functional response theory for molecules*, in: Recent Advances in Density Functional Methods, D. P. Chong, (Ed.), p. 155, Singapore, World Scientific. 1995.
55. V. Chernyak and S. Mukamel, *Density-matrix representation of nonadiabatic couplings in time-dependent density functional (TDDFT) theories*, J. Chem. Phys., **112**, 3572–3579, 2000.

56. C. P. Hu, H. Hirai, and O. Sugino, *Nonadiabatic couplings from time-dependent density functional theory: Formulation in the Casida formalism and practical scheme within modified linear response*, J. Chem. Phys., **127**, 064103, 2007.
57. Robert Send and Filipp Furche, *First-order nonadiabatic couplings from time-dependent hybrid density functional response theory: Consistent formalism, implementation, and performance.*, J Chem Phys, **132**, no. 4, 044107, Jan 2010.
58. S Tretiak, V Chernyak, and S Mukamel, *Excited electronic states of carotenoids: Time-dependent density-matrix-response algorithm*, Int J Quantum Chem, **70**, no. 4-5, 711–727, Jan 1998.
59. V Chernyak and S Mukamel, *Bosonized squeezed-state coupled-cluster approach to electron correlations in nonlinear spectroscopy*, J Chem Phys, **111**, no. 10, 4383–4396, Jan 1999.
60. Sergei Tretiak and Shaul Mukamel, *Density Matrix Analysis and Simulation of Electronic Excitations in Conjugated and Aggregated Molecules*, Chem. Rev., **102**, no. 9, 3171–3212, 2002.
61. O Berman and S Mukamel, *Quasiparticle density-matrix representation of nonlinear time-dependent density-functional response functions*, Phys Rev A, **67**, no. 4, 042503, Jan 2003.
62. Mark E. Casida, *Time-dependent density-functional theory for molecules and molecular solids*, Journal of Molecular Structure: THEOCHEM, **914**, no. 1-3, 3 – 18, 2009, Time-dependent density-functional theory for molecules and molecular solids.
63. Andrei Ipatov, Felipe Cordova, Loïc Joubert Doriol, and Mark E. Casida, *Excited-state spin-contamination in time-dependent density-functional theory for molecules with open-shell ground states*, Journal of Molecular Structure: THEOCHEM, **914**, no. 1-3, 60 – 73, 2009, Time-dependent density-functional theory for molecules and molecular solids.

Hybrid Car-Parrinello Molecular Dynamics / Molecular Mechanics Simulations: A Powerful Tool for the Investigation of Biological Systems

Emiliano Ippoliti*[1], Jens Dreyer*[1], Paolo Carloni[1,2], and Ursula Röthlisberger[3]

[1] Computational Biophysics
German Research School for Simulation Sciences[§]

[2] Institute for Advanced Simulation
Forschungszentrum Jülich, D-52425 Jülich, Germany
E-mail: p.carloni@grs-sim.de

[3] Computational Chemistry and Biochemistry
EPFL – École Polytechnique Fédérale de Lausanne, 1015 Ecublens, Switzerland
E-mail: ursula.roethlisberger@epfl.ch

Hybrid Car-Parrinello molecular dynamics/molecular mechanics (CPMD/MM) simulations are now extensively used to investigate biological systems. Here, after introducing some basics general aspects of hybrid quantum mechanics/molecular mechanics methods, we provide an in-depth discussion of the CPMD/MM method. We mention possible pitfalls and main limitations of the approach. We close these lecture notes with a couple of recent applications to systems of biological and pharmacological relevance.

1 Introduction

Density functional theory (DFT) is a widely applied quantum chemical method for the investigation of biological systems. It scales favorably with the number of electrons and the accuracy of the employed exchange-correlation functionals, which contains all the intricacies of the many-body problem, is constantly improving[1–5]. Its scope was further enlarged in 1985, when Car and Parrinello (CP) proposed a unified scheme for DFT and molecular dynamics (MD)[6]. By treating the electronic degrees of freedom as dynamical variables they managed to describe the time evolution of molecular systems (presently up to almost 2000 atoms)[7] without resorting to a force field[8,9]. The method enabled new types of realistic simulations for many different kinds of systems.[a]

Most systems of biological relevance are large: for instance, a system containing a protein in aqueous solution may consist of 10^4 to 10^5 atoms. To deal with these systems, hybrid Car-Parrinello molecular dynamics/molecular mechanics (CPMD/MM) schemes have been introduced. These follow the original quantum mechanical/molecular mechanical[b] (QM/MM) approach proposed by Warshel and Levitt[11]: a region of interest (e.g. an enzymatic active site) is described at the DFT level, mechanically and electrostatically coupled

[§]Joint venture of RWTH Aachen University and Forschungszentrum Jülich, Germany.
*These authors contributed equally to this work.
[a]Currently, Born–Oppenheimer approaches to first principles MD[8] are also widely and efficiently used[10].
[b]Several current QM/MM schemes, as the CPMD/MM one, are actually performing molecular dynamics at finite temperatures and not just geometry optimizations as the term "molecular mechanics" could suggest.

with the rest of the system treated using biomolecular force fields like AMBER[12], GROMOS[13] or CHARMM[14]. Most current CPMD/MM applications in biophysics employ the approach developed by Rothlisberger and co-workers[16], in which the CPMD program[17] is used for the QM part and the classical part, calculated with routines from Gromos96 code[13], is described either by GROMOS or AMBER force field.[c] The next section opens with a discussion of the general features of QM/MM methods. It is followed by a detailed description of the CPMD/MM approach. The paper closes with a very brief discussion of a couple of applications to biomolecular systems. A significant portion of the material presented in these lecture notes has been already reported by some of us (PC and UR) in Ref. 15.

2 Methods

2.1 General Features of QM/MM Methods

There are two fundamentally different ways to carry out calculations on a system that has been partitioned into a QM and a MM region.

In the **subtractive scheme**, the QM calculation is performed on an isolated QM system and the environment effects (i.e. the influence of the MM system on the QM system) are estimated at the lower level of theory by the difference between two MM calculations, one treating the entire system (QM+MM) and one the QM region only. In this approach, the total energy of the embedded system is written as

$$E = E^{QM}(QM) + E^{MM}(QM + MM) - E^{MM}(QM). \tag{1}$$

The force F_I acting on atom I at position R_I reads:

$$F_I = -\frac{\partial E}{\partial \mathbf{R}_I} = -\frac{\partial E^{QM}(QM)}{\partial \mathbf{R}_I} - \frac{E^{MM}(QM + MM)}{\partial \mathbf{R}_I} + \frac{E^{MM}(QM)}{\partial \mathbf{R}_I} \tag{2}$$

for the force F_I acting on atom I at position R_I. A QM/MM implementation that uses a subtractive scheme is the integrated molecular orbital molecular mechanics (IMOMM) scheme developed by Maseras and Morokuma[18] available in the Gaussian program[19]. Note that in a subtractive scheme, all calculations are performed with "pure" (either fully QM or fully MM) Hamiltonians. The advantage of such an approach lies in the fact that there is no QM/MM interface that has to be dealt with. The disadvantage is that the environment influence is often described at a very simple level: The electrostatic interactions between the QM and the MM parts is described entirely at the force field level, i.e. by Coulomb interactions between effective point charges. Such an electrostatic coupling between QM and MM part is called "mechanical coupling". This indicates that, in such an approach, the electrons do not feel anything of the classical electrostatic field of the environment and all the electrostatic interactions between QM and MM part act solely on the level of the atoms. The influence of the environment as described by the lower level method is only a reasonable estimate for the environment effect at the higher level if the two descriptions are not too different.

[c] A similar hybrid approach has been also implemented in the CP2K code[41].

To minimize the difference in the treatment of the two regions, the original IMOMM scheme has been extended to three (respectively multiple) layers (ONIOM)[20]. A typical ONIOM calculation consists for example of a first-principles (DFT or wavefunction-based method) region, adjacent to a layer treated with a semi-empirical method followed by a third layer treated at the molecular mechanics level.

In the **additive scheme**, more often used than the subtractive scheme, the system is described by a single hybrid Hamiltonian

$$H = H_{QM} + H_{MM} + H_{QM/MM}, \tag{3}$$

where H_{QM} is the quantum Hamiltonian, H_{MM} is the molecular mechanics Hamiltonian, and $H_{QM/MM}$ is the interaction Hamiltonian between QM and MM system. The lowest eigenvalue of the Hamiltonian in Eq. 3 determines the total energy E of the mixed quantum/classical system

$$E = E_{QM} + E_{MM} + E_{QM/MM}. \tag{4}$$

The advantage of an additive scheme is that the QM calculation can be directly executed in the presence of the classical environment in such a way that the electron density of the QM system is optimized in (and polarized by) the external electrostatic field of the surroundings. The prize for this is that the real system is replaced by a somewhat artificial, heterogeneous construct, in which different parts of the system are described at largely disparate levels, i.e. one part of the system is represented in electronic detail, whereas all the surroundings is reduced to a purely classical (mechanical and electrostatic) description. In this way, an abrupt QM/MM border is created. One of the drastic consequences of this approach is the fact that when passing from the QM to the MM zone of the system the electrons suddenly cease to exist. Such a simplified description can necessarily only constitute a somewhat crude representation of the true uniform system. The rest of these notes deals only with this scheme.

2.2 Comparison between Full QM and QM/MM Calculations

To identify where the main approximations enter and see how severe they are, let us consider the case when the entire system (QM+MM) is described uniformly at the DFT level. The total (electronic plus core-core interaction[d]) energy of such a system is given by the density functional[22]

$$E = T[\rho] + \int_\Omega V^{ex}(\mathbf{r})\rho(\mathbf{r}) + \frac{1}{2}\int\int \frac{\rho(\mathbf{r}_1)\rho(\mathbf{r}_2)}{r_{12}} \mathrm{d}\mathbf{r}_1 \mathrm{d}\mathbf{r}_2 + E_{xc}[\rho] + \frac{1}{2}\sum_I \sum_J \frac{Z_I Z_J}{R_{IJ}} \tag{5}$$

where T and E_{xc} are the kinetic and the exchange-correlation energy density functionals, respectively; V^{ex} is the external electrostatic potential created by the positively charged nuclei; r represents the electronic coordinates, while r_{12} refers to interelectronic and R_{IJ}

[d] Here, we are implicitly assuming that the Born-Oppenheimer approximation[21] is applied and that only the electrons are dealt with at quantum level while the nuclei are still described as point-like charges moving according to the classical Newtonian laws. This level of approximation turns out to be adequate for most biophysical applications.

to internuclear distances; Z_I and Z_J represent the nuclear (or core[e]) charge of atom I and J, respectively.

Now, we partition the system into two parts, A and B, with respective densities ρ_A and ρ_B. The total density ρ can be expressed (see also Ref. 23) as

$$\rho(\mathbf{r}) = \rho_A(\mathbf{r}) + \rho_B(\mathbf{r}). \tag{6}$$

Analogous to Eq. 4 the total energy is given by

$$E = E_A + E_B + E_{A-B} \tag{7}$$

with

$$\begin{aligned} E = T[\rho_A] + T[\rho_B] + T^{NL} + \int_\Omega V^{ex}(\mathbf{r})\rho_A(\mathbf{r}) + \int_\Omega V^{ex}(\mathbf{r})\rho_B(\mathbf{r}) + \\ + \frac{1}{2}\int\int \frac{\rho_A(\mathbf{r}_1)\rho_A(\mathbf{r}_2)}{r_{12}}\mathrm{d}\mathbf{r}_1\mathrm{d}\mathbf{r}_2 + \frac{1}{2}\int\int \frac{\rho_B(\mathbf{r}_1)\rho_B(\mathbf{r}_2)}{r_{12}}\mathrm{d}\mathbf{r}_1\mathrm{d}\mathbf{r}_2 + \\ + \frac{1}{2}\int\int \frac{\rho_A(\mathbf{r}_1)\rho_B(\mathbf{r}_2)}{r_{12}}\mathrm{d}\mathbf{r}_1\mathrm{d}\mathbf{r}_2 + E_{xc}[\rho_A] + E_{xc}[\rho_B] + E_{xc}^{NL} + \frac{1}{2}\sum_I\sum_J \frac{Z_I Z_J}{R_{IJ}}. \end{aligned} \tag{8}$$

The terms T^{NL} and E_{xc}^{NL} are:

$$T^{NL} = T[\rho_A + \rho_B] - T[\rho_A] - T[\rho_B] \tag{9}$$

$$E_{xc}^{NL} = E_{xc}[\rho_A + \rho_B] - E_{xc}[\rho_A] - E_{xc}[\rho_B]. \tag{10}$$

These terms account for the nonlinearity of the kinetic energy and the exchange-correlation density functionals, respectively.[f] They are only zero if ρ_A and ρ_B are spatially well separated.

For the particular case that we describe part A of the system with another approach than part B, it is useful to separate also the external potential V^{ex} into contributions from the nuclear cores of A and those of B:

$$V^{ex}(r) = V_A^{ex}(r) + V_B^{ex}(r). \tag{11}$$

E_A and E_B in Eq. 7 are given by the terms

$$\begin{aligned} E_Y = T[\rho_Y] + \int_\Omega V^{ex}(\mathbf{r})\rho_Y(\mathbf{r}) + \frac{1}{2}\int\int \frac{\rho_Y(\mathbf{r}_1)\rho_Y(\mathbf{r}_2)}{r_{12}}\mathrm{d}\mathbf{r}_1\mathrm{d}\mathbf{r}_2 + \\ + E_{xc}[\rho_Y] + \frac{1}{2}\sum_{I\in Y}\sum_{J\in Y}\frac{Z_I Z_J}{R_{IJ}} \end{aligned} \tag{12}$$

where $Y = A$ or B, respectively. Often the nuclear charges Z_I and Z_J are expanded into Gaussian shaped charge distributions with width R_c of the form

$$Z_I = \int_\Omega \rho_I^{nucl}(\mathbf{r} - \mathbf{R}_I)\mathrm{d}r = \int_\Omega \frac{Z_I}{R_c^3}\pi^{-2/3}\exp\left[-\frac{|\mathbf{r} - \mathbf{R}_I|^2}{R_c^2}\right] \tag{13}$$

[e] In some schemes only the outermost electrons of each atom (usually the *valence* electrons) are described by a wave function. All the other electrons are described implicitly by introducing pseudopotentials (see later) and the nuclear charge is replaced by the core charge, i.e. the difference between the atomic number and the number of explicit valence electron. Often this modified nucleus is refer to as "core".

[f] The term E_{xc}^{NL} in Eq. 10 arises also in the construction of *ab initio* atomic pseudopotentials when the system has to be partitioned into valence and core densities. In this case E_{xc}^{NL} is called *nonlinear core correction*[24].

and the three Coulomb terms can be summarized into one expression, which depends on the combined nuclear and electronic charge density $\rho^{el+nucl} = \rho^{el} + \rho^{nucl}$

$$\int_\Omega V_Y^{ex}(\mathbf{r})\rho_Y(\mathbf{r})\mathrm{d}r + \frac{1}{2}\int\int \frac{\rho_Y(\mathbf{r}_1)\rho_Y(\mathbf{r}_2)}{r_{12}}\mathrm{d}\mathbf{r}_1\mathrm{d}\mathbf{r}_2 + \frac{1}{2}\sum_{I\in Y}\sum_{J\in Y}\frac{Z_I Z_J}{R_{IJ}} = \\ = \frac{1}{2}\int\int \frac{\rho_Y^{el+nucl}(\mathbf{r}_1)\rho_Y^{el+nucl}(\mathbf{r}_2)}{r_{12}}\mathrm{d}\mathbf{r}_1\mathrm{d}\mathbf{r}_2. \quad (14)$$

The interface term E_{A-B} describes the interaction between A and B and therefore contains all the remaining terms

$$E_{A-B} = T^{NL} + \int_\Omega V_B^{ex}(\mathbf{r})\rho_A(\mathbf{r}) + \int_\Omega V_A^{ex}(\mathbf{r})\rho_B(\mathbf{r}) + \\ + \frac{1}{2}\int\int \frac{\rho_A(\mathbf{r}_1)\rho_B(\mathbf{r}_2)}{r_{12}}\mathrm{d}\mathbf{r}_1\mathrm{d}\mathbf{r}_2 + E_{xc}^{NL} + \frac{1}{2}\sum_{I\in A}\sum_{J\in B}\frac{Z_I Z_J}{R_{IJ}}. \quad (15)$$

For the special case where part A is treated with a QM and part B with an MM method, the first two energy terms in Eq. 4 correspond to

$$E_{QM} = T[\rho_{QM}] + \frac{1}{2}\int\int \frac{\rho_{QM}^{el+nucl}(\mathbf{r}_1)\rho_{QM}^{el+nucl}(\mathbf{r}_2)}{r_{12}}\mathrm{d}\mathbf{r}_1\mathrm{d}\mathbf{r}_2 + E_{xc}[\rho_{QM}] \quad (16)$$

$$E_{MM} = T[\rho_{MM}] + \frac{1}{2}\int\int \frac{\rho_{MM}^{el+nucl}(\mathbf{r}_1)\rho_{MM}^{el+nucl}(\mathbf{r}_2)}{r_{12}}\mathrm{d}\mathbf{r}_1\mathrm{d}\mathbf{r}_2 + E_{xc}[\rho_{MM}]. \quad (17)$$

E_{MM} is delegated to the classical force field. Clearly, none of the current force fields for biomolecular simulations can provide an exact match of the terms in Eq. 17. However, we will try to point out which typical analytical expressions are currently in use to mimic the physical effects described by E_{MM}.

As electrons are not considered explicitly, force fields are parameterized to single (or to the average of several) configurations with fixed electron density distributions. Therefore, the kinetic energy term in Eq. 17 can be considered as an additive constant that is not taken explicitly into account. The effect of the exchange-correlation energy functional is often replaced by a pair-additive van der Waals term:

$$E_{xc} \approx E^{vdW} = \sum_{I'}\sum_{J'} 4\varepsilon_{I'J'}\left(\left(\frac{\sigma_{I'J'}}{R_{I'J'}}\right)^{12} - \left(\frac{\sigma_{I'J'}}{R_{I'J'}}\right)^{6}\right). \quad (18)$$

The electrostatic potential due to the combined electronic and ionic charge distribution is approximated via effective point charges, usually located at atomic positions:

$$\frac{1}{2}\int\int \frac{\rho_{MM}^{el+nucl}(\mathbf{r}_1)\rho_{MM}^{el+nucl}(\mathbf{r}_2)}{r_{12}}\mathrm{d}\mathbf{r}_1\mathrm{d}\mathbf{r}_2 \approx \frac{1}{2}\sum_{I'}\sum_{J'}\frac{q_{I'}q_{J'}}{R_{I'J'}}. \quad (19)$$

The set of effective (often empirical) point charges commonly used in biomolecular force fields cannot be expected to faithfully reproduce the left hand side of Eq. 19, i.e. to be fully consistent with the electronic structure method used for the QM part. However, due to the extremely cumbersome work involved in the development of a general and transferable force field for complex biological systems, people usually prefer to employ existing

parameterizations instead of constructing a fully *ab initio* derived force field. In addition, it turns out that although the magnitude of effective point charges used in different force fields can vary largely, the average electrostatic potentials seem to be in surprisingly good agreement with each other as well as with DFT descriptions[25].

In spite of this somewhat reassuring caveat, the fact remains that real electronic charge distributions are far from mere assemblies of point charges. The point charge approximation breaks completely down in the description of covalent chemical bonds that are characterized by highly inhomogeneous and highly directional distributions of the electron density. Clearly, simple electrostatic/van der Waals descriptions such as those in Eqs. 18 and 19 cannot reproduce the intricacies of chemical bonding. In most force fields, the interaction between nearest, second nearest and third nearest neighbor atoms linked by chemical bonding are therefore mimicked by mechanical bond, angle and torsional angle terms of the following typical form:

$$E_{MM}^{bonded} = \sum_{b} \frac{1}{2} k_b (R_{I'J'} - b_0)^2 + \sum_{\theta} \frac{1}{2} k_\theta (\theta_{I'J'K'} - \theta_0)^2 + \\ + \sum_{\phi} \sum_{n} k_n [1 + \cos(n\phi_{I'J'K'L'}) - \phi_0] \tag{20}$$

where the first term runs over all bonds b with harmonic force constant k_b and equilibrium bond length b_0, the second term runs over all bonding angles θ with harmonic force constant k_θ and equilibrium bonding angle θ_0, whereas the last term is a sum over all dihedral angle interactions ϕ with multiplicity n and corresponding force constants k_n and phases ϕ_0. For the atoms connected via bonded terms, the nonbonded (electrostatic and van der Waals) interactions are either omitted or scaled down (so called exclusion rules).

Using Eqs. 18–20 the interaction energy in Eq. 15 becomes

$$E_{QM/MM} = \sum_{I'} \int_{\Omega} \frac{q_{I'}}{\mathbf{R}_{I'} - \mathbf{r}} \rho_{QM}^{el+nucl}(\mathbf{r}) \mathrm{d}\mathbf{r} + \\ + \sum_{I'} \sum_{I} 4\varepsilon_{I'I} \left(\left(\frac{\sigma_{I'I}}{R_{I'I}} \right)^{12} - \left(\frac{\sigma_{I'I}}{R_{I'I}} \right)^{6} \right) + \\ + \sum_{b} \frac{1}{2} k_b (R_{I'I} - b_0)^2 + \sum_{\theta} \frac{1}{2} k_\theta (\theta_{I''J''K''} - \theta_0)^2 + \\ + \sum_{\phi} \sum_{n} k_n [1 + \cos(n\phi_{I''J''K''L''}) - \phi_0] \tag{21}$$

where I runs over QM and I' over MM atoms and at least one atom of the triple $(I''J''K'')$ and quadruples $(I''J''K''L'')$ of bonded atoms is a QM atom. In this formulation, the effective classical point charges act as an external field to the QM calculation, i.e. the electron density of the QM part is polarized by the classical environment (in contrast to the subtractive approach described earlier).

Both the van der Waals term and the bonded terms are acting on atomic positions only, i.e. are not part of the total electronic potential and thus are not directly felt by the electrons. If we want to achieve a closer model of a full QM description, the deviations caused by the actual MM representation have to be compensated by a correction term ΔV in the total

potential that the electrons of the QM part experience

$$V_{tot} = V_{QM} + V_{QM/MM} + \Delta V \tag{22}$$

$$\Delta V = \Delta V^{NL} + \Delta V^{el} \tag{23}$$

where ΔV^{el} accounts for the error in the electrostatic terms (deviation of the classical electrostatic potential from the QM reference and reduction of the electronic density distribution to a point charge representation) whereas the nonlinear correction term ΔV^{NL} results from the nonlinearity corrections in Eqs. 9 and 10. Thus this term is a mere artifact of the density partitioning and is not present in a system treated at the uniform level. To keep this term minimal, the somewhat trivial but important condition has to be fulfilled: the QM part has to be chosen in such a way that the electronic wave functions are localized to this region. If this condition cannot be fulfilled, the correction term ΔV^{NL} gains in importance (see paragraph about pitfalls and limitations).

How can we assess the importance of the correction term ΔV in practice? Ideally, one would like that the electron density in the QM region, ρ_{QM}, matches as closely as possible the electron density in the same region produced by a full QM representation of the system (ρ_{true}). According to the Hohenberg-Kohn theorem[22], if the two densities are identical, all the properties we calculate for the QM region are identical to those of the real system. In other words, if we determine the correction potential ΔV in such a way that the total electronic potential in a QM/MM simulation V_{tot} minimizes the density difference

$$\int_{\Omega'} (\rho_{true}(\mathbf{r}) - \rho_{QM}(\mathbf{r}))^2 \tag{24}$$

where Ω' is a suitably chosen volume of the QM region, our QM/MM simulation approaches the full QM reference results in an optimal way (see also paragraph about pitfalls and limitations).

2.3 CPMD/MM Method: Basics

The Car-Parrinello method[6] can be extended into a QM/MM scheme using a mixed Lagrangian of the form[16]:

$$\mathcal{L} = \frac{1}{2}\mu \sum_i \int \mathrm{d}\mathbf{r}\, \dot{\psi}_i^*(\mathbf{r})\dot{\psi}_i(\mathbf{r}) + \frac{1}{2}\sum_I M_I \dot{\mathbf{R}}_I^2 - E_{MM} - E_{QM/MM} - E_{QM} + \\ + \sum_{i,j} \Lambda_{i,j} \left(\int \mathrm{d}\mathbf{r}\, \psi_i^*(\mathbf{r})\psi_j(\mathbf{r}) - \delta_{i,j} \right) \tag{25}$$

where μ is the fictitious mass associated with the electronic degrees of freedom, ψ_i are the Kohn-Sham one particle orbitals, M_I is the mass of atom I and $\Lambda_{i,j}$ are Lagrange multipliers that enforce orthonormality of the Kohn-Sham orbitals. The energy of the QM system E_{QM} is given by the Kohn-Sham energy density functional[26]

$$E_{QM} = E_{KS}[\psi_i, \mathbf{R}_I] = -\frac{1}{2}\int \mathrm{d}\mathbf{r}\, \psi_i^*(\mathbf{r})\nabla^2\psi_i(\mathbf{r}) + \int \mathrm{d}\mathbf{r}\, V^{ex}(\mathbf{r})\rho_{QM}(\mathbf{r}) + \\ + \frac{1}{2}\int \mathrm{d}\mathbf{r}\, \mathrm{d}\mathbf{r}' \rho_{QM}(\mathbf{r}) \frac{1}{|\mathbf{r}-\mathbf{r}'|} \rho_{QM}(\mathbf{r}') + E_{xc}[\rho_{QM}(\mathbf{r})] \tag{26}$$

where for the spin unpolarized case, the electron density $\rho_{QM}(\mathbf{r})$ is given by the sum of the densities of the doubly occupied one-particle states:

$$\rho_{QM} = 2\sum_i \psi_i^*(\mathbf{r})\psi_i(\mathbf{r}). \tag{27}$$

The purely classical part E_{MM} is described by a standard biomolecular force field:

$$E_{MM} = E_{MM}^{bonded} + E_{MM}^{non-bonded} \tag{28}$$

as given by Eqs. 18–20. The interaction between the QM and MM parts, $E_{QM/MM}$, is included in the form of Eq. 21 with the only exception of the harmonic bond interactions between QM and MM atoms which are omitted from the classical description and treated at the QM level.

Standard implementations of Car-Parrinello MD simulations use plane wave basis sets. In this case, due to the high intrinsic flexibility of a plane wave basis set (in contrast to e.g. the minimal basis sets used in semi-empirical QM/MM calculations), special care has to be taken that the CPMD/MM interface is described in an accurate and consistent way. In our case, the quantum/classical correction term ΔV consists of specifically designed monovalent pseudopotentials to represent bonds between QM and MM parts of the system[27] and of modified screened Coulomb potentials for the interaction of the quantum electron density with close by classical point charges[28].

In the context of a plane wave based Car-Parrinello scheme, a direct evaluation of the first term of Eq. 21 is prohibitive as it involves of the order of $N_r \times N_{MM}$ operations, where N_r is the number of real space grid points (typically $\sim 100^3$) and N_{MM} is the number of classical atoms (usually of the order of 10,000 or more in systems of biochemical relevance). Therefore, the interaction between the QM system and the more distant MM atoms is included via a Hamiltonian term explicitly coupling the multipole moments of the quantum charge distribution with the classical point charges. This two level electrostatic coupling scheme can also be refined to an intermediate third layer that makes efficient use of variational D-RESP charges[28,29]. Highly efficient schemes based on a dual grid approach[30] or a multigrid approach with Gaussian expansion[31] have also been proposed in this context.

The hybrid CPMD/MM implementation of Ref. 17 establishes an interface between the Car-Parrinello code CPMD and the classical force fields GROMOS96 and AMBER in combination with a particle-particle-particle mesh (P3M) treatment of the long-range electrostatic interactions[32]. With this implementation, efficient and consistent simulations of complex systems (of the order of 10^5 atoms) can be performed. In these calculations, the steric and electrostatic effects of the surroundings are taken into account explicitly.

2.4 CPMD/MM Method: Limitations

The most stringent current limitation is the short time scale accessible via CPMD/MM simulations of the order of tens to hundreds of picoseconds which severely restricts the accuracy of time-averaged properties, such as binding free energies. Possible remedies for this problem are: (i) resorting to semi-empirical methods that allow sampling for hundreds of picoseconds[33], (ii) employing multiple time step sampling for the QM and MM parts[34], (iii) using enhanced sampling approaches such as metadynamics[35], introducing

either a classical[36,69] or electronic[27,37] bias potential, (iv) or exploiting a linear response approximation with respect to a reference potential[11].

The issue of the accuracy of DFT is also very important. A particular problem is the adequate description of London dispersion forces. Several methods have been developed to cure this problem, e.g. the addition of an effective atom-centered non-local term to the exchange-correlation potential may cure this significant drawback without additional computational cost[38]. The dispersion correction most commonly used with the CPMD/MM scheme are dispersion corrected atom-centered potentials (DCACPs) that are directly included in the electronic Hamiltonian[39,40]. Another well-known issue of most DFT calculations is the underestimation of energy barriers associated with proton transfer events and other chemical reactions. The implementation of mixed localized basis sets other than plane-waves might enable the use of hybrid exchange-correlation functionals, such as B3LYP[41] or meta-hybrid functionals such as e.g. the MXX suite[42–45], which might help improve the accuracy of the results.

2.5 Pitfalls

The main intrinsic approximations of a QM/MM approach lie in the reduction of the real electron density distribution of the MM part to a mere point charge representation and the neglect of the kinetic energy and exchange-correlation corrections Eqs. 9 and 10 on the electronic level. All three of these terms are particularly severe in the neighborhood of a covalent chemical bond, where the electron density distribution is far from isotropic and the densities of the QM and MM part are strongly overlapping. In force field descriptions, these deficiencies in the description of chemical bonding are remedied by including the special bonding terms given in Eq. 20. However, these terms are a function of atomic coordinates only and do not influence the electronic potential in a direct way. One of the most current problems in QM/MM simulations thus occurs when the border between QM and MM parts has to run across a chemical bond: this is called the *link atom problem*. For QM/MM simulations of biological systems this is essentially always the case. In fact, a typical QM/MM partitioning for such systems includes only a portion of a biological macromolecule. The latter must then be cut into a QM and a MM region. As electrons cease to exist when passing from the QM to the MM region, the QM system contains unsaturated valencies and has to be made chemically inert.

This can be done in the spirit of Eq. 22 by introducing an explicit correction term in the total electronic potential felt by the QM electrons. For the case of a QM/MM bond cut, the simplest way is to use a monovalent pseudopotential situated at the position of the first MM atom to represent the correction potential in Eq. 23. This pseudopotential is usually constructed in such a way that the electrons of the QM region are scattered correctly by the classical environment. It is a common choice in CPMD/MM simulations to employ analytic, nonlocal pseudopotentials of e.g. the Gödecker type[46]

$$V^{eff}(\mathbf{r},\mathbf{r}') = V^{loc}(\mathbf{r})\delta(\mathbf{r}-\mathbf{r}') + \sum_l V_l^{nloc}(\mathbf{r},\mathbf{r}')$$

$$V^{loc}(\mathbf{r}) = -\frac{Z_I}{r}\mathrm{erf}\left[\frac{\mathbf{r}}{r_{loc}\sqrt{2}}\right] + \exp\left[-\frac{r^2}{2r_{loc}^2}\right]\Theta$$

$$\Theta = \left(C_1 + C_2 \left(\frac{r}{r_{loc}} \right)^2 + C_3 \left(\frac{r}{r_{loc}} \right)^4 + C_4 \left(\frac{r}{r_{loc}} \right)^6 \right)$$

$$V_l^{nloc}(\mathbf{r}, \mathbf{r}') = \sum_{m=-l}^{+l} Y_{l,m}(\mathbf{r}) \sum_{i,j=1}^{3} p_i^l(\mathbf{r}) h_{i,j}^l p_j^l(\mathbf{r}') Y_{l,m}^*(\mathbf{r}')$$

$$p_{lh} \propto r^{l+2(h-1)} \exp\left[-r^2/2r_l^2\right]$$

to represent the MM atoms involved in QM/MM bond cuts. The adjustable parameters r_{loc}, r_l, $h_{i,j}^l$ and C_1 to C_4 are determined in analogy to Eq. 24 by minimizing the density penalty

$$F[\rho_{QM}(\mathbf{r}), \{\sigma_i\}] = \int_{\Omega'} \mathrm{d}(\mathbf{r}) |\rho_{ref}(\mathbf{r}) - \rho_{QM}(\mathbf{r}, \{\sigma_i\})|^2 \tag{29}$$

where the σ_i's are the set of adjustable parameters and ρ_{ref} is a reference density that approximates ρ_{true} in Eq. 24. ρ_{ref} is usually determined from a QM/MM calculation of the system with extended QM part[47].

There are many other *ad hoc* procedures in use to cure the link atom problem. Commonly used strategies are to add capping atoms (hydrogen or fluorine) or to represent the last QM atom with frozen frontier orbitals[48]. However, hydrogen capping introduces new atoms into the QM system that are not present in the real system. As a consequence, the QM portion is chemically not identical with the real system (e.g. the true system may contain C-C bonds at the boundary that are now described with C-H bonds that clearly have different electronic and chemical properties). Furthermore, additional degrees of freedom have been introduced and interactions of these nonexistent ghost atoms with the classical environment have to be carefully removed. Some of these drawbacks are remedied by the use of frozen frontier orbitals for the boundary atoms. In this way, no additional physical interactions and degrees of freedom are introduced and the QM part retains its original composition. However, frozen orbitals have to be determined via calculations on small model systems and, as the name says, they remain frozen when transferred into the real environment. Specially parameterized pseudopotentials such as the ones described above, on the other hand, have the additional flexibility to adjust to changes in the environment.

Another possible artifact in QM/MM simulations, in particular in combination with extended and highly flexible basis sets (such as e.g. plane waves) is the *electron spill-out* problem. As shown in Eq. 21, the exchange interactions between QM and MM part are taken into account on the level of atomic pair interactions only. Once again, these terms do not directly affect the electrons of the QM part. For a proper description of the electronic structure of the QM region an electronic correction term ΔV_{xc}^{NL} has to be included. As we have seen, this term is especially important for regions with overlapping or nearly overlapping densities between QM and MM parts, which is particularly the case for the nearby atoms surroundings the QM region. Due to the fact that the MM part contains no explicit electrons, the electrons of the QM part are no longer repelled by the closed-shell cores of the MM region. As a result of this missing Pauli repulsion, the electrons of the QM part can artificially localize on nearby positively charged classical point charges. This phenomenon is called electron spill-out. This effect can be avoided by using Gaussian smeared (screened) classical charges[g] or by replacing the classical point charge potential

[g] Attention: drastic artifacts are possible by choosing too large widths for the Gaussian broadening.

by suitably constructed ionic pseudopotentials with screened electrostatic interactions[16]. The latter solution is the one implemented in CPMD and the one used in the applications mentioned in this article. In particular

$$E^{el}_{QM/MM} = \sum_{I \in MM} q_I \int \mathrm{d}\mathbf{r}\, \rho(\mathbf{r}) \nu_I(|\mathbf{r} - \mathbf{R}_I|) \tag{30}$$

where q_I is the classical point charge located at R_I and

$$\nu_I(|\mathbf{r} - \mathbf{R}_I|) = \frac{r_c^4 - r^4}{r_c^5 - r^5} \tag{31}$$

(with r_c chosen as the covalent radius of atom I) is a Coulomb interaction potential modified at short-range in such a way as to avoid spill-out of the electron density to nearby positively charged classical point charges.

Other potential sources of problems are possible incompatibilities between the QM and MM descriptions, such as imbalances in the electrostatic interactions that can lead to artificial preferences of e.g. substrate-QM, respectively substrate-MM interactions. Another problem is the consistent application of the classical exclusion rules for nonbonded interactions. In most force field definitions, nonbonded interactions (such as van der Waals and electrostatics) are not taken into account for nearby bonded neighbors. Such a selective neglect of particular pair interactions is not easily transferable to a many-body QM description. A consistent approach is however possible via mapping of the many-body electronic Hamiltonian to a pair additive point charge representation[29].

3 Applications to Biological Systems

3.1 Bioinorganic Chemistry of Parkinson's Disease: Copper Binding to α-Synuclein

Parkinson's disease (PD) is the second most common neurodegenerative disease in adults, affecting about 5 million people worldwide[49]. It is characterized by a loss of dopaminergic neurons and the presence of proteinaceous fibrillar aggregates (Lewy bodies) in the surviving motor neurons[50]. The most abundant components of the Lewy bodies are amyloid fibrils consisting of the protein α-synuclein (AS)[51–54]. It has been demonstrated that metal ions such as copper or iron bind to AS and accelerate its fibrillation *in vitro*[55,56]. We performed CPMD/MM simulations on AS/Cu(II) adducts - along with experimental spectroscopic investigations by our experimental collaborators - to elucidate the structural determinants of the adducts[57]. We performed 4ps-long CPMD/MM simulations on 18 representative conformers identified previously[58] using DFT for the QM part and the AMBER force field[12] for the MM part. The Cu(II) ion binds to the N-terminal Met-1 and Asp-2 backbone nitrogens, the Asp-2 carboxylate side-chain, and a water molecule (Fig. 1). The QM part shown in Fig. 1 contained 27 atoms, including Cu(II), the N-terminal Met-1 backbone unit and its side-chain up to the C_β atom, the Asp-2, and the water molecule coordinating the copper ion. Valences of terminal carbon atoms were capped by adding hydrogen atoms to the QM region.

The calculated average structural parameters point to a distorted tetragonal preferential coordination geometry for Cu(II). Specific Cu(II) binding in the N-terminus region, which

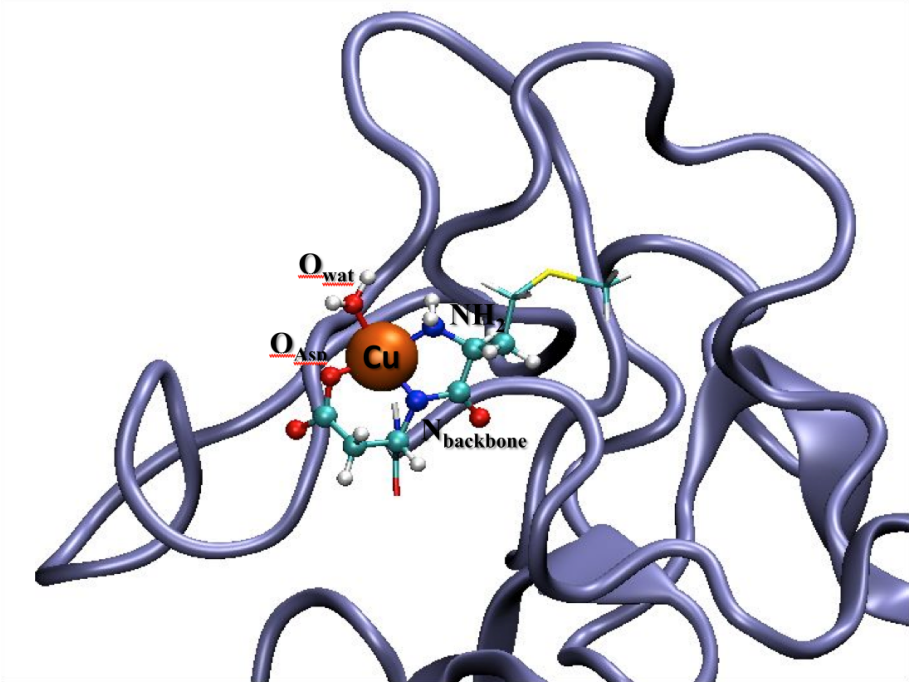


Figure 1. Representative structure of one of the AS/CU(II) conformers, as obtained by CP/MM simulations. The Cu(II) coordination geometry is distorted tetragonal.

is the highest affinity binding site, was shown to be crucial for the metal-mediated AS fibrillation process. The calculated absorption spectrum in accord with experimental data shows a characteristic band around 620 nm.

Thus, from the combined theoretical and experimental study new insight into the structural binding specificity and aggregation enhancement mediated by Cu(II) was obtained.

3.2 Optical Properties of Indole in Water Solution

Optical properties of chromophores play a central role for a precise and non-destructive interrogation of a variety of biochemical events. These include fundamental and important processes such as transient interactions between biomolecules (proteins or nucleic acids), protein dynamics, fibrillation and plaque formation associated with the development of neurodegenerative diseases, or high-throughput screening in drug discovery. Understanding how optical properties of chromophores are tuned by the biomolecular and/or solvent environment is therefore of fundamental importance, yet this information is so far mostly lacking.

Many proteins contain naturally fluorescent amino acid residues such as phenylalanine, tyrosine or tryptophan. In particular, the fluorescence of tryptophan residues has been shown to be particularly well-suited to monitor protein dynamics[59]. Tryptophan emission usually dominates protein emission as it absorbs at the longest wavelength and exhibits the

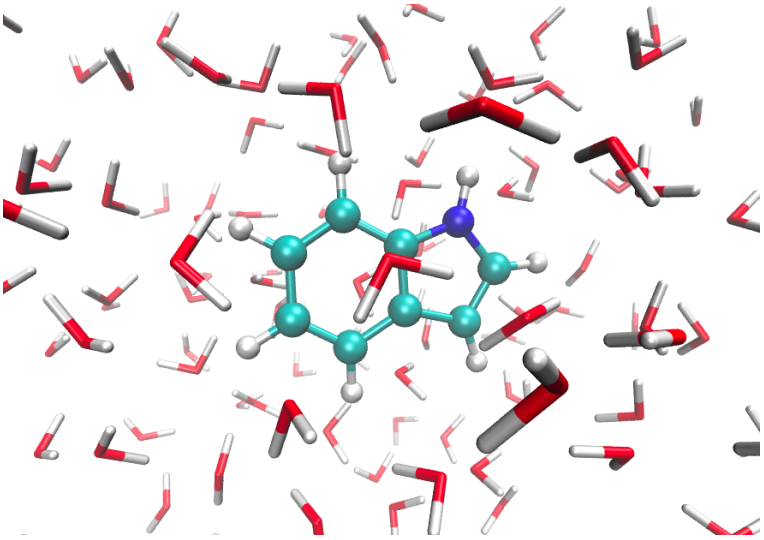


Figure 2. Indole in water solution. The solute is treated at the QM level, whilst the solvent is described by the AMBER force field[12].

largest extinction coefficient. Indole, the chromophore of tryptophan, absorbs at about 280 nm and emits near 340 nm. Changes of the emission spectrum occur in response to structural or polarity environmental changes, e.g. the emission may be blueshifted if tryptophan is buried in the proteins interior, whereas a redshift may occur if the chromophore is at an exposed surface accessible to water solvent molecules, e.g. upon protein unfolding.

Here, we investigate the physical origin of the redshift observed in the optical absorption spectrum of indole in going from the gas phase to aqueous solution[60,61]. We use CPMD/MM simulations interfaced with time-dependent DFT (TDDFT)[63] methods[62,64,65] as well as many-body (GW-BSE)[66] approaches[60,61].

Our calculations demonstrate that the experimentally observed and computationally confirmed solvatochromic redshift of the optical absorption spectrum in water is a consequence of the combination of two effects: the geometrical distortion of the indole molecule in the solvent as well as the electrostatic interaction with the water molecules' electric dipoles. Both effects, and their sum, depend on the particular configuration of the system; this emphasizes the need of including both altogether and of averaging over several snapshots.

These studies open the way to further applications on other biorelevant molecules, such as fluorescent probes in their target proteins, for which the evaluation of the optical shift enables the understanding of the nature of their environment.

4 Concluding Remarks

Nowadays, CPMD/MM simulations are a rather established tool for the investigation of adiabatic ground state reactions in biological environments, such as, for instance, the solvation of biologically relevant molecules[67]. One of the main benefits of the CPMD/MM

approach is its ability to simulate complex reactions from first principles. This approach, which includes temperature effects, can benefit from the use of statistical mechanics methods[68–71] to investigate rare events, such as enzymatic reaction mechanisms. Recent reviews report examples of enzymatic reactions investigated with this method[72–74]. CPMD/MM applications are also of importance to study drug action. Indeed, the interaction between a ligand and its target might at times depend on the electronic structure in such a subtle way that is difficult to capture with force field based MD. A recent review reports CPMD/MM applications that address this issue[75]. The method has also recently provided valuable insights on DNA damage[76–78].

The CPMD/MM approach has also been extended to the description of electronically excited states[79–81,62,60], and nonadiabatic dynamics[82] which enables the investigation of photochemical reactions, e.g. in photoactive proteins and photochemically linked substrate target interactions. In excited state QM/MM schemes, the excited states are either described via multiconfigurational wave function-based quantum chemical methods, or by many-body perturbation theory or through excited state extensions of density functional theory[83–85], as, for instance, the study described in Sec. 3.2. Whereas the former two types of approaches are still limited to fairly small systems and relatively small basis sets, which can compromise accuracy, the latter one is also extendable to fairly large systems.

Further CPMD/MM developments also allow access to many other molecular properties beyond optical spectra such as, e.g. NMR chemical shifts[25], that are useful to make direct contact with experimental data and facilitate verification of the simulation results for complex systems.

The interested reader is referred to the book of Marx and Hutter[8] for a comprehensive survey of recent developments in the field.

Acknowledgments

PC and UR wish to thank all their coauthors of the papers referred to in this review, and Michael L. Klein and Michele Parrinello for their continued support over the last two decades.

References

1. J. P. Perdew, A. Ruzsinszky, J. Tao, V. N. Staroverov, G. E. Scuseria and G. I. Csonka, *Prescription for the design and selection of density functional approximations: More constraint satisfaction with fewer fits*, J. Chem. Phys. **123**, 062201, 2005.
2. Y. Zhao, N. E. Schultz and D. G. Truhlar, *Design of Density Functionals by Combining the Method of Constraint Satisfaction with Parametrization for Thermochemistry, Thermochemical Kinetics, and Noncovalent Interactions*, J. Chem. Theory Comput. **2**, 364, 2006.
3. G. E. Scuseria and V. N. Staroverov in C. E. Dykstra, G. Frenking, K. S. Kim and G. E. Scuseria (Eds.), *Chapter 24 - Progress in the development of exchange-correlation functionals*, Theo. Appl. Comput. Chem. , 669, 2005.
4. Y. Zhao and D. G. Truhlar, *Density Functionals with Broad Applicability in Chemistry*, Accounts Chem. Res. **41**, 157, 2008.

5. G. I. Csonka, J. P. Perdew and A. Ruzsinszky, *Global Hybrid Functionals: A Look at the Engine under the Hood*, J. Chem. Theory Comput. **6**, 3688, 2010.
6. R. Car and M. Parrinello, *Unified approach for molecular dynamics and density-functional theory*, Phys. Rev. Lett. **55**, 2471, 1985.
7. P. Carloni, *Ab initio simulation of a biological ion channel*, PRACE project, 2011. `http://www.prace-ri.eu/PRACE-1st-Regular-Call`.
8. D. Marx and J. Hutter, *Ab initio Molecular Dynamics: Basic Theory and Advanced Methods*, Cambridge University Press, 2009.
9. *Special issue: Parrinello Festschrift*, Chem. Phys. Chem. **6**, 1669, 2005.
10. A. Hassanali, M. K. Prakash, H. Eshet and M. Parrinello, *On the recombination of hydronium and hydroxide ions in water* Proc. Natl. Acad. Sci. USA **108**, 20410, 2011.
11. A. Warshel and M. Levitt, *Theoretical studies of enzymic reactions - dielectric, electrostatic and steric stabilization of carbonium - ion in reaction of lysozyme*, J. Mol. Biol. **103**, 227, 1976.
12. W. D. Cornell, P. Cieplak, C. I. Bayly, K. M. Gould, K. M. Merz, D. M. Ferguson, G. L. Seibel, U. C. Singh, P. K. Weiner, and P. A. Kollmann, *A 2nd generation force-field for the simulation of proteins, nucleic acids, and organic molecules*, J. Am. Chem. Soc. **117**, 5179, 1995.
13. W. R. P. Scott, P. H. Hünenberger, I. G. Tironi, A. E. Mark, S. R. Billeter, J. Fennen, A. E. Torda, T. Huber, P. Krüger, and W. F. van Gunsteren, *The GROMOS biomolecular simulation program package*, J. Phys. Chem. A **103**, 3596, 1999.
14. A. D. MacKerell, Jr. J. Wiorkiewicz-Kuczera, and M. Karplus, *An all-atom empirical energy function for the simulation of nucleic acids*, J. Am. Chem. Soc. **117**, 11946, 2000.
15. U. Rothlisberger, and P. Carloni, *Drug-target binding investigated by quantum mechanical/molecular mechanical (QM/MM) methods*, from "Computer Simulations in Condensed Matter Systems: From Materials to Chemical Biology, Vol 2" Lecture Notes in Physics **704**, 449, 2006.
16. A. Laio, J. VandeVondele, and U. Rothlisberger, *A Hamiltonian electrostatic coupling scheme for hybrid Car-Parrinello molecular dynamics simulations*, J. Chem. Phys. **116**, 6941, 2002.
17. CPMD, IBM Corp. (1990-2001) Copyright MPI fuer Festkörperforschung, Stuttgart (1997-2001). `http://www.cpmd.org`.
18. F. Maseras and K. Morokuma. *IMOMM - A new integrated ab initio plus molecular mechanics geometry optimization scheme of equilibrium structures and transition states*, J. Comput. Chem. **16**, 1170, 1995.
19. `http://www.gaussian.com`.
20. M. Svensson, S. Humbel, R. D. J. Froese, T. Matsubara, S. Sieber, and K. Morokuma *ONIOM: A multilayered integrated MO+MM method for geometry optimizations and single point energy predictions. A test for Diels-Alder reactions and Pt(P(t-Bu)3)2^{+} + H_2 oxidative addition*, J. Phys. Chem. **100**, 19357, 1996.
21. M. Born and J. R. Oppenheimer, Ann. Physik **84**, 457, 1927.
22. P. Hohenberg and W. Kohn *Inhomogeneous electron gas*, Phys. Rev. B **136**, 864, 1964.
23. T. A. Wesolowski, *Frozen density-functional approach for ab initio calculations of solvated molecules*, J. Phys. Chem. **97**, 8050, 1993.

24. S. G. Louie, S. Froyen, and M. L. Cohen, *Non-linear ionic pseudopotentials in spin-density-functional calculations*, Phys. Rev. B **26**, 1738, 1982.
25. D. Sebastiani and U. Rothlisberger, *Advances in Density Functional Based Modelling Techniques: Recent Extensions of the Car-Parrinello Approach* in P. Carloni, F. Alber "Medicinal Quantum Chemistry", Series: Methods and Principles in Medicinal Chemistry. Series Editors: R. Mannhold, H. Kubiny, G. Folkers, Wiley-VCH, Weinheim. 2003.
26. W. Kohn, and L. J. Sham, *Self-consistent equations including exchange and correlation*, Phys. Rev. **140**, 1133, 1965.
27. J. VandeVondele and U. Rothlisberger, *Accelerating rare reactive events by means of a finite electronic temperature*, J. Am. Chem. Soc. **124**, 8163, 2002
28. A. Laio, J. VandeVondele, and U. Rothlisberger, *D-RESP: Dynamically generated electrostatic potential derived charges from QM/MM simulations*, J. Phys. Chem. B **106**, 7300, 2002.
29. A. Laio, F. Gervasio, M. Sulpizi, and U. Rothlisberger, *A variational definition of electrostatic potential derived charges* J. Phys. Chem. **108**, 7983, 2004.
30. D. Yarne, M. E Tuckerman, and G. J. Martyna, *A dual length scale method for plane-wave-based simulations studies of chemical systems modeled using mixed* ab initio*/empirical force field descriptions*, J. Chem. Phys. **115**, 3531, 2001.
31. T. Laino, F. Mohamed, A. Laio, and M. Parrinello, *An efficient real space multigrid QM/MM electrostatic coupling*, J. Comp. Theory Chem. **1**, 1176, 2005.
32. P. Hünenberger, *Optimal charge-shaping functions for the particle-particle-particle-mesh (P3M) method for computing electrostatic interactions in molecular simulations*, J. Chem. Phys. **113**, 10464, 2000.
33. R. Rajamani, K. J. Naidoo, and J. L. Gao, *Implementation of an adaptive umbrella sampling method for the calculation of multidimensional potential of mean force of chemical reactions in solution*, J. Comp. Chem. **24**, 1775, 2003.
34. T. K. Woo, P. Margl, P. E. Bloechl, and T. Ziegler *Sampling phase space by a combined QM/MM ab initio Car-Parrinello molecular dynamics method with different (multiple) time steps in the quantum mechanical (QM) and the molecular mechanical (MM) domains* J. Phys. Chem. A **106**, 1173, 2002.
35. A. Laio and M. Parrinello, *Escaping free-energy minima*, Proc. Nat. Ac. Sc. **99**, 12562, 2003.
36. J. VandeVondele and U. Rothlisberger *Efficient multidimensional free energy calculations for ab initio molecular dynamics using classical bias potentials*, J. Chem. Phys. **113**, 4863, 2000.
37. L. Guidoni and U. J. Rothlisberger, *Scanning reactive pathways with orbital biased molecular dynamics*, J. Chem. Theo. Comp. **1**, 554, 2005.
38. S. Grimme, *Density functional theory with London dispersion corrections*, Wiley Interdiscip. Rev. Comput. Mol. Sci. **1**, 211, 2011.
39. O.A. von Lilienfeld, I. Tavernelli, D. Sebastiani, and U. Rothlisberger, *Optimization of Effective Atom Centered Potentials for London Forces in Density Functional Theory*, Phys. Rev. Lett. **93**, 15300, 2004.
40. I. C. Lin, M.D. Coutinho-Neto, C. Felsenheimer, O.A. von Lilienfeld, I. Tavernelli, and U. Rothlisberger, *A library of dispersion corrected atom-centered potentials for generalized gradient approximation functionals: elements H, C, N, O, He, Ar, and Kr*, Phys. Rev. B **75**, 205131, 2007.

41. http://www.cp2k.org.
42. Y. Zhao, and D. G. Truhlar, *Hybrid Meta Density Functional Theory Methods for Thermochemistry, Thermochemical Kinetics, and Noncovalent Interactions: The MPW1B95 and MPWB1K Models and Comparative Assessments for Hydrogen Bonding and van der Waals Interactions*, J. Phys. Chem. A **108**, 6908, 2004.
43. Y. Zhao, B. J. Lynch, and D. G. Truhlar, *Doubly Hybrid Meta DFT: New Multi-Coefficient Correlation and Density Functional Methods for Thermochemistry and Thermochemical Kinetics* J. Phys. Chem. A **108**, 4786, 2004.
44. Y. Zhao, and D. G. Truhlar, *The M06 Suite of Density Functionals for Main Group Thermochemistry, Thermochemical Kinetics, Noncovalent interactions, Excited States, and Transition Elements: Two New Functionals and Systematic Testing of Four M06 Functionals and Twelve Other Functionals*, Theor. Chem. Acc. **120**, 215, 2008.
45. R. Peverati and D. G. Truhlar, *Improving the Accuracy of Hybrid Meta-GGA Density Functionals by Range Separation*, J. Phys. Chem. Lett. **2**, 2810, 2011.
46. C. Hartwigsen, S. Goedecker, and J. Hutter, *Relativistic separable dualspace Gaussian pseudopotentials from H to Rn* Phys. Rev. B **58**, 3641, 1998.
47. O. A. von Lilienfeld, I. Tavernelli, U. Rothlisberger, and D. Sebastiani, *Variational optimization of effective atom-centered potentials for molecular systems*, J. Chem. Phys. **122**, 014113, 2005.
48. X. Assfeld and J. L. Rivail, *Quantum chemical computations on parts of large molecules: The* ab initio *local self consistent field method*, Chem. Phys. Lett. **263**, 100, 1996.
49. L. A. Shehadeh, K. Yu, L. Wang, A. Guevara, C. Singer, J. Vance, and S. Papapetropoulos, *SRRM2, a Potential Blood Biomarker Revealing High Alternative Splicing in Parkinson's Disease*, PLoS ONE **5**, e9104, 2010.
50. J. M. Fearnley and A. J. Lees, *Ageing and Parkinson's disease: Substantia nigra regional selectivity*, Brain **114**, 2283, 1991.
51. M. G. Spillantini, M. L. Schmidt, V. M.-Y. Lee, J. Q. Trojanowski, R. Jakes, M. Goedert, α*-Synuclein in Lewy bodies*, Nature **388**, 839, 1997.
52. M. Goeder, α*-synuclein and neurodegenerative diseases*, Nature Rev. Neurosci. **2**, 492, 2001.
53. H. Snyder and B. Wolozin, *Pathological proteins in Parkinsons disease: Focus on the proteasome*, J. Mol. Neurosci. **24**, 425, 2004.
54. M. R. Cookson, α*-Synuclein and neuronal cell death*, Mol. Neurodegeneration **4**, 9, 2009.
55. V. N. Uversky, J. Li and A. L. Fink, *Metal-triggered Structural Transformations, Aggregation, and Fibrillation of Human* α*-Synuclein*, J. Biol. Chem. **276**, 44284, 2001.
56. R. M. Rasia, C. W. Bertoncini, D. Marsh, W. Hoyer, D. Cherny, M. Zweckstetter, C. Griesinger, T. M. Jovin, and C. O. Fernández, *Structural characterization of copper(II) binding to* α*-synuclein: Insights into the bioinorganic chemistry of Parkinson's disease*, Proc. Natl. Acad. Sci. USA **102**, 4294, 2005.
57. A. Binolfi, E. E. Rodriguez, D. Valensin, N. D'Amelio, E. Ippoliti, G. Obal, R. Duran, A. Magistrato, O. Pritsch, M. Zweckstetter, G. Valensin, P. Carloni, L. Quintanar, C. Griesinger, C. O. Fernández, *Bioinorganic Chemistry of Parkinson's Disease: Structural Determinants for the Copper-Mediated Amyloid Formation of Alpha-Synuclein*, Inorg. Chem. **49**, 10668, 2010.

58. F. E. Herrera, A. Chesi, K. E. Paleologou, A. Schmid, A. Munoz, M. Vendruscolo, S. Gustincich, H. A. Lashuel and P. Carloni, *Inhibition of alpha-synuclein fibrillization by dopamine is mediated by interactions with five C-terminal residues and with E83 in the NAC region* PloS One **3**, e3394, 2008.
59. M. Hof, R. Hutterer, V. Fidler, Eds., *Fluorescence Spectroscopy in Biology*, Springer Verlag, Springer Series on Fluorescence, Vol. 3, 2005.
60. A. M. Conte, E. Ippoliti, R. Del Sole, P. Carloni, O. Pulci, *Many-Body Perturbation Theory Extended to the Quantum Mechanics/Molecular Mechanics Approach: Application to Indole in Water Solution* J. Chem. Theory Comp. **5**, 1822, 2009.
61. A. Mosca-Conte, E. Ippoliti, R. Del Sole, P. Carloni, O. Pulci, *Many-body meets QM/MM: Application to indole in water solution*, Phys. Status Solidi B **247**, 1920, 2010.
62. I. Tavernelli, U. F. Röhrig, U. Rothlisberger , *Molecular dynamics in electronically excited states using time-dependent density functional theory*, Mol. Phys. **103**, 963, 2005.
63. M. Marques, M. A. L.; Gross, E. K. U. Time-Dependent Density Functional Theory. In A Primer in Density Functional Theory; Fiolhais, C.; Nogueira, F.; Marques, M. A. L., Eds.; Springer-Verlag: Berlin, 2003; Vol. 620, pp 144 184
64. M. Sulpizi, P. Carloni, J. Hutter, U. Rothlisberger, *A hybrid TDDFT/MM investigation of the optical properties of aminocoumarins in water and acetonitrile solution*, Phys. Chem. Chem. Phys. **5**, 4798, 2003.
65. M. Sulpizi, U. Röhrig, J. Hutter, U. Rothlisberger, *Optical properties of molecules in solution via hybrid TDDFT/MM simulations*, Int. J. Quantum Chem. **101**, 671, 2005.
66. G. Onida, L. Reining, A. Rubio, *Electronic excitations: density-functional versus many-body Green's-function approaches* Rev. Mod. Phys. **74**, 601, 2002.
67. J. Sun, D. Bousquet, H. Forbert and D. Marx, *Glycine in aqueous solution: solvation shells, interfacial water, and vibrational spectroscopy from ab initio molecular dynamics*, J. Chem. Phys. **133**, 114508, 2010.
68. B. Ensing, M. De Vivo, Z. W. Liu, P. Moore, M. L. Klein, *Metadynamics as a tool for exploring free energy landscapes of chemical reactions*, Accounts Chem. Res. **39**, 73-81, 2006.
69. V. Leone, F. Marinelli, P. Carloni and M. Parrinello, *Targeting biomolecular flexibility with metadynamics* Curr. Opin. Struct. Bio. **20**, 148, 2010.
70. A. Barducci, M. Bonomi and M. Parrinello, *Metadynamics*, Wiley Interdiscip. Rev. Comput. Mol. Sci. **1**, 826, 2011.
71. X. Biarns, S. Bongarzone, A. Vargiu, P. Carloni and P. Ruggerone, *Molecular motions in drug design: the coming age of the metadynamics method*, J. Comput. Aid. Mol. Des. **25**, 395, 2011.
72. X. Biarnés, A. Ardèvol, A. Planas, C. Rovira, A. Laio and M. Parrinello *The conformational free energy landscape of β-D-glucopyranose. Implications for substrate preactivation in beta-glucoside hydrolases*, J. Am. Chem. Soc. **129**, 10686, 2007.
73. P. Vidossich, M. Alfonso-Prieto, X. Carpena, I. Fita, P. C. Loewen and C. Rovira, *The dynamic role of distal side residues in heme hydroperoxidase catalysis. Interplay between X-ray crystallography and ab initio MD simulations*, Arch Biochem Biophys. **500**, 37, 2010.

74. M. Dal Peraro, A. J. Vila, and P. Carloni, *Catalytic Mechanism of Metallo beta-Lactamases: Insights from Calculations and Experiments*, in Quantum Biochemistry, C. F. Matta, Ed., WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim, 2010.
75. E. Brunk, N. Ashari, P. Athri, P. Campomanes, F. F. de Carvalho, B. F. E. Curchod, P. Diamantis, M. Doemer, J. Garrec, A. Laktionov, M. Micciarelli, M. Neri, G. Palermo, T. J. Penfold, S. Vanni, I. Tavernelli, U. Rothlisberger, *Pushing the Frontiers of First-Principles Based Computer Simulations of Chemical and Biological Systems*, CHIMIA **65**, 667, 2011.
76. F. L. Gervasio, A. Laio, M. Iannuzzi, M. Parrinello, *Influence of DNA structure on the reactivity of the guanine radical cation* Chem. Eur. J. **10**, 4846, 2004.
77. M. Boero, F. L. Gervasio, M. Parrinello *Charge localisation and hopping in DNA* Mol. Simulat. **33**, 57, 2007.
78. Y. A. Mantz, F. L. Gervasio, T. Laino, M. Parrinello, *Solvent effects on charge spatial extent in DNA and implications for transfer* Phys. Rev. Lett. **99**, 058104, 2007.
79. G. Grönhof, M. Bouxin-Cademartory, B. Hess, S. P. De Visser, H. J. C. Berendsen, M. Olivucci, A. E. Mark, and M. A. Robb *Photoactivation of the photoactive yellow protein: Why photon absorption triggers a trans-to-cis isomerization of the chromophore in the protein*, J. Am. Chem. Soc. **126**, 4228, 2004.
80. M. E. Moret, E. Tapavizca, L. Guidoni, U. F. Röhrig, M. Sulpizi, I. Tavernelli, and U. Rothlisberger, *Quantum mechanical / molecular mechanical (QM/MM) Car-Parrinello simulations in excited states*, CHIMIA **59**, 493, 2005.
81. U. F. Röhrig, I. Frank, J. Hutter, and U. Rothlisberger *QM/MM Car-Parrinello molecular dynamics study of the solvent effects on the ground state and on the first excited singlet state of acetone in water*, Chem. Phys. Chem. **4**, 1177, 2003.
82. E. Tapavicza, I. Tavernelli, and U. Rothlisberger, *Trajectory surface hopping within linear response time-dependent density functional theory*, Phys. Rev. Lett. **98**, 023001, 2007.
83. L. González, D. Escudero and L. Serrano-Andrés, *Progress and Challenges in the Calculation of Electronic Excited States* ChemPhysChem **13**, 28, 2011.
84. Y. Ma, M. Rohlfing, and C. Molteni, *Modeling the Excited States of Biological Chromophores within Many-Body Green's Function Theory* J. Chem. Theory Comput. **6**, 257, 2010.
85. D. Rocca, D. Lu and G. Galli, *Ab initio calculations of optical absorption spectra: Solution of the Bethe-Salpeter equation within density matrix perturbation theory* J. Chem. Phys. **133**, 164109, 2010.

Simulation Techniques for Studying the Impact of Force on (Bio)chemical Processes

Frauke Graeter[1,2] **and Wenjin Li**[2]

[1] Heidelberg Institute for Theoretical Studies
Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany
E-mail: frauke.graeter@h-its.org

[2] CAS-MPG Partner Institute and Key Laboratory for Computational Biology
Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences
320 Yue Yang Road, Shanghai 200031, China
E-mail: liwenjin@picb.ac.cn

1 Introduction: How Force Affects Chemical Bonds

Mechanochemistry, less well-known than thermochemistry, electrochemistry and photochemistry, studies the impact of mechanical force on chemical reactions. Mechanical force is recently recognized as a catalyst of chemical reactions, as an increasing number of studies suggest force to accelerate bimolecular reactions. How does a stretching force acting on a chemical bond alters the molecule's reactivity or reaction mechanism? How does a force weaken (or in contrast stabilize) a reactive molecular entity and thereby steers a reaction along tilted free energy landscapes and different pathways?

Thiol-disulfide exchange has been recently studied intensively under mechanical force. At first, the reduction of a disulfide bond, which was engineered into an immunoglobulin domain of titin (I27), by dithiothreitol (DTT) was investigated by AFM[1]. The reduction rate was found to exponentially increase with the applied stretching force in the range of 100 to 600 pN. This tendency that a bond opens more readily when it is pulled, is intuitively expected. Recently, we have performed computational studies of the same redox system in a force range up to 2000 pN, and revealed a shift of the transition state structure at high forces[19]. Δx_{r}, the distance along the pulling coordinate between the reactant and the transition state, is an important quantity to characterize the force-dependence of a chemical reaction. When different reducing agents were used in the experiments, the measured Δx_{r} varied from 0.23 Å to 0.46 Å[4]. This is an indicator of structural differences among the transition states of disulfide reduction catalysed by different agents. Surprisingly, the disulfide reduction rate shows a more complex force dependency when an enzyme (thioredoxin) is used as catalyst. It is found to decrease at first upon application of small forces, and then to increase at larger forces when an enzyme (thioredoxin) is used as catalyst[2,3]. As another example of a mechanochemical reaction, Brantley *et al.*[14] recently reported that mechanical force selectively transformed triazoles into their azide and alkyne precursors with high fidelity, a potential step forward to develop mechanoresponsive materials.

Interestingly, mechanical force can change the pathways of chemical reactions. Pathways, which are thermochemically difficult or impossible, can take place under external force. 1,2-disubstituted benzocyclobutene can occur as a trans- and cis-isomer. Under thermal activation, the Woodward-Hoffman rules[7] predict that both trans- and cis- isomer undergo conrotatory ring opening and yield the intermediates of E,E-isomer and E,Z-isomer,

respectively. However, mechanical force is reported to induce a disrotatory ring opening in the cis-isomer and a conrotatory ring opening in the trans-isomer. Thus, both isomer yield the same intermediate of E,E-isomer[5]. A later computational work in Martinez's group again supports a mechanical-activated disrotatory ring opening under cis-pulling[6].

Mechanophores, mechanically sensitive chemical groups, are of importance in the design of new materials. Mechanical force can induce the ring-opening of the colourless spiropyran into the coloured merocyanine, which will be reversed if exposed to visible light[12]. Davis *et al.* have synthesized a spiropyran-linked polymer, which changes its color under tensile loading[13]. In this way, one can visualize the damages of materials under stress.

How have such force-dependent reactivities been interpreted? Bell's model assumes that the structure of a transition state is force-independent, which seems not always true in at least, e.g., the disulfide reduction by DTT[19]. Bell's model is also found to be inadequate to describe the relationship between the rate and the external force in a forced unfolding study of I27 by AFM and a forced unzipping of DNA hairpin in a nanopore[8,10,11]. Recently, a new model has been proposed in Hummer's group[8]

$$r(F) = r_0(1 - vF\Delta x_{\mathrm{r}}/\Delta G^{\ddagger})^{1/\mathrm{v}-1}\exp\{\beta\Delta G^{\ddagger}[1-(1-vF\Delta x_{\mathrm{r}}/\Delta G^{\ddagger})^{1/\mathrm{v}}]\}. \quad (1)$$

Here, v characterizes the shape of the potential energy landscape and equals 1/2 or 2/3, $\Delta G^{\ddagger}$ donates the apparent free energy of activation, and β corresponds to $1/kT$, where k is the Boltzmann constant and T is the temperature. The rate at zero force is r_0. This model takes the effect of the application of force on the structure of transition state into account.

However, this is a phenomenological model that assumes a the force to alter the free energy landscape along exactly a single degree of freedom, x. However, the molecular detailed impact of force onto the reacting molecule into various degrees of freedom, which even might be orthogonal to the direction of force application, can not be excluded. To understand the effect of force on the reactivity, quantum chemical calculations are needed, in which the interplay of a stretching force and the electronic structure can be monitored. Recently, much progress has been made in this direction. Among others, Marx et al. have developed a framework to assess the tilting of energy landscapes by force and applied it to the ring opening of cyclobutene into a trans isomer discussed above and to the opening of cyclopropane[15,16].

In these lecture notes, we are focusing on our approach of using transition path sampling and rate calculations to quantify force-induced reactivity, and on our application to thiol/disulfide exchange[19]. We note that we have recently extended our approach of hybrid quantum/classical mechanical simulations (QM/MM) to the direct calculation of redox potentials of disulfide bonds under force[18].

2 Methods

2.1 Combined Quantum Mechanical/Molecular Mechanical Simulations

Molecular mechanical (MM) force fields are built on empirical potentials, and thus are not capable to describe systems that involve, for example, covalent bond formation or breaking. Such systems require quantum mechanics (QM) for a precise treatment. However, QM

can treat only relatively small systems, which consist of tens or several hundreds of atoms. If one requires to simulate chemical reactions that occur in a large system, e.g., enzymatic reactions, combined quantum mechanical/molecular mechanical (QM/MM) simulation is a method of choice. In QM/MM simulations, the chemical reaction center and its surrounding atoms are treated by QM, while the remainder is described by MM. Therefore, QM/MM combines the accuracy of QM with the low computational cost of MM. For more details on QM/MM, we recommend two recent reviews: one by Lin and Truhlar[17], and the other by Senn and Thiel[20].

2.2 Transition Path Sampling

Many processes like chemical reactions or protein folding can be simplified to processes with two stable states that are separated by a single high energy barrier. As depicted in Fig. 1a, region A and B are the two stable states, and the energy barrier is highlighted in the middle. For chemical reactions, region A and B represent the reactant and product states, respectively. In this example, the multi-dimensional space of the system is projected onto two order parameters, $R1$ and $R2$, both of which change during the reactions. Examples for order parameters are given further below. A reactive trajectory (shown as a black solid line) leads to the rare but crucial transition between A and B. The system spends considerably longer times in the two free energy wells of the reactant and product than in the high free energy states between the two. Thus, while the transition of interest might only take a few 100 fs, the dwell time of the system in A or B might be in the microsecond to second time scale. Transition path sampling (TPS) has been developed to enhance the sampling of the rare reactive trajectories, which are otherwise hardly harvested by conventional simulations[21–25].

2.2.1 Sampling the Transition Path Ensemble

The idea of transition path sampling is to sample a new transition path based on an existing (old) one (a transition path refers to a reactive trajectory) with a Monte Carlo procedure, and the new pathway is made sure to be equally weighted with the old one in the transition path ensemble. In principle, there are many strategies to do it. Here, we take the shooting move in deterministic simulation as an example to give readers a concrete concept of what TPS does.

a) Defining the probability of a reactive path. In molecular simulations, the time evolution of a system is represented by an ordered sequence of states, $X(T) \equiv \{X_0, X_{\Delta \mathrm{t}}, X_{2\Delta \mathrm{t}}, ...X_{\mathrm{T}}\}$ (see Fig. 1a, black solid line). Here, Δt is the time increment. $X(T)$ consists of $L = T/\Delta t + 1$ states, and its starting point is X_0. For deterministic dynamics, the probability of a trajectory equals to the probability of the initial state in a given ensemble, $\rho(X_0)$. Therefore, the probability of trajectory $X(T)$ to be a reactive trajectory is given below:

$$P_{\mathrm{AB}}(X(T)) = h_A(X_0)\rho(X_0)h_{\mathrm{B}}(X_{\mathrm{T}})/Z_{\mathrm{AB}}(T) \tag{2}$$

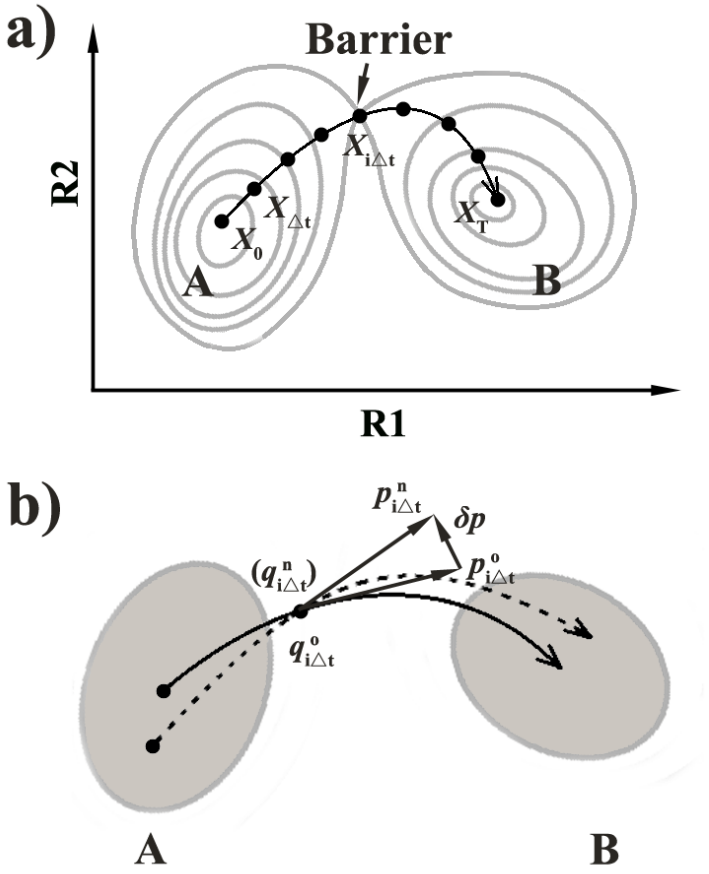


Figure 1. Schematical description of the free energy landscape of a system and the shooting move in TPS. a) A typical free energy landscape of a process is shown with two stable states (labelled with A and B) and a barrier in the middle. R1 and R2 are two arbitrary coordinates. A transition pathway (black solid line) connecting states A and B is given as well. The transition path is represented by an ordered sequence of states $\{X_0, X_{\Delta t}, X_{2\Delta t}, ...X_T\}$. b) An example of shooting moves. The two filled grey areas represents the states A and B mentioned above. A state $\{q^o_{i\Delta t}, p^o_{i\Delta t}\}$ is randomly chosen from an old transition path (solid line). The momentum $p^o_{i\Delta t}$ is perturbed to be $p^n_{i\Delta t}$, where $p^n_{i\Delta t} = p^o_{i\Delta t} + \delta p$, while the coordinate is unchanged with $q^o_{i\Delta t} = q^n_{i\Delta t}$. From the newly generated state $\{q^n_{i\Delta t}, p^n_{i\Delta t}\}$, a new transition path (dashed line) is obtained by evolving the system backward in time to zero and forward in time to T.

Here, $h_A(X)$ ($h_B(X)$) is the characteristic functions of region A (B). $h_A(X)$ equals 1 if state X lies in A, and it equals zero otherwise. $Z_{AB}(T)$ is the normalizing factor, the sum of all the possible reactive trajectories with length T in a given ensemble.

$$Z_{AB}(T) \equiv \int dX_0 h_A(X_0)\rho(X_0)h_B(X_T) \tag{3}$$

b) Sampling the transition path ensemble by shooting. In a transition path ensemble, the distribution of transition paths is given in Eq. 2. To make sure that the correctly weighted transition paths are sampled, the following two probabilities should equal: the probability to generate a new transition path from a old one $P_{\rm gen}(X^{\rm o}(T) \to X^{\rm n}(T))$, and the probability to generate the old transition path from the new one $P_{\rm gen}(X^{\rm n}(T) \to X^{\rm o}(T))$. In a shooting move, a state $X^{\rm o}_{{\rm i}\Delta{\rm t}}$, ${\rm i} \in [0, L]$, is randomly chosen. Then, a new state $X^{\rm n}_{{\rm i}\Delta{\rm t}}$ is generated by adding a small perturbation to $X^{\rm o}_{{\rm i}\Delta{\rm t}}$. Here, the superscript o and n refer to the old path and the new path, respectively. Note that a state X consists of the coordinate q and the momentum p, $X = \{q, p\}$, the perturbation can be added to q or/and p. In practice, it's convenient to keep q untouched and change p by δp. As illustrated in Fig. 1b, the selected state $X^{\rm o}_{{\rm i}\Delta{\rm t}} = \{q^{\rm o}_{{\rm i}\Delta{\rm t}}, p^{\rm o}_{{\rm i}\Delta{\rm t}}\}$ in a old transition path (the solid line in Fig. 1b) is changed to $X^{\rm n}_{{\rm i}\Delta{\rm t}} = \{q^{\rm n}_{{\rm i}\Delta{\rm t}}, p^{\rm n}_{{\rm i}\Delta{\rm t}}\}$, where $p^{\rm n}_{{\rm i}\Delta{\rm t}} = p^{\rm o}_{{\rm i}\Delta{\rm t}} + \delta p$. Starting with $X^{\rm n}_{{\rm i}\Delta{\rm t}}$, one can evolve the system backward in time to 0 and forward in time to T, then a new transition path is generated if it initials from region A and ends in region B (the dashed line in Fig. 1b). The probability to generate a new transition path from a old one is the product of four parts, the probability of the old path in the given ensemble, the probability to generate $X^{\rm n}_{{\rm i}\Delta{\rm t}}$ from $X^{\rm o}_{{\rm i}\Delta{\rm t}}$ ($P_{\rm gen}(X^{\rm o}_{{\rm i}\Delta{\rm t}} \to X^{\rm n}_{{\rm i}\Delta{\rm t}})$), the probability of that the new path is reactive, and the probability to accept the new transition path $P_{\rm acc}(X^{\rm o}(T) \to X^{\rm n}(T))$.

$$\begin{aligned} P_{\rm gen}(X^{\rm o}(T) \to X^{\rm n}(T)) = {} & P_{\rm AB}(X^{\rm o}(T)) P_{\rm gen}(X^{\rm o}_{{\rm i}\Delta{\rm t}} \to X^{\rm n}_{{\rm i}\Delta{\rm t}}) h_A(X^{\rm n}_0) h_B(X^{\rm n}_{\rm T}) \\ & \times P_{\rm acc}(X^{\rm o}(T) \to X^{\rm n}(T)) \end{aligned} \tag{4}$$

Similarly, for generating the old path from the new one, we have

$$\begin{aligned} P_{\rm gen}(X^{\rm n}(T) \to X^{\rm o}(T)) = {} & P_{\rm AB}(X^{\rm n}(T)) P_{\rm gen}(X^{\rm n}_{{\rm i}\Delta{\rm t}} \to X^{\rm o}_{{\rm i}\Delta{\rm t}}) h_A(X^{\rm o}_0) h_B(X^{\rm o}_{\rm T}) \\ & \times P_{\rm acc}(X^{\rm n}(T) \to X^{\rm o}(T)) \end{aligned} \tag{5}$$

The *detailed balance* of moves in trajectory space requires $P_{\rm gen}(X^{\rm o}(T) \to X^{\rm n}(T)) = P_{\rm gen}(X^{\rm n}(T) \to X^{\rm o}(T))$, which gives

$$\frac{P_{\rm acc}(X^{\rm o}(T) \to X^{\rm n}(T))}{P_{\rm acc}(X^{\rm n}(T) \to X^{\rm o}(T))} = \frac{P_{\rm AB}(X^{\rm n}(T)) P_{\rm gen}(X^{\rm n}_{{\rm i}\Delta{\rm t}} \to X^{\rm o}_{{\rm i}\Delta{\rm t}}) h_A(X^{\rm o}_0) h_B(X^{\rm o}_{\rm T})}{P_{\rm AB}(X^{\rm o}(T)) P_{\rm gen}(X^{\rm o}_{{\rm i}\Delta{\rm t}} \to X^{\rm n}_{{\rm i}\Delta{\rm t}}) h_A(X^{\rm n}_0) h_B(X^{\rm n}_{\rm T})} \tag{6}$$

This condition can be satisfied using a Metropolis criterion[26]

$$P_{\rm acc}(X^{\rm o}(T) \to X^{\rm n}(T)) = min[1, \frac{P_{\rm AB}(X^{\rm n}(T)) P_{\rm gen}(X^{\rm n}_{{\rm i}\Delta{\rm t}} \to X^{\rm o}_{{\rm i}\Delta{\rm t}}) h_A(X^{\rm o}_0) h_B(X^{\rm o}_{\rm T})}{P_{\rm AB}(X^{\rm o}(T)) P_{\rm gen}(X^{\rm o}_{{\rm i}\Delta{\rm t}} \to X^{\rm n}_{{\rm i}\Delta{\rm t}}) h_A(X^{\rm n}_0) h_B(X^{\rm n}_{\rm T})}] \tag{7}$$

Note that the old path is reactive, i.e., $h_A(X^{\rm o}_0) = 1$ and $h_B(X^{\rm o}_{\rm T}) = 1$. Eq. 7 can be simplified as

$$P_{\rm acc}(X^{\rm o}(T) \to X^{\rm n}(T)) = h_A(X^{\rm n}_0) h_B(X^{\rm n}_{\rm T}) \times min[1, \frac{\rho(X^{\rm n}_{{\rm i}\Delta{\rm t}}) P_{\rm gen}(X^{\rm n}_{{\rm i}\Delta{\rm t}} \to X^{\rm o}_{{\rm i}\Delta{\rm t}})}{\rho(X^{\rm o}_{{\rm i}\Delta{\rm t}}) P_{\rm gen}(X^{\rm o}_{{\rm i}\Delta{\rm t}} \to X^{\rm n}_{{\rm i}\Delta{\rm t}})}] \tag{8}$$

Here, we apply Eq. 2 and the fact that the probability of the states on the same path in deterministic dynamics are the same. Although Eq. 8 is obtained base on the deterministic dynamics, it can be inferred base on a general dynamics[25]. In the implementation of shooting moves, normally a symmetric generation probability is ensured, and thus

$P_{\mathrm{gen}}(X^{\mathrm{o}}_{\mathrm{i}\Delta\mathrm{t}} \rightarrow X^{\mathrm{n}}_{\mathrm{i}\Delta\mathrm{t}}) = P_{\mathrm{gen}}(X^{\mathrm{n}}_{\mathrm{i}\Delta\mathrm{t}} \rightarrow X^{\mathrm{o}}_{\mathrm{i}\Delta\mathrm{t}})$. Specific strategies are always applied to ensure that states $X^{\mathrm{o}}_{\mathrm{i}\Delta\mathrm{t}}$ and $X^{\mathrm{n}}_{\mathrm{i}\Delta\mathrm{t}}$ are within the same microcanonical ensemble, i.e., $\rho(X^{\mathrm{o}}_{\mathrm{i}\Delta\mathrm{t}}) = \rho(X^{\mathrm{n}}_{\mathrm{i}\Delta\mathrm{t}})$. Thus, the acceptance probability becomes

$$P_{\mathrm{acc}}(X^{\mathrm{o}}(T) \rightarrow X^{\mathrm{n}}(T)) = h_A(X^{\mathrm{n}}_0)h_B(X^{\mathrm{n}}_{\mathrm{T}}) \tag{9}$$

This equation states that any new trajectory will be accepted if it initials from region A and ends in region B.

2.2.2 Computing Rate Constants

In this section, we explain how to obtain rate constants from the transition path ensemble[24]. Given a system with two stable states A and B, which is separated by a single high energy barrier, molecules transit from one state to the other at equilibrium, while the populations of states remain unchanged. Since such transitions are rare, the time correlation function, $C(t)$, relates to the reaction time of the system ($\tau_{\mathrm{rxn}} \equiv (k_{\mathrm{AB}} + k_{\mathrm{BA}})^{-1}$) via the following formula[27]

$$C(t) \approx \langle h_{\mathrm{B}} \rangle (1 - exp\{-t/\tau_{\mathrm{rxn}}\}) \tag{10}$$

If the time required for a system to cross the energy barrier and commit to the other stable state (τ_{mol}) is far smaller than the reaction time of the system (i.e., $\tau_{\mathrm{mol}} << \tau_{\mathrm{rxn}}$), $C(t)$ scales linearly in the intermediate time region, and we have

$$C(t) \approx k_{\mathrm{AB}} t, \tau_{\mathrm{mol}} < t << \tau_{\mathrm{rxn}} \tag{11}$$

For a system at equilibrium, $C(t)$ characterizes the conditional probability to find the system in state B at time t if it was in state A at time zero, and is defined as follows

$$C(t) \equiv \frac{\langle h_{\mathrm{A}}(X_0) h_{\mathrm{B}}(X_{\mathrm{t}}) \rangle}{\langle h_{\mathrm{A}}(X_0) \rangle} \tag{12}$$

Here, $\langle ... \rangle$ is the ensemble average of all initial states. In deterministic dynamics, $C(t)$ can be written in terms of the probability of all initial states $\rho(X_0)$:

$$C(t) = \frac{\int dX_0 \rho(X_0) h_{\mathrm{A}}(X_0) h_{\mathrm{B}}(X_{\mathrm{t}})}{\int dX_0 \rho(X_0) h_{\mathrm{A}}(X_0)} \tag{13}$$

Eq. 11 and Eq. 13 together provide a way to calculate the forward reaction rate constant k_{AB} by molecular simulations. One can simply run a large set of simulations that start with states in region A and are with the same time length of t, and then counts the probability of the end state in region B, which gives the value of $C(t)$. The derivative of $C(t)$ over time gives the rate constant. However, this apparently involves numerous computational efforts.

If region B can be defined by an order parameter $\lambda(X)$, and the distribution of the end states, i.e., $X(t)$, along the order parameter $P(\lambda, t)$ is known, $C(t)$ is simply the integral of $P(\lambda, t)$ along λ over the region B.

$$C(t) = \int_{\lambda_{\min}}^{\lambda_{\max}} d\lambda P(\lambda, t). \tag{14}$$

Here, $\lambda_{\min}$ and $\lambda_{\max}$ are the lower and upper bound of region B along λ. $P(\lambda, t)$ is given by

$$P(\lambda, t) = \frac{\int dX_0 \rho(X_0) h_{\mathrm{A}}(X_0) \delta[\lambda - \lambda(X(t))]}{\int dX_0 \rho(X_0) h_{\mathrm{A}}(X_0)}, \tag{15}$$

where $\delta(X)$ is Dirac's delta function. $P(\lambda, t)$ can be divided into several overlapped windows, and its distribution in each window can be estimated separately. The distribution of $P(\lambda, t)$ over the whole range of λ is then obtained by connecting all windows. In each window, transition path sampling can be applied to enhance the sampling of paths that connect region A and the window region. Therefore, computational efforts to compute $C(t)$ are dramatically reduced.

The above-mentioned method can only compute $C(t)$ in time t at a time, and the evaluation of K_{AB} requires $C(t)$ at different times to be evaluated. Therefore, it is laborious. Fortunately, $C(t)$ can be factorized to be written as

$$C(t) = \frac{\langle h_{\mathrm{B}}(t) \rangle_{\mathrm{AB}}}{\langle h_{\mathrm{B}}(t') \rangle_{AB}} \times C(t'), 0 < t < T \tag{16}$$

where $\langle ... \rangle_{\mathrm{AB}}$ denotes an average on the ensemble of the reactive paths, which start in region A and visit region B within the time length of T. T is the time length of the transition path. $\langle h_B(t) \rangle_{\mathrm{AB}}$ is then the proportion of reactive paths whose configuration at time t belonged to region B, and can be estimated by a single transition path sampling run. Only $C(t')$, the $C(t)$ at time t' ($t' < T$), is needed to be evaluated.

2.2.3 Committor Probability

Starting with a structure, the system will visit first either region A or region B depending on its initial momenta. If the initial momenta are drawn from an appropriate distribution, e.g., Boltzmann distribution, the committor probability of the structure defines the probability of the system to visit first region B, P_{B}. Importantly, transition states are configurations that have equal probability to visit regions A and B, that is to say, transition states are the structures with a committor probability of $1/2$. Therefore, the estimation of P_{B} of structures provides a way to identify transition states. Computing P_{B} of a given structure can be performed as follows. Initiate a finite number N of fleeting trajectories from the structure with a momentum drawn from Boltzmann distribution, count the number x of trajectories which reach region B first, and then P_{B} is given by x/N in a standard deviation $\delta = [P_{\mathrm{B}}(1 - P_{\mathrm{B}})/N]^{1/2}$.

3 Application: Disulphide Bond Reduction

The reduction of a protein disulfide bond to two thiol groups by a small reducing molecule such as dithiothreitol (DTT) showed an increase in reaction rate with mechanical force[1].

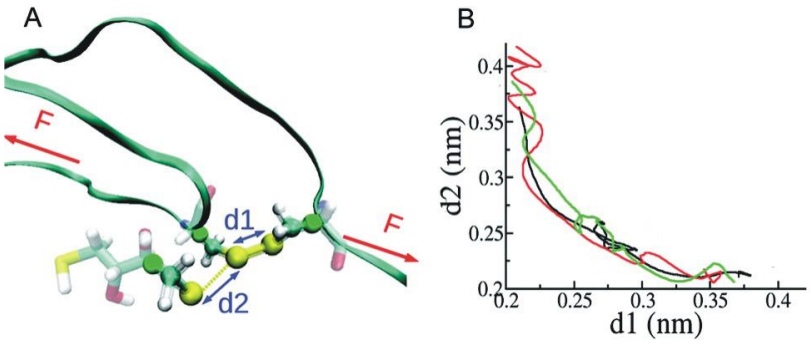


Figure 2. Disulfide bond reduction under force. (A) The system comprises the protein disulfide bond and DTT, which attacks one of the sulfur atoms of the protein. Atoms treated by QM are shown as spheres. Other atoms including the surrounding water molecules (not shown) are treated by MM. The way of force application is indicated by arrows. More specifically, a constant force was applied to the terminal C_α-atoms of the protein. (B) Three representative dynamic transitions obtained from TPS are projected onto $d1$ and $d2$. As can be seen, the DTT sulfur first approaches the disulfide bond, which is followed by (not preceded by or in parallel with) the lengthening of the cleaving bond. Adopted from Ref. 19.

The exponential increase was in line with the Bell model and interpreted in the light of this one-dimensional model. To elucidate the underlying mechanism and its potential force-dependency we set out to simulate with high-performance computing methods this reaction. We chose the methods outlined above, namely a QMMM description of the system and TPS including reaction rate and committor probability calculations.

For the transition path sampling, a choice must be made for the definition of reactants and products. During the reaction, as shown in Fig. 2, the sulfur-sulfur bond length of the disulfide group in the protein, $d1$, extends, while a new disulfide bond between DTT and protein, described by the distance $d2$ shortens. We chose these two order parameters to define the reactant and product states. More specifically, $\lambda < -0.12\,\mathrm{nm}$ and $\lambda > 0.12\,\mathrm{nm}$ were specifying the regions of reactant and product state, respectively, where $\lambda = d1 - d2$.

The obtained reaction rates are shown in Fig. 3. In agreement with experiments, rates increase with force. In fact, when using the Bell model, as in the experiments, within the force range of 0-500 pN, we obtained a distance between reactant and transition state, which is quantitatively in accordance with experiments. Interestingly, however, the linear dependency between the logarithm of rates and the forces is violated at larger forces, and can be readily fitted with the Dudko-Hummer model instead (see Eq. 4). This non-linearity of the rates can be interpreted by a shift in the transition state by the mechanical force. In other words, force tilts the free energy landscape such that the transition state moves along the reaction coordinate. What, if so, are the order parameters to describe this shift?

To answer this question, we next set out to determine the transition state ensemble from committor probabilities. As described in more detail in the Methods section, we were initiating trajectories with random velocities from our reactive trajectories, and counted the arrivals in either state A or B. Structures on the $d1$ and $d2$ surface that committed nearly equally to the reactants and products, i.e. with P_B and P_A lying between 0.4 and 0.6, were considered as conformations belonging to the transition state. For comparably low forces, we find transition states to be symmetric, with $d1 \sim d2$, force renders transition states

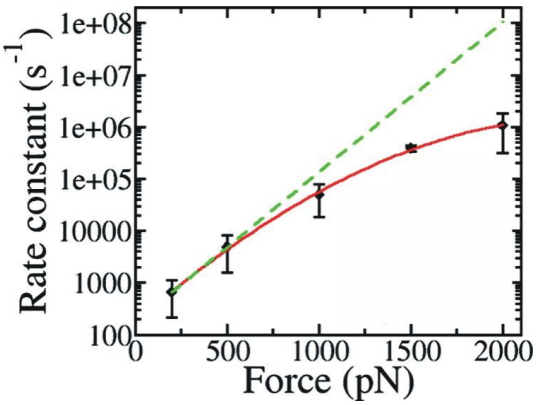


Figure 3. Reaction rates increase with force according to our simulations (diamonds). The increase observed between 0 and 500 pN is quantitatively in line with the experiments. However, at larger forces, the rates diverge from a linear dependence between the logarithm of rates and the forces, as suggested by Bell's model (dashed line), but instead follow the Dudko-Hummer model described above (solid line). Adopted from Ref. 19.

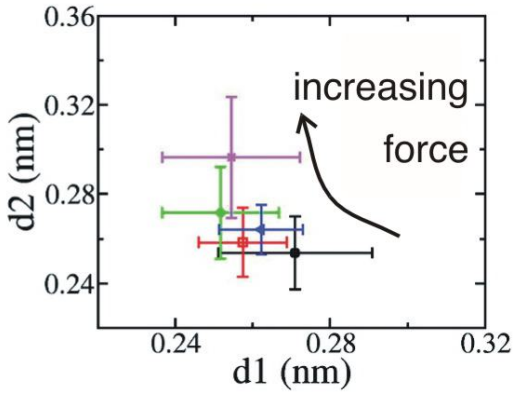


Figure 4. Transition states change upon force application. In contrast to what is commonly assumed and the basis of the Bell model, we find a shift in transition state along the two order parameters with force. Transition states have been obtained from committor probability calculations. Adopted from Ref. 19.

increasingly asymmetric. More specifically, the attacking sulfur does not have to approach the disulfide bond as much anymore (measured by $d2$) to reach the free energy barrier, and induces the bond cleavage ($d1$ increase) already at larger distances. In other words, the transition states moves towards the reactant states under force application.

Importantly, our results suggest more than a single order parameter to play an important role in describing the process of the reaction. While experiments measure end-to-end distance, thereby largely probing changes in $d1$, also the distance between the two reacting molecules, a length that experiments are blind for, contributes to the reaction coordinate.

4 Outlook

The combination of TPS and QMMM has proven very useful in the context of mechanochemistry. The advantage of TPS over other methods that yield free energy profiles is the direct calculation of rates, without the need of assuming a certain attempt frequency or alike. This renders TPS and rate calculations an optimal choice for simulations that aim at explaining experimentally measured force-dependent rates as they are obtained from force spectroscopy experiments. An ultimate aim of such and similar simulations is to quantitatively link the externally applied force of AFM or optical tweezer experiments or the internally produced force of a strained ring mechanophore to the internal stress at the reaction center of the molecule. Internal stress here refers to the distortion of the reactive bonds, angles and electronic orbitals away from their equilibrium states. Such an internal force distribution analysis has been recently developed in our group for classical mechanical force fields[28]. It makes use of the inter-atomic forces in the structure, comprising both bonded and non-bonded forces, to reveal the internal stress in the molecular scaffold. It is reminiscent of finite element methods used in engineering to detect the stress distribution in macroscopic objects such as cars or wheels. The concepts will be given during the lecture, and details are given in the relevant publications[28]. The force distribution analysis might prove useful for studying mechanochemical effects as an approach complementary to studying the free energies or rates of transitions.

References

1. Wiita, A. P. and Ainavarapu, S. R. and Huang, H. H. and Fernandez, J. M. Force-dependent chemical kinetics of disulfide bond reduction observed with single-molecule techniques, Proc. Natl. Acad. Sci. U.S.A. **103**, 7222–7227, 2006.
2. Wiita, A. P. and Perez-Jimenez, R. and Walther, K. A. and Graeter, F. and Berne, B. J. and Holmgren, A. and Sanchez-Ruiz, J. M. and Fernandez, J. M. *Probing the chemistry of thioredoxin catalysis with force*, Nature **450**, 124–127, 2007.
3. Perez-Jimenez, R. and Li, J. and Kosuri, P. and Sanchez-Romero, I. and Wiita, A. P. and Rodriguez-Larrea, D. and Chueca, A. and Holmgren, A. and Miranda-Vizuete, A. and Becker, K. and Cho, S. H. and Beckwith, J. and Gelhaye, E. and Jacquot, J. P. and Gaucher, E. A. and Gaucher, E. and Sanchez-Ruiz, J. M. and Berne, B. J. and Fernandez, J. M. *Diversity of chemical mechanisms in thioredoxin catalysis revealed by single-molecule force spectroscopy*, Nat. Struct. Mol. Biol. **16**, 890–896, 2009.
4. Koti Ainavarapu, S.R. and Wiita, A.P. and Dougan, L. and Uggerud, E. and Fernandez, J.M. *Single-molecule force spectroscopy measurements of bond elongation during a bimolecular reaction*, Journal of the American Chemical Society **130**, 6479–6487, 2008.
5. Hickenboth, C.R. and Moore, J.S. and White, S.R. and Sottos, N.R. and Baudry, J. and Wilson, S.R. *Biasing reaction pathways with mechanical force*, Nature **446**, 423–427, 2007.
6. Ong, M.T. and Leiding, J. and Tao, H. and Virshup, A.M. and Martinez, T.J. *First principles dynamics and minimum energy pathways for mechanochemical ring opening of cyclobutene*, Journal of the American Chemical Society **131**, 6377–6379, 2009.

7. Woodward, R.B. and Hoffmann, R. *The conservation of orbital symmetry*, Angewandte Chemie International Edition in English **8**, 781–853, 1969.
8. Dudko, O.K. and Hummer, G. and Szabo, A. *Theory, analysis, and interpretation of single-molecule force spectroscopy experiments*, Proceedings of the National Academy of Sciences **105**, 15755, 2008.
9. Bell, G.I. *Models for the specific adhesion of cells to cells*, Science **200**, 618, 1978.
10. Schlierf, M. and Rief, M. *Single-molecule unfolding force distributions reveal a funnel-shaped energy landscape*, Biophysical Journal **90**, L33–L35, 2006.
11. Mathé, J. and Visram, H. and Viasnoff, V. and Rabin, Y. and Meller, A. *Nanopore unzipping of individual DNA hairpin molecules*, Biophysical Journal **87**, 3205–3212, 2004.
12. Minkin, V.I. *Photo-, thermo-, solvato-, and electrochromic spiroheterocyclic compounds*, Chemical Reviews **104**, 2751–2776, 2004.
13. Davis, D.A. and Hamilton, A. and Yang, J. and Cremar, L.D. and Van Gough, D. and Potisek, S.L. and Ong, M.T. and Braun, P.V. and Martínez, T.J. and White, S.R. and others *Force-induced activation of covalent bonds in mechanoresponsive polymeric materials*, Nature **459**, 68–72, 2009.
14. Brantley, J.N. and Wiggins, K.M. and Bielawski, C.W. *Unclicking the Click: Mechanically Facilitated 1, 3-Dipolar Cycloreversions*, Science **333**, 1606–1609, 2011.
15. Ribas-Arino J., and Shiga M., and Marx D. *Understanding covalent mechanochemistry*, Angew Chem Int Ed Engl. **23**, 4190–3, 2009.
16. Dopieralski P., and Ribas-Arino J., and Marx D. *Force-transformed free-energy surfaces and trajectory-shooting simulations reveal the mechano-stereochemistry of cyclopropane ring-opening reactions*, Angew Chem Int Ed Engl. **31**, 7105–8, 2011.
17. Lin, H. and Truhlar, D.G. *QM/MM: what have we learned, where are we, and where do we go from here?*, Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta) **117**, 185–199, 2007.
18. Baldus I., and Graeter, F. *Mechanical force can fine-tune redox potentials of disulfide bonds*, Biophysical Journal, in press, 2011.
19. Li, W., and Graeter, F. *Atomistic Evidence of How Force Dynamically Regulates Thiol/Disulfide Exchange*, Journal of the American Chemical Society **132**, 16790–5, 2010.
20. Senn, H.M. and Thiel, W. *QM/MM methods for biomolecular systems*, Angewandte Chemie International Edition **48**, 1198–1229, 2009.
21. Dellago, C. and Bolhuis, P.G. and Csajka, F.S. and Chandler, D. *Transition path sampling and the calculation of rate constants*, The Journal of Chemical Physics **108**, 1964, 1998.
22. Dellago, C. and Bolhuis, P.G. and Chandler, D. *Efficient transition path sampling: Application to Lennard-Jones cluster rearrangements*, The Journal of Chemical Physics **108**, 9236, 1998.
23. Bolhuis, P.G. and Dellago, C. and Chandler, D. *Sampling ensembles of deterministic transition pathway*, Faraday Discuss **110**, 421–436, 1998.
24. Dellago, C. and Bolhuis, P.G. and Chandler, D. *On the calculation of reaction rate constants in the transition path ensemble*, The Journal of Chemical Physics **110**, 6617, 1999.
25. Dellago, C. and Bolhuis, P.G. and Geissler, P.L. *Transition path sampling*, Advances in Chemical Physics, 1–78, 2002.

26. Metropolis, N. and Rosenbluth, A.W. and Rosenbluth, M.N. and Teller, A.H. and Teller, E. and others *On the calculation of reaction rate constants in the transition path ensemble*, The Journal of Chemical Physics **21**, 1087, 1953.
27. Chandler, D. *Statistical mechanics of isomerization dynamics in liquids and the transition state approximation*, The Journal of Chemical Physics **68**, 2959, 1978.
28. Stacklies, W., and Vega, C., and Wilmanns, M., and Graeter, F. *Mechanical network in titin immunoglobulin from force distribution analysis* PLoS Comput Biol. **5**, e1000306, 2009.

Coarse Grained Models for Multiscale Simulations of Biomolecular Systems

Christine Peter

Max Planck Institute for Polymer Research
Ackermannweg 10, 55128 Mainz, Germany
E-mail: peter@mpip-mainz.mpg.de

By systematically connecting simulation models on different levels of resolution multiscale simulation seeks to go beyond the time- and length-scale limits of high resolution models while at the same time retaining microscopic information. Development of multiscale simulation models involves systematic scale bridging where interaction functions in the lower resolution model are derived based on structural and dynamic properties of the high resolution simulation. In the present review I will focus on the systematic development of coarse grained (CG) models that are consistent with an underlying classical atomistic forcefield description. Systematic coarse graining needs to address questions such as: What is meant by consistency? What are suitable targets in such a parametrization process? Can one achieve both structural and thermodynamic agreement between atomistic and CG level? To which extent are the CG models developed at a certain thermodynamic state point / in a given chemical environment transferable to different situations? In discussing these questions I will particularly focus on problems arising in the context of biomolecular systems.

1 Introduction

Many problems and questions in biological and other soft matter systems are governed by phenomena and interactions on a wide range of length and timescales. For example, studying the folding and aggregation of biomacromolecules such as proteins or nucleic acids, the formation of virus shells, the self-assembly of lipid bilayers or structure formation in biomaterials requires length and time scales that are clearly beyond simulations with an all-atom (or even higher) resolution. To investigate these systems, so-called coarse grained (CG) models have been developed, where several atoms are grouped into superatoms – coarse grained particles[1]. Often, this reduction of degrees of freedom is accompanied by leaving out solvent degrees of freedom, resulting in a drastic decrease of the number of particles that are treated explicitly in molecular dynamics (MD) or Monte Carlo (MC) simulations. CG models give access to longer time and length scales for several reasons: (i) there is a smaller number of particles in the system, reducing the computational cost; (ii) typically, CG potentials are softer than atomistic ones, which allows to use a larger simulation timestep; (iii) due to the smoother energy landscape, the dynamics in CG systems is faster. This last aspect implies that one typically has to determine a timescaling factor between the simulation timescale (in Lennard Jones units) and the corresponding real world time (or the corresponding timescale in a higher resolution system). A detailed discussion of these dynamic aspects with further references can be found in Ref. 2.

Many CG models are generic, i.e. they were not developed to model a specific chemical system but rather with the aim to study a physical phenomenon such as folding or aggregation in general. One example are generic CG lipid models which have been successfully employed to study the self assembly of micelles, bilayers and other structures[3–7].

Generic CG models have also been employed to study folding and aggregation of peptides and proteins[8–20].

In the present review, I will focus on a different group of models: CG models for multiscale simulation purposes[21–30]. Multiscale simulations employ models at different levels of resolution (in the present case classical atomistic and CG models) sequentially or simultaneously. In the first case, the resolution of the entire system is changed in the course of the simulation. In the second case parts of the system coexist at different resolution levels in a hybrid fashion. In adaptive resolution approaches individual particles (groups of particles) can change their resolution during the course of the simulation[31,32]. Typically, in multiscale approaches, the various levels are systematically linked, to allow for a seamless change of resolution. In this framework, CG models are often systematically built up in relation to an atomistic description with the aim to keep the models on the two levels consistent. I will introduce methods to generate such CG models which are based on atomistic reference sampling, and I will discuss several of the typical questions and challenges one faces in such a coarse graining endeavor: What is meant by consistency? What are suitable targets in such a parametrization process? Can one achieve both structural and thermodynamic agreement between atomistic and CG level? To which extent are the CG models developed at a certain thermodynamic state point / in a given chemical environment transferable to different situations?

2 Deriving CG Interaction Potentials

A wide range of approaches have been developed that aim at consistency between a CG model and either experimental data or simulations of accurate high resolution models. Typically, these approaches are divided into thermodynamics-based and so called structure-based ones. In thermodynamic coarse graining approaches, individual elements of the CG interaction function are separately parameterized based on thermodynamic reference data such as solvation free energies and partitioning data, liquid densities, surface tension, etc[33–44]. (These are usually experimental reference data, but in a multiscale simulation approach, the reference data can of course also be obtained from an atomistic simulation, to keep the CG and atomistic level thermodynamically consistent.)

In another group of approaches one numerically generates CG interaction functions with the aim to reproduce the configurational phase space sampled in an atomistic reference simulation. These approaches may rely on different types of reference properties such as structure functions[21,45,46,22,47–49,29,50–54], mean forces[55,23,56–59], or relative entropies[60–62].

In a multiscale approach, one first needs to define the relationship between the two levels of resolution. This is typically done via mapping functions which determine the CG Cartesian coordinates of each site as a linear combination of coordinates for the atoms that are "involved" in the site (that could be via a center-of-mass or a center-of-geometry mapping or some other geometric construction). This means the CG coordinates $\mathbf{R}$ are constructed from the atomistic coordinates $\mathbf{r}$ via

$$\mathbf{R} = \mathbf{M}\mathbf{r} \tag{1}$$

where $\mathbf{M}$ is an $n \times N$ matrix (n and N being the number of particles in the atomistic and CG system, respectively). In the (canonical) sampling of the atomistic and CG systems

with respective interaction potentials $U^{at}(\mathbf{r})$ and $U^{CG}(\mathbf{R})$ the corresponding configuration functions $P^{at}(\mathbf{r})$ and $P^{CG}(\mathbf{R})$ are given by

$$P^{at}(\mathbf{r}) = Z_{at}^{-1} \exp[-\beta U^{at}(\mathbf{r})] \tag{2}$$

and

$$P^{CG}(\mathbf{R}) = Z_{CG}^{-1} \exp[-\beta U^{CG}(\mathbf{R})] \tag{3}$$

with $Z_{at} = \int \exp[-\beta U^{at}(\mathbf{r})] \mathrm{d}\mathbf{r}$ and $Z_{CG} = \int \exp[-\beta U^{CG}(\mathbf{R})] \mathrm{d}\mathbf{R}$ being the respective partition functions and $\beta = 1/k_B BT$.

If one analyses the atomistically sampled system in CG coordinates one can determine the probability distribution of sampling atomistic coordinates that map to a given CG coordinate $\mathbf{r}$)

$$P^{at}(\mathbf{R}) = \langle \delta(\mathbf{M}\mathbf{r} - \mathbf{R}) \rangle \tag{4}$$

(Here, I follow the notation used by Noid and collaborators, e.g. in Refs. 63, 64). The angular brackets indicate canonical sampling of the atomistic system (i.e. according to $P^{at}(\mathbf{r})$)

One can formulate the aim of many systematic coarse graining approaches in the following way: to reproduce the part of phase space which is sampled by the atomistic system. Following this, one possible definition of consistency between atomistic and CG level of resolution is that the two models are consistent if the canonical configurational distribution sampled by the CG model $P^{CG}(\mathbf{R})$ is equal to the probability distribution $P^{at}(\mathbf{R})$ obtained after mapping the atomistic system to CG coordinates.

In a canonical ensemble, independent degrees of freedom q are Boltzmann distributed and the Boltzmann inverse of $P(q)$

$$U(q) = -k_B T \ln P(q) \tag{5}$$

is a many-dimensional potential of mean force (PMF), which – used as a potential in a (for example CG) simulation – reproduces the distribution $P(q)$. This means that Boltzmann inversion of $P^{at}(\mathbf{R})$ defines (uniquely up to an additive constant) a (high-dimensional) CG potential

$$U_{PMF}^{CG}(\mathbf{R}) = -k_B T \ln P^{at}(\mathbf{R}) + const \tag{6}$$

which will result in a sampling of CG configurations which is consistent with the atomistic reference simulation. This high-dimensional, many-body CG potential is not a conventional potential energy function but a configuration-dependent free energy (PMF). This means it contains both energetic and entropic contributions from the configurational sampling in the high-resolution model and the mapping between high-resolution and CG model (Eq. 4). Therefore, the resulting CG model is state point dependent and not necessarily readily transferable. While it is conceptually easy to formulate the PMF as a solution of the systematic coarse graining task, it is practically unfeasible. In most cases the PMF cannot be easily determined, and even if it were possible, the resulting high-dimensional potentials are computationally prohibitive. In addition, $U_{PMF}^{CG}(\mathbf{R})$ is a function of $\mathbf{R}$, i.e. this PMF as is can only be applied to a system which is identical in size to the atomistic reference system, a limitation, that defeats the purpose of coarse graining. Therefore, one has to decompose the PMF into simpler independent terms, approximate it by simpler interaction functions (ideally ones that resemble interaction functions typically used in molecular

mechanics forcefields, i.e. short range bonded contributions and pair potentials or similar). Conceptually, one can decompose the PMF into a series of many-body terms (up to an N-body term, where N is the number of particles on the system). However, this itself does not solve the problem since these multi-body interactions are again computationally unfeasible.

$$U^{CG}_{PMF}(\mathbf{R}) = \sum_{i,j} U_2(r_{ij}) + \sum_{i,j,k} U_3(r_{ij}, r_{jk}, r_{ik}) + \cdots + const \tag{7}$$

$$\approx \sum_{i,j} V_{\text{eff}}(r_{ij}) + const \tag{8}$$

In Eq. 8 one approximates the series by an effective pair interaction which also contains contributions from the higher order terms in Eq. 7 (some approaches also include three body terms for systems where this is necessary[65]). There are many approaches to this task of determining effective CG interactions, and all the resulting CG models are (only) approximations to $U^{CG}_{PMF}(\mathbf{R})$.

I will describe some of these methods in more details below, but before that I would like to introduce a separation that is frequently made, namely a separation into nonbonded and bonded degrees of freedom. This separation is based on the assumption that the total potential energy can be separated into bonded and nonbonded contributions:

$$U^{CG} = U^{CG}_B + U^{CG}_{NB} \tag{9}$$

Practically, this separation is usually always realized, in the sense that the bonded interaction functions typically comprise of two, three, and four-body terms representing bonds, angles, and dihedrals in the same molecule (i.e. shorter ranged, no pair list construction, etc) , while the nonbonded interaction functions usually consist of long-range pair potentials (and in some cases three-body terms).

Some approaches try to – as cleanly as possible – separate bonded and nonbonded terms during the parametrization process, while others determine parameters for all types of interactions simultaneously[22]. In polymeric systems, a clean separation can be achieved by obtaining bond, angle and torsion distributions from sampling an isolated polymer chain with exclusions of long-range nonbonded interactions[21,66,48,67]. This ensures that nonbonded interactions that will be explicitly present in the final CG model are excluded during the conformational sampling from which the bonded interactions are derived, which strictly avoids double counting. This approach is problematic for cases where the environment has a large influence on the conformations of the molecule since it assumes that the CG environment in the final model "manages" to exert precisely the same influence as the original atomistic environment (which is of course the aim of the coarse graining of nonbonded interaction but may be practically difficult, see below). One example where this is problematic are biomolecules in water where the solvent has a particularly important influence on the conformational equilibrium sampled by the molecule[68,69]. Nevertheless one often tries to keep bonded and nonbonded interactions as separate as possible, even if the above mentioned clean separation (without any double counting) is impossible, aiming at a certain modularity and transferability[30]. Transferability in the context of bonded interactions refers to the possibility to reuse potentials which have been determined for shorter fragments in longer chains (or in chains with a different overall sequence in the case of polypeptides[30]). Modularity refers to the possibility to (re)use bonded interaction

functions with different types of nonbonded CG models (from different coarse graining methodologies) with no or very little reparametrization. Bond, angle, and torsion potentials obtained via Boltzmann inversion (see below) can for example also be used in combination with nonbonded interaction functions determined via force matching[70,71]. One advantage of keeping bonded and nonbonded interactions as separate as possible is particularly eminent for biological macromolecules where correlations along the polymer chain (in the case of proteins most obviously demonstrated in the Ramachandran plot) play a decisive role. Capturing them in the CG model often requires special intramolecular interaction functions. With a separation into bonded and nonbonded interactions, studies regarding these aspects of conformational sampling are are not limited to one type of nonbonded interactions (e.g. structure-based coarse-graining) approaches but also hold for approaches where nonbonded interactions are parameterized differently, e.g. based on force matching or thermodynamic data.

In the following, I will first introduce one possibility to determine bonded interaction potentials and illustrate several aspects of conformational sampling with the help of an example, a short peptide in aqueous solution. After that I will introduce different approaches to determine nonbonded interactions.

2.1 Bonded Interactions, Conformational Sampling

In coarse graining methods where the parameterization is based on atomistic reference simulations, one first maps the atomistically sampled conformations to CG coordinates. From the latter, one obtains reference distributions (in CG degrees of freedom), for intramolecular interactions these are typically bond, angle and dihedral distributions. In structure-based approaches these distributions are Boltzmann inverted to obtain the corresponding potentials of mean force:

$$\begin{aligned} U^{CG}(r,T) &= -k_B T \ln(P(r,T)/r^2) + const_r \\ U^{CG}(\theta,T) &= -k_B T \ln(P(\theta,T)/\sin\theta) + const_\theta \\ U^{CG}(\varphi,T) &= -k_B T \ln(P(\varphi,T)) + const_\varphi \end{aligned} \tag{10}$$

Whether these potentials of mean force can be directly used as (tabulated) CG interaction potentials depends on several conditions. One condition is that the degrees of freedom are uncorrelated, i.e. the probability distribution describing the conformations factorizes into the corresponding contributions:

$$P(r,\theta,\varphi,T) = P(r,T)P(\theta,T)P(\varphi,T) \tag{11}$$

How well this assumption holds may (for a given molecule) depend on the choice of the CG mapping scheme, as was for example nicely shown for the case of two CG polystyrene models[72,73]. The assumption of uncorrelated DOFs is particularly problematic for biological systems with distinct secondary structures. Here, certain correlations between intramolecular degrees of freedom are characteristic for the conformations adopted by the molecule and need to be accounted for in the CG model[74–77,51,78,19]. A second condition is that bonded and nonbonded interactions can be separated according to Eq. 9. If there is a coupling of bonded and nonbonded interactions that cannot be prevented by simple avoiding of double counting of interactions, then it is very likely that the PMFs in Eq. 10 cannot be directly used as interaction potentials.

For the above reasons, potentials obtained from Boltzmann inversion according to Eqs. 10 may not succeed at producing the correct conformational equilibrium of the peptide in the CG model (i.e. after combining all covalent potentials and nonbonded interactions including the solvent). In that case one can introduce an additional refining step, which is completely analogous to the iterative procedure commonly used for nonbonded interactions which will be in details discussed below. For example for an angular DOF, θ, the iterative refinement is done as follows:

$$U_{i+1}(\theta) = U_i(\theta) + k_B T \ln \left[\frac{P_i(\theta, T)}{P_{target}(\theta, T)} \right] \tag{12}$$

Here $P_{target}(\theta, T)$ is the reference angular distribution from atomistic simulation and $P_i(\theta, T)$ is the current distribution after the ith iteration.

Peptides in aqueous solution (in the present case short oligoalanine fragments) can serve as a good example for which one can illustrate both the effect of coupling between bonded and nonbonded interactions in CG models and the influence of correlated degrees of freedom. The inset of Fig. 1 shows the investigated system, a capped ALA_3 peptide which has in the CG description the form of a linear chain of seven beads of two types, one for the peptide group (PEP, brown transparent beads in Fig. 1) and one representing the α and β carbon atoms (CAB, blue transparent beads in Fig. 1)[69]. For this system we developed two CG models, one with an implicit and one with an explicit CG solvent description, where each solvent particle was represented by a single CG particle. For both models, bond, angle and torsion potentials were at first obtained directly from Boltzmann inversion (Eq. 10).

In the case of the implicit solvent description, the CG peptide with these potentials reproduced all atomistic reference bond, angle and torsion distributions. In contrast, this was not the case in presence of explicit CG solvent. Here, the angles centered at the three CAB beads were severely distorted compared to the reference. The observed discrepancy could be corrected by an iterative refinement of the corresponding CG angle potential (Eq. 12). It was observed that the iterative correction to the angle potential did not have any negative effect on the sampling of other CG degrees of freedom such as bonds or other torsions. The so obtained explicit solvent CG model for ALA_3 was now also capable of reproducing all local conformational properties, i.e. all bond, angle and torsion distributions, just as the implicit solvent CG model. However, other (a little less local) properties not used for parametrization such as the distance distribution between two peptide groups separated by five CG bonds or the angle distribution between three consecutive peptide groups were not reproduced anymore. The distribution of this 1,3,5 PEP-PEP-PEP angle potential is shown in the left panel of Fig. 1. For longer chains this results in a non negligible discrepancy in the end to end distance distribution of the peptide, making the initial models rather unsatisfactory. This can be seen in the right panel of Fig 1: compare the atomistic reference (black solid line) to the end to end distances sampled by the implicit solvent CG model (blue dotted line) and the explicit solvent CG model with iteratively corrected CAB angle potentials (blue crosses).

To overcome this problem, additional CG potentials (1,5 PEP-PEP distance or 1,3,5 PEP-PEP-PEP angle potentials) along the peptide backbone were needed. This can be explained by the fact that the different regions in the 1,3,5 PEP-PEP-PEP angle distribution can be linked to typical secondary structure elements sampled by the polypeptide chain, namely to α-helical and β-strand chain segments. α-helical chain segments correspond

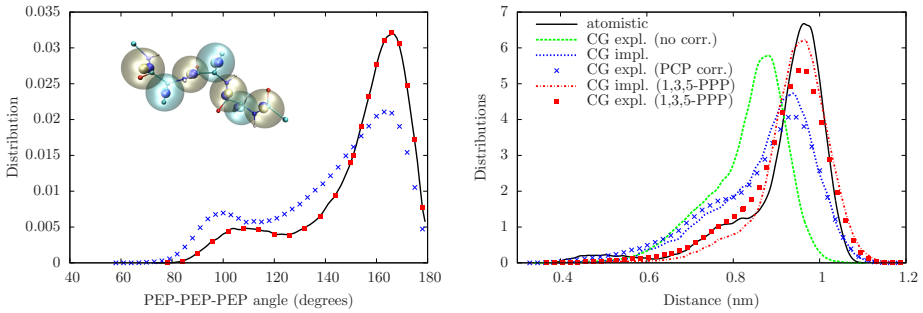


Figure 1. Conformational properties in a CG model of Ala_3 (Inset: molecular structure and CG mapping. Small beads: atomistic united atom representation; large transparent beads: CG model with PEP (brown) and CAB (blue) bead types).
Left panel: Angle distributions between three consecutive peptide groups (1,3,5 PEP-PEP-PEP angle) with different CG models. Atomistic reference: black solid line; explicit-solvent CG model with iterated PEP-CAB-PEP potential: blue crosses; explicit-solvent CG model with additional 1,3,5 PEP-PEP-PEP angle potential (iterative Boltzmann inversion): red squares.
Right panel: End-to-end distance distributions sampled with different CG models. Atomistic reference: black solid line; explicit-solvent CG model with bond, angle, torsion potentials from Boltzmann inversion without iterative refinement: green dashed line; implicit-solvent CG model with bond, angle, torsion potentials from Boltzmann inversion: blue dotted line; explicit-solvent CG model with iterated PEP-CAB-PEP potential: blue crosses; implicit-solvent CG model with additional 1,3,5 PEP-PEP-PEP angle potential (Boltzmann inversion, no iterative refinement): red dot dashed line; explicit-solvent CG model with additional 1,3,5 PEP-PEP-PEP angle potential (iterative Boltzmann inversion): red squares. The colors indicate the effect of the different refinement steps. Green: model with chain conformations where coupling between bonded and nonbonded (solvent) interactions causes deviations in local properties (PEP-CAB-PEP angle). Blue: models with correct bond, angle, and torsion distributions, but where conformations involving a 1,3,5 PEP-PEP-PEP segment deviate. Red: models where local chain conformations up to 1,3,5 PEP-PEP-PEP segments are correct.

to a 1,3,5 PEP-PEP-PEP angle around 100 degrees (i.e. corresponding to the shoulder in the distribution in Fig. 1) while β-strand chain segments correspond to the main peak around 170 degrees. This is not unexpected since the 1,3,5 PEP-PEP-PEP angle covers two CAB groups, i.e. adjacent sets of Ramachandran angles which should be correlated in secondary structure elements. Since secondary structure formation is intimately linked with intrachain correlations, most importantly between different backbone dihedral angles, the factorization into independent bond angles and torsions (Eq. 11) which is assumed in the Boltzmann inversion procedure (Eq. 10) does not hold for these biomolecular systems. For the chosen mapping scheme, correct backbone conformational sampling could be achieved by imposing a 1,3,5 PEP-PEP-PEP angle potential between consecutive peptide units along the backbone, and the resulting CG model was well capable to distinguish α-helical and β-structural chain segments and reproduce these conformations[69] (red squares in Fig 1). It should be noted that similar potential terms that account for correlations along the backbone and propensities for the formation of secondary structure elements have been discussed for other CG models, an excellent review of different approaches can be found in Ref. 19.

The right panel of Fig 1 shows the end-to-end distance distribution for all the explicit solvent and the implicit solvent CG models in comparison with the atomistic reference (black solid line) and summarizes the effects of the different refinement steps on the con-

formations of the Ala_3 chain: in the case of the CG model with explicit solvent and bond, angle and torsion potentials from direct Boltzmann inversion we had observed coupling between bonded and nonbonded (solvent) interactions which causes deviations in local properties, in this case the PEP-CAB-PEP angle. This leads to a large shift of the end-to-end distribution towards more compact structures (green dashed line). This artifact can be overcome by iterative refinement of the PEP-CAB-PEP angle. The resulting explicit-solvent model shows an end-to-end distance distribution which is very similar to the implicit solvent CG model obtained from Boltzmann inversion without iteration (where by construction coupling between solvent and angle cannot occur). These are distributions (blue dotted line and blue crosses) characteristic for models with correct individual bond, angle, and torsion distributions, but distortions in conformations involving 1,3,5 PEP-PEP-PEP segments causing the observed shift towards shorter end-to-end distances compared to the atomistic reference. Finally, the line and red symbols belong to end-to-end distance distributions characteristic for models where local chain conformations up to 1,3,5 PEP-PEP-PEP segments are correct (both for the explicit and the implicit solvent case, for implicit solvent, a 1,3,5 PEP-PEP-PEP has been determined based on Boltzmann inversion of the corresponding angle distribution from the atomistic reference, for the explicit solvent case, again an additional iterative refinement of this potential was needed according to Eq. 12). These distributions agree very well with the atomistic reference, which points out that the local properties of 1,3,5 PEP-PEP-PEP segments are crucial for a CG model to be able to describe and distinguish secondary structure propensities of a peptide chain. (Note that these are still very local properties which are not related to hydrogen bonding effects that drive for example the formation of an α-helix.)

This oligoalanine example shows that it is advantageous to keep the parametrization of bonded and nonbonded interactions separate, since in an "all-in-one iteration" procedure such a physically very important effect might have been overlooked. Secondly, the principal need for these potentials (or something analogous to account for conformational correlations) is independent form the methodology with which nonbonded interactions are determined. This means a bonded-interactions study provides the insight, which potentials are absolutely required to ensure correct sampling of chain conformations (especially the "special" potentials such as the1,3,5 angle), which ones are coupled to the nonbonded interactions, and which ones are coupled to each other. (This knowledge also suggests a certain sequence in the parametrization procedure.) Note, that while clearly not all CG procedures use (iterative) Boltzmann inversion for the bonded interactions, many of them do, even if they use other parametrization strategies for the nonbonded ones[70,71].

2.2 Parameterizing Nonbonded Interactions

As already mentioned, there are different approaches towards nonbonded interactions in systematic coarse graining, mostly divided into parametrizations based on thermodynamic targets and structure-motivated approaches that aim at reproducing the atomistic configuration space sampling (Eq. 4). As shown above, in these (in the widest sense) structure-based approaches CG interaction functions are parameterized so that they approximate the many-body PMF (Eq. 6) from the atomistic sampling (Eq. 8). The various structural approaches differ in the type of target functions they derive from the atomistic system. I will review two of them here: one, where one parameterizes against simpler structure functions (for

example pair correlation functions, i.e. pair potentials of mean force) instead of the many-body PMF and a second one, where the CG model is parameterized so that it reproduces (approximates) the mean forces on the CG sites (i.e. it uses derivatives of Eq. 6).

2.2.1 Structure-Based Coarse Graining: Iterative Boltzmann Inversion

"Traditional" structure-based methods provide CG interactions that reproduce predefined target structure properties – often a set of radial distribution functions[21,45,46,22,47–49,29,50–54]. This means that the many-body PMF (Eq. 6 and 7) is replaced as a target by a set of simpler structural correlation functions. If the interactions in the CG model are statistically independent or only weakly coupled then direct Boltzmann inversion determines each term in the potential immediately from the corresponding distribution function[79,80,21,81] (as we have seen already above for bonded interaction terms). For non-bonded interactions in dense systems this is typically not the case. This means that the individual distribution functions and their corresponding potentials of mean force (e.g. a radial distribution function of a simple liquid $g_{target}(r)$ and is Boltzmann inverse, the pair PMF, $V_0^{CG}(r) = -k_BT \ln g_{target}(r)$) cannot be directly used as interaction function since they correspond not only to the interaction potential but also the correlated contributions from the surroundings. These (multibody) effects of the environment need to be removed from the PMF to generate an effective pair potential that reproduces the target structure (for example the pair correlation function in the liquid) in an analogous manner as in Eq. 8. It can be shown that such a pair potential is unique (up to an arbitrary constant)[82] and exists[83,84,60,85]. There are several numerical methods to generate this pair potential (tabulated interaction function).

Iterative Boltzmann inversion (IBI)[86,87,47] is a natural extension of the Boltzmann inversion method. Here, a numerical CG potential is iteratively refined until the target structure is reproduced within a predefined error. Each step in the iteration procedure is a CG simulation with potential $V_i^{CG}(r)$ which yields an RDF $g_i(r)$ that differs from the target $g_{target}(r)$. The potential is then modified by a correction term $\Delta V(r)$ according to

$$V_{i+1}^{CG}(r) = V_i^{CG}(r) + \Delta V(r) = V_i^{CG}(r) + k_BT \ln \frac{g_i(r)}{g_{target}(r)} \tag{13}$$

Sometimes the potential correction $\Delta V(r)$ is multiplied with a prefactor $0 < \lambda \leq 1$ to avoid overshooting in the numerical procedure. The iterative procedure is initiated often with the (pair) potential of mean force $V_0^{CG}(r) = -k_BT \ln g_{target}(r)$, but that is not mandatory, different starting potentials might be useful, in particular for more complex mixed systems, where the iterative procedure may be unstable, because intermediate CG models for example lead to phase separation. An illustration of the IBI method and the typical types of potentials that are obtained can be found in Fig. 2 (left and middle panel).

IBI is by no means the only numerical method that solves the above task. Another numerical scheme is the so called inverse Monte Carlo (or more recently renamed Newton inversion) method[45,46,49,29] which should according to Henderson's theorem lead to the same numerical solution for the pair potential corresponding to a given pair correlation function. While in IBI the potential update ΔV is ad hoc, it is computed in IMC using rigorous statistical mechanical arguments. For details see Ref. 45. In the case of multi-component systems, where several pair potentials need to be updated, IMC accounts for

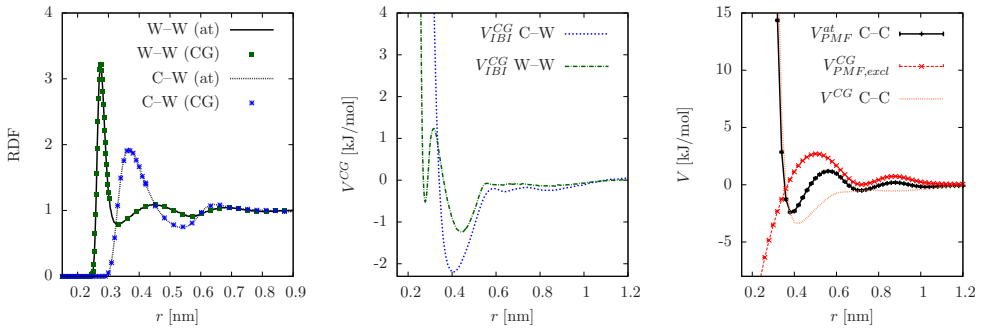


Figure 2. Illustration of structure-based coarse graining that aim at reproducing pair PMFs from atomistic target. Example system: methane (denoted as C) in water (denoted as W), both mapped to a single site.
left panel: RDFs of atomistic and CG system
middle panel: Tabulated potentials obtained from IBI
right panel: Illustration of the subtraction procedure to determine CG solute-solute (here C-C) potentials. Black line: atomistic target potential of mean force between two methane molecules in aqueous solution. Red dashed line: CG potential of mean force with excluded direct interactions between the two solute particles (note that the solute-solute exclusion leads to the missing short range repulsion in the PMF). Red solid line: resulting solute-solute potential according to Eq. 16.

correlations between observables, i.e. the updates for the different potentials are interdependent. In contrast, for IBI, each potential is updated independently, which might lead to oscillations and convergence problems in the iteration procedure. The disadvantage of IMC on the other hand is a high computational cost and problems with numerical stability, for a detailed comparison see Ref. 88. Related to IMC there are several other recent developments, e.g. a molecular renormalization group approach[50–52] or an approach that relies on relative entropies[60–62].

While the above structure-based methods by construction *exactly* (within the error of the numerical procedure) reproduce the local pair structures and thus are well-suited to reinsert atomistic coordinates, it is not a priori clear whether they are equally well suited to reproduce thermodynamic properties (pressure, phase behavior, etc.) of the reference system. Note also that CG models based on pair correlation functions do not necessarily reproduce higher-order (e.g. three-body) structural correlations[88] since the pair correlation functions as structural targets are just an approximation to the total conformational distribution function obtained from the atomistic sampling, $P^{at}(\mathbf{R})$ (Eq. 4). This means that if higher order correlations are a crucial part of the many-body PMF, models based on pair structures may fail to represent these, and it may even be possible that models which are limited to pair potentials may fail to reproduce these correlations irrespective of the parametrization methodology. One example where this is studied in details is the example of liquid water[89,88,65]. Recently Noid and coworkers have analyzed these aspects in details using concepts from liquid state theory[64].

One more note concerning Henderson's theorem: even though there is in principle one *exact* solution for the effective pair potential that reproduces a given pair correlation function, different potentials might give a reasonably close representation of the structure, i.e. the above inverse problem is mathematically ill-posed[88,90]. This effect becomes even

more pronounced in complex systems where several interaction functions corresponding to several RDFs need to be numerically determined. This ill-posedness can to some extent be turned into an advantage since it allows to impose thermodynamic constraints in the parametrization procedure. This will result in interaction functions which do *not exactly* reproduce the target structure but give a very close representation while at the same time produce a desired thermodynamic behavior. One example for this are pressure correction terms[47,89]. Here, an additional linear pressure correction is applied during the iterative Boltzmann inversion procedure with

$$\Delta V_{i,P}^{CG}(r) = A_i \left(1 - \frac{r}{r_{cut}}\right) \tag{14}$$

where r_{cut} is the radial cutoff distance of the nonbonded interaction and the constant A is determined via the virial expression for the pressure to

$$-\left[\frac{2\pi N\rho}{3r_{cut}} \int_0^{r_{cut}} r^3 g_i(r) \mathrm{d}r\right] A_i \approx (P_i - P_{target})\, V \tag{15}$$

V is the volume of the system, P_i the pressure of the CG model in the i-th iteration, and P_{target} the target pressure. Details can be found in Ref. 89.

It is to be expected that there will be more development in this direction (using other types of thermodynamic constraints) since in particular for complex soft matter system the balancing of structural and thermodynamic behavior in CG models is an ongoing field of research[53,54].

2.2.2 Extension to Dilute Solute/Solvent Systems

In mixed systems where one component is very dilute (from now on termed solute), e.g. biomolecules in aqueous solution, iterative Boltzmann inversion and similar methods are problematic. While one can easily compute the solvent-solvent and the solute-solvent radial distribution functions, and therefore determine the corresponding CG potentials with for example IBI, this is not so straightforward for the interactions between the low concentration component (solutes). (Note that for simplicity I will here only discuss solutes that are represented by a single CG bead.) Here, obtaining the PMF through brute force sampling of a radial distribution function is not advisable. In that case one should compute the solute-solute pair PMF (between two solute particles) with an advanced sampling method such as umbrella sampling or thermodynamic integration (using distance constraints)[91,92].

When solvent degrees of freedom are not explicitly present in the CG system, this solute-solute PMF can be used directly as effective solute-solute nonbonded interaction since the environmental (solvent) effects within the PMF are not explicitly represented through solvent degrees of freedom in the CG model.This direct use of the PMF has for example been employed for an implicit solvent model of aqueous electrolyte solution, i.e. implicit solvent ion models[46,93,94,50,95]. Also for other other types of solutes the solute-solute PMF has been used as interaction potential in implicit solvent models[68,96].

However, if solvent, for example water, explicitly exists in the CG system, new effective solute-solute nonbonded pair interactions are needed from which the solvent contributions are removed in the same way they are removed by IBI in other systems. However, due to the sampling problem of the PMF between the solute (dilute) component, an iterative

procedure is prohibitive for the solute-solute interactions. To solve this problem, an approximate method has been developed by Villa et al[97,98]. Here, the CG solvent-solvent and solute-solvent interactions are first determined, for example through normal IBI. Now the pair PMF between the solutes $V_{PMF}^{at}(r)$ is computed (from atomistic umbrella sampling or thermodynamic integration) and used as a target, in other words the resulting CG model is parameterized to reproduce the solute-solute association strength observed in the atomistic system. In order to remove the solvent contribution from $V_{PMF}^{at}(r)$, a subtraction procedure is employed. One conducts a separate PMF calculation (again with umbrella sampling or thermodynamic integration), this time in a CG system, where the (previously determined) CG solvent-solvent and solute-solvent interactions are present but no direct interaction between the solute particles is turned on. The resulting PMF $V_{PMF,excl}^{CG}(r)$ *only* consists of the environmental contributions (in the CG environment). By subtracting $V_{PMF,excl}^{CG}(r)$ from the target PMF one obtains the missing direct pair interaction

$$V^{CG}(r) = V_{PMF}^{at}(r) - V_{PMF,excl}^{CG}(r) \tag{16}$$

which by construction reproduces the target PMF. The method is illustrated in Fig. 2. The left and middle panels show parametrization of the solvent-solvent interactions by IBI, and the right panel shows the different contributions in Eq. 16. Properties of the resulting solute/solvent systems will be discussed below in the context of transferability[98].

Note, that this subtraction procedure is not necessarily limited to CG solvent-solvent or solute-solvent interactions determined by IBI. In principle also other types of CG solvent-solvent or solute-solvent interactions could be used to determine $V_{PMF,excl}^{CG}(r)$. If one then applies Eq. 16, one obtains an effective solute-solute interaction $V^{CG}(r)$ which reproduces the atomistically observed solute-solute association strength (i.e. $V_{PMF}^{at}(r)$) in the particular CG solvent that was chosen.

2.2.3 Force Matching and Related Methods

An alternative method to construct CG potentials from an atomistic reference sampling is force matching (also termed MS-CG/multiscale coarse graining). This method has been successfully applied to a multitude of biomolecular and other soft matter systems[55,23,56–59]. Here, the CG forcefield is determined such that the difference between the (instantaneous) CG forces and the forces in the underlying atomistic system is minimized. Thus, force matching uses a variational (i.e. non-iterative) approach for constructing the CG potential based on the atomistic reference simulation (in this case the recorded forces from the atomistic simulation). The numerical implementation of this variational principle works in such a way that the exact many-body PMF (Eq. 6) is represented by a linear combination of basis functions that are functions of the CG site coordinates. This means the CG force field depends on M parameters $g_1, \cdots, g_M$. These parameters can be prefactors of analytical functions, tabulated values of the interaction potentials, or coefficients of splines used to describe these potentials. These M parameters are optimized so that the CG model reproduces the forces in the atomistic system (after mapping) as closely as possible. To this end, the (CG) reference forces on the N CG sites obtained from the atomistic system are computed by properly reweighting the forces on the atoms (i.e. applying the mapping scheme on the level of forces). By doing this on L snapshots from the atomistic simulation one gets $N \times L$ reference forces. Now, one optimizes the M parameters of the CG

force field in such a way that the deviation of the CG forces from these reference forces is minimized. This means one optimizes the parameters $g_1, \cdots, g_M$ in the following $N \times L$ equations:

$$f_{il}^{CG}(g_1, \cdots g_M) = f_{il}^{ref} \tag{17}$$

Noid et. al have shown that the coarse-grained CG potentials obtained from force matching are an approximate variational solution for the exact many-body potential of mean force for the coarse-grained sites (Eq. 6) and have thus established the link between force matching and various structure-based methods[99,63,100]. It should be noted though, that also in the case of force matching the CG force field is an approximation to the high dimensional PMF within the limitations of the types of CG forces chosen (for example pair forces that can be either derived from analytical or from numerical tabulated potentials). This also implies that a CG model obtained from force matching does not by construction reproduce the pair correlation functions in the system, and the reproduction of local structural properties such as pair distributions may (or may not) be rather weak. An exact reproduction of the underlying atomistic problem by force matching potentially requires the introduction of higher order (e.g. three-body) interactions. Recently, Noid and coworkers have extended the force matching method and demonstrated that the CG force field can be directly determined from structural correlation functions obtained from the atomistic system instead of the forces[63]. Their theoretical approach also allows an assessment of the correlations between different interactions that are neglected by straightforward Boltzmann inversion and allows to quantify the importance of many-body correlations in CG models.

3 Systematic Coarse Graining: Challenges

From the preceding sections we have seen that there are different approaches to systematically parameterize CG models which by construction will not be equally well suited to reproduce thermodynamic and structural properties of the system. It is not a priori clear whether structure-based potentials reproduce macroscopic thermodynamic properties and, vice versa, if thermodynamics-based potentials reproduce microscopic structural properties. Yet, the interplay of structure and thermodynamics is crucial for the investigation of structure formation processes, in particular for biomolecular aggregation in aqueous solution where partitioning and phase separation play a decisive role. All CG models (in fact also all classical atomistic forcefields) are state-point dependent and cannot necessarily be – without reparametrization – transferred to different thermodynamic conditions or a different chemical environment compared to the one where they had been derived. This means 'transferability' can refer to a change in temperature, density, concentration, system composition, phase, etc., but also a change in chemical environment, e.g. the change of length or sequence of an aminoacid chain. Structure-motivated CG models which approximate the high dimensional PMF obtained from an atomistic reference are by construction heavily state point dependent, and several studies have addressed questions regarding their ability to reproduce thermodynamic properties. One system that has been of particular interest in this context is liquid water[84,89,101]. The reason is on the one hand of course its immense importance in all questions regarding biomolecular systems. In addition, it is of particular methodological interest because for single bead models of water it is known that three-body correlations play a decisive role and the potential compromise between reproducing

pair- or higher order structural correlations is particularly relevant for the properties of the model[89,88,65]. Different studies have been carried out that compare structure-motivated and thermodynamics-based CG models[102,103,90]. While CG models where the parametrization targets had been solvation and partitioning properties are particularly well suited to reproduce processes where for example hydrophilicity/hydrophobicity arguments play a decisive role, they do not per se reproduce the structure of the system[102,90]. Related to their ability to reproduce the thermodynamic properties of certain chemical units, these models exhibit considerable transferability and can often be applied to a variety of molecular systems and a range of thermodynamic conditions. Motivated by these observations, intensive research is currently carried out to derive CG potentials that are both thermodynamically as well as structurally consistent with the underlying higher-resolution description, thus ensuring for example state point transferability[104,98,58,53,54]. Another current line of research is related to the approach to determine CG potentials for dilute components in solute/solvent systems (see above), where pair potentials of mean force on atomistic and CG level are subtracted in order to reproduce the solute-solute association behavior[68,98]. Here, new coarse graining approaches are developed which rely on a thermodynamic cycle to obtain effective CG pair potentials[105,90].

3.1 Transferability of CG Models

As discussed in the previous section, the balance between thermodynamics and structure is intimately related to the question of transferability. In particular, binary mixtures have been widely used as model systems to explore various aspects of the transferability of CG models for biomolecular systems[106–108,93–95,39,104,98]. The transferability to different concentrations of liquid mixtures or solutions is of vital importance for simulation of processes such as (bio)molecular aggregation which are characterized by spatially varying structure and fluctuating concentrations.

It has been mentioned before that effective pair potentials account for multibody effects, for example, three body interactions. For this reason, they are only to a limited extent additive, which limits the transferability of the potentials[106,98]. Understanding the physical nature of non-additivity in the system of interest can help to make a CG model transferable. In principle, there are various possibilities to approach the question of transferability of effective pair potentials: (i) One applies a model derived at/optimized for a given state point unaltered to a range of state points "nearby"; in that case, one has to carefully investigate the range in which this is permitted[109,67]. (ii) One creates a new set of potentials for each state point one wants to investigate[109]. (iii) One specifically designs a single CG model with the aim to be transferable (for example specific density dependent potentials[107,108,58], CG models that are designed to be applicable for a range of mixture compositions[39,104], or CG models that are capable of capturing a liquid crystalline phase transition[53,54]). (iv) One uses a model derived at one state point and (analytically) modifies it to be applicable to different conditions (one example being the rescaling of potentials in order to apply them to a different temperature[110]).

In the following, I will – in a little more detail – discuss examples for two above scenarios. These examples illustrate that understanding the physical basis behind the limitations in transferability can help to design transferable models.

Recently, Villa et al. proposed a CG model for mixed systems of benzene in water[98].

The CG benzene-benzene potential had been parameterized on the basis of the benzene-benzene PMF of two benzene molecules in aqueous solution, i.e., at "infinite" dilution (as described above). Benzene-water mixtures of different composition have been studied with this CG model and analyzed using Kirkwood-Buff theory of solutions[111]. Kirkwood-Buff theory provides a link between local structural information and thermodynamic properties of the solution. The CG model, parametrized at infinite dilution of benzene, reproduces the Kirkwood-Buff integrals of mixtures at various concentrations obtained with the detailed-atomistic model. It was found that this CG model can reproduce the changes in the benzene chemical potential and the activity coefficients of the mixtures over a range of mixture compositions (up to concentrations where benzene and water demix in the atomistic reference simulation). This is shown in the left panel of Fig. 3. A possible explanation is that hydrophobic interactions between benzene solutes are short-ranged, and the multibody correlations involved in hydrophobic association can be described by pairwise additive effective potentials (category (i) of the above list). The observed transferability of the potential supports the idea that hydrophobic interactions between small molecules are pairwise additive. Villa et al. also found that a different CG model for benzene-benzene interactions that had been derived for pure benzene (via IBI) is neither suited to describe benzene-benzene interactions in aqueous solution at different concentrations nor a phase-separated benzene/water system with a bulk benzene layer[98] (see also Fig. 3). This example illustrates the necessity of a careful choice of reference state points.

In the second example, the situation is different. Here, the transferability of CG (in this case implicit-solvent) ion models in aqueous solution has been investigated. Due to long-range electrostatic interactions, the ions affect the behavior of water increasingly with increasing ion concentration. More specifically, the presence of many ions reduces the orientational fluctuations of the water molecules and thus the dielectric permittivity of the solvent. Therefore, effective ion-ion potentials parametrized at infinite dilution are not directly transferable to higher salt concentrations. Hess et al. developed a reduced-resolution (in this case implicit-solvent) potential for aqueous electrolyte solutions where an ion-concentration-dependent Coulomb term was added to the (ion-specific) pair interaction. Thus, by using a concentration-dependent dielectric permittivity of water, part of the multibody effects in the system were accounted for in the ion-ion pairwise interaction in the implicit solvent model[93,94]. This approach reproduced the NaCl solution osmotic properties and the ion coordination up to a concentration of 2.8 M (mol/L). While in the case of the CG model of benzene/water mixtures[98] the short-range hydrophobic interactions parameterized at infinite dilution were directly transferable to higher benzene concentrations, the ion-ion interactions determined at infinite dilution had to be split into a short ranged ion-specific and a long-range electrostatic part. The interactions were then made transferable by keeping the short-ranged part constant and analytically modifying the long-ranged electrostatic part (category (iv) of the above list). Shen et al. have further investigated the structure and osmotic properties of electrolyte solutions over a wide range of concentrations[95]. Using a concentration-dependent dielectric constant one obtains also very good structural properties of the electrolyte solution at low and intermediate salt concentrations while for larger salt concentrations multibody ion-ion correlations put a limit to straightforward transferability (see right panel of Fig. 3). Guided by this structural analysis the transferability of the implicit-solvent model could be improved also for high ion concentrations. One obtains transferable implicit-solvent effective pair potentials which are

both structurally and thermodynamically well consistent with an explicit solvent reference model.

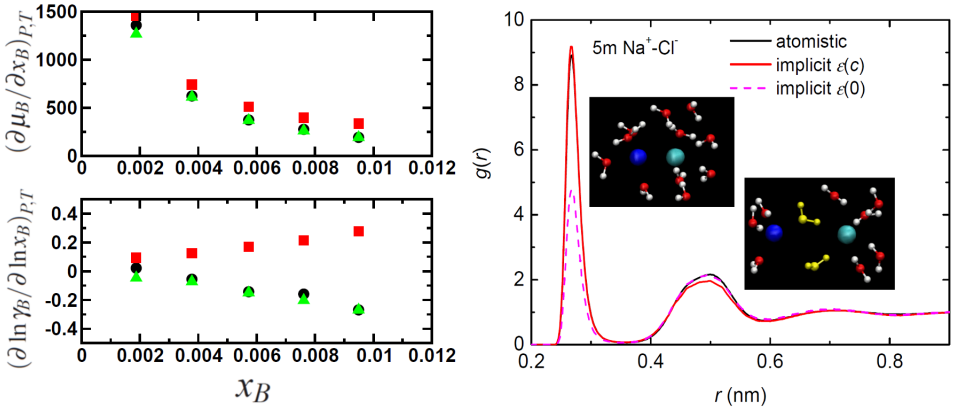


Figure 3. Transferability of coarse grained (CG) nonbonded interaction potentials
Left panels: Hydrophobic molecules in aqueous solution[98]: Derivative of chemical potential (upper panel) and activity coefficient (lower panel) as a function of the benzene mole fraction in benzene/water mixtures from atomistic simulations (black symbols) and two types of CG models, one derived from a benzene/water system at infinite dilution (green) and one from bulk benzene (red).
Right panel: Electrolyte solutions[95]: Radial distribution function (RDF) of Na-Cl (5m concentration) in atomistic/explicit solvent (black line), and CG/implicit solvent simulations. CG simulation without concentration dependent dielectric constant: pink dashed line; CG simulation made transferable with concentration dependent dielectric constant: red line. The snapshots indicate typical structures in explicit solvent corresponding to the first two peaks in the RDF (first peak: contact ion pair; second peak: solvent shared ion pair).

Acknowledgments

I would like to thank all my collaborators and the members of the multiscale modeling group at the Max Planck Institute for Polymer Research for many stimulating discussions about how to systematically develop CG models, in particular Nico van der Vegt and Kurt Kremer for many fruitful collaborations. I am also extremely grateful to all the developers of the VOTCA package for initiating and providing a wonderful new platform to use, test and develop a large variety of coarse graining methods[88]. I also want to thank Olga Bezkorovaynaya for providing figures for this manuscript and Biswaroop Mukherjee for carefully reading it. Financial support from the German Science Foundation within the Emmy Noether program is gratefully acknowledged.

References

1. Gregory A Voth, (Ed.), *Coarse-graining of condensed phase and biomolecular systems*, CRC Press, Boca Raton, FL, 2009.

2. D. Fritz, K. Koschke, V. A. Harmandaris, Nico F. A. van der Vegt, and K. Kremer, *Multiscale modeling of soft matter: scaling of dynamics*, Phys. Chem. Chem. Phys., **13**, 10412–10420, 2011.
3. C.F. Lopez, S.O. Nielsen, P.B. Moore, J.C. Shelley, and M.L. Klein, *Self-assembly of a phospholipid Langmuir monolayer using a coarse-grained molecular dynamics simulations.*, J. Phys.: Condens. Matter, **14**, 9431–9444, 2002.
4. I. R. Cooke, K. Kremer, and M. Deserno, *Tunable generic model for fluid bilayer membranes*, Phys. Rev. E, **72**, no. 1, 011506, 2005.
5. M. Müller, K. Katsov, and M. Schick, *Biological and synthetic membranes: What can be learned from a coarse-grained description?*, Physics Reports, **434**, 113 – 176, 2006.
6. B. J. Reynwar, G. Illya, V. A. Harmandaris, M. M. Müller, K. Kremer, and M. Deserno, *Aggregation and vesiculation of membrane proteins by curvature-mediated interactions*, Nature, **447**, no. 7143, 461 – 464, 2007.
7. M. L. Klein and W. Shinoda, *Large-scale molecular dynamics simulations of self-assembling systems*, Science, **321**, 798–800, 2008.
8. N. Go, *Theoretical-studies of protein folding*, Annu. Rev. Biophys. Bioeng., **12**, 183 – 210, 1983.
9. D. Thirumalai and D. K. Klimov, *Deciphering the timescales and mechanisms of protein folding using minimal off-lattice models*, Curr. Opin. Struct. Biol., **9**, no. 2, 197 – 207, 1999.
10. A. Liwo, P. Arlukowicz, C. Czaplewski, S. Oldziej, J. Pillardy, and H. A. Scheraga, *A method for optimizing potential-energy functions by a hierarchical design of the potential-energy landscape: Application to the UNRES force field*, Proc. Natl. Acad. Sci. USA, **99**, no. 4, 1937 – 1942, 2002.
11. G. Favrin, A. Irback, and S. Wallin, *Folding of a small helical protein using hydrogen bonds and hydrophobicity forces*, Proteins, **47**, 99–105, 2002.
12. T. Head-Gordon and S. Brown, *Minimalist models for protein folding and design*, Curr. Opin. Struct. Biol., **13**, no. 2, 160 – 167, 2003.
13. H. D. Nguyen and C. K. Hall, *Molecular dynamics simulations of spontaneous fibril formation by random-coil peptides*, Proc. Natl. Acad. Sci. USA, **101**, no. 46, 16180 – 16185, 2004.
14. N. V. Buchete, J. E. Straub, and D. Thirumalai, *Development of novel statistical potentials for protein fold recognition*, Curr. Opin. Struct. Biol., **14**, no. 2, 225 – 232, 2004.
15. C. Clementi, *Coarse-grained models of protein folding: toy models or predictive tools?*, Curr. Opin. Struc. Biol., **18**, 10–15, 2008.
16. P. Derreumaux and N. Mousseau, *Coarse-grained protein molecular dynamics simulations*, J. Chem. Phys., **126**, 025101, 2007.
17. G. Bellesia and J. E. Shea, *Self-assembly of beta-sheet forming peptides into chiral fibrillar aggregates*, J. Chem. Phys., **126**, 245104, 2007.
18. T. Bereau and M. Deserno, *Generic coarse-grained model for protein folding and aggregation*, J. Chem. Phys., **130**, 235106, 2009.
19. V. Tozzini, *Minimalist models for proteins: a comparative analysis*, Q. Rev. Biophys., **43**, 333–371, 2010.
20. C. Wu and J.-E. Shea, *Coarse-grained models for protein aggregation*, Curr. Opin. Struc. Biol., **21**, 209–220, 2011.

21. W. Tschöp, K. Kremer, J. Batoulis, T. Burger, and O. Hahn, *Simulation of polymer melts. I. Coarse-graining procedure for polycarbonates*, Acta Polym., **49**, no. 2-3, 61 – 74, 1998.
22. F. Müller-Plathe, *Coarse-graining in polymer simulation: From the atomistic to the mesoscopic scale and back*, ChemPhysChem, **3**, 754 – 769, 2002.
23. G. S. Ayton, W. G. Noid, and G. A. Voth, *Multiscale modeling of biomolecular systems: in serial and in parallel*, Curr. Opin. Struct. Biol., **17**, 192 – 198, 2007.
24. P. L. Freddolino, A. Arkhipov, A. Y. Shih, Y. Yin, Z. Chen, and K. Schulten, "Application of residue-based and shape-based coarse graining to biomolecular simulations.", in: Coarse-Graining of Condensed Phase and Biomolecular Systems, G. A. Voth, (Ed.). Chapman and Hall/CRC Press, Taylor and Francis Group, 2008.
25. P. Sherwood, B. R. Brooks, and M. S. P. Sansom, *Multiscale methods for macromolecular simulations*, Curr. Opin. Struc. Biol., **18**, 630–640, 2008.
26. C. Peter and K. Kremer, *Multiscale simulation of soft matter systems - from the atomistic to the coarse-grained level and back*, Soft Matter, **5**, 4357–4366, 2009.
27. T. Murtola, A. Bunker, I. Vattulainen, M. Deserno, and M. Karttunen, *Multiscale modeling of emergent materials: biological and soft matter*, Phys. Chem. Chem. Phys., **11**, 1869–1892, 2009.
28. C. Peter and K. Kremer, *Multiscale simulation of soft matter systems*, Faraday Discuss, **144**, 9–24, 2010.
29. A. Lyubartsev, A. Mirzoev, L. J. Chen, and A. Laaksonen, *Systematic coarse-graining of molecular models by the Newton inversion method*, Faraday Discuss, **144**, 43–56, 2010.
30. O. Engin, A. Villa, C. Peter, and M. Sayar, *A Challenge for Peptide Coarse Graining: Transferability of Fragment-Based Models*, Macromol. Theory Simul., **20**, 451–465, 2011.
31. M. Praprotnik, L. Delle Site, and K. Kremer, *Adaptive resolution molecular-dynamics simulation: Changing the degrees of freedom on the fly*, J. Chem. Phys., **123**, no. 22, 224106, 2005.
32. M. Praprotnik, L. Delle Site, and K. Kremer, *Multiscale Simulation of Soft Matter: From Scale Bridging to Adaptive Resolution*, Annu. Rev. Phys. Chem., **59**, no. 1, 545–571, 2008.
33. S. O. Nielsen, C. F. Lopez, G. Srinivas, and M. L. Klein, *A coarse grain model for n-alkanes parameterized from surface tension data*, J. Chem. Phys., **119**, 7043–7049, 2003.
34. S. J. Marrink, A. H. deVries, and A. E. Mark, *Coarse Grained Model for Semiquantitative Lipid Simulations*, J. Phys. Chem. B, **108**, 750–760, 2004.
35. S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. de Vries, *The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations*, J. Phys. Chem. B, **111**, 7812–7824, 2007.
36. W. Shinoda, R. DeVane, and M. L. Klein, *Multi-property fitting and parameterization of a coarse grained model for aqueous surfactants*, Mol. Simulat., **33**, 27–36, 2007.
37. L. Monticelli, S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman, and S. J. Marrink, *The MARTINI Coarse-Grained Force Field: Extension to Proteins*, J. Chem. Theor. Comput., **4**, 819–834, 2008.

38. B. M. Mognetti, L. Yelash, P. Virnau, W. Paul, K. Binder, M. Mueller, and L. G. Macdowell, *Efficient prediction of thermodynamic properties of quadrupolar fluids from simulation of a coarse-grained model: The case of carbon dioxide*, J. Chem. Phys., **128**, 104501, 2008.
39. B. M. Mognetti, P. Virnau, L. Yelash, W. Paul, K. Binder, M. Müller, and L. G. Macdowell, *Coarse-grained models for fluids and their mixtures: Comparison of Monte Carlo studies of their phase behavior with perturbation theory and experiment*, J. Chem. Phys., **130**, 044101, 2009.
40. C. A. López, A. J. Rzepiela, A. H. de Vries, L. Dijkhuizen, P. H. Hünenberger, and S. J. Marrink, *Martini Coarse-Grained Force Field: Extension to Carbohydrates*, J. Chem. Theory Comput., **5**, 3195–3210, 2009.
41. R. DeVane, W. Shinoda, P. B. Moore, and M. L. Klein, *Transferable Coarse Grain Nonbonded Interaction Model for Amino Acids*, J. Chem. Theory Comput., **5**, 2115–2124, 2009.
42. R. DeVane, M. L. Klein, C.-C. Chiu, S. O. Nielsen, W. Shinoda, and P. B. Moore, *Coarse-Grained Potential Models for Phenyl-Based Molecules: I. Parametrization Using Experimental Data*, J. Phys. Chem. B, **114**, 6386–6393, 2010.
43. X. He, W. Shinoda, R. DeVane, and M. L. Klein, *Exploring the utility of coarse-grained water models for computational studies of interfacial systems*, Mol. Phys., **108**, 2007–2020, 2010.
44. S. O. Yesylevskyy, L. V. Schafer, D. Sengupta, and S. J. Marrink, *Polarizable Water Model for the Coarse-Grained MARTINI Force Field*, PLoS Comput. Biol., **6**, e1000810, 2010.
45. A. P. Lyubartsev and A. Laaksonen, *Calculation of effective interaction potentials from radial-distribution functions - a reverse Monte-Carlo approach*, Phys. Rev. E, **52**, 3730 – 3737, 1995.
46. A. P. Lyubartsev and A. Laaksonen, *Osmotic and activity coefficients from effective potentials for hydrated ions*, Phys. Rev. E, **55**, 5689–5696, 1997.
47. D. Reith, M. Putz, and F. Müller-Plathe, *Deriving effective mesoscale potentials from atomistic simulations*, J. Comp. Chem., **24**, 1624 – 1636, 2003.
48. C. Peter, L. Delle Site, and K. Kremer, *Classical simulations from the atomistic to the mesoscale: coarse graining an azobenzene liquid crystal*, Soft Matter, **4**, 859–869, 2008.
49. T. Murtola, M. Karttunen, and I. Vattulainen, *Systematic coarse graining from structure using internal states: Application to phospholipid/cholesterol bilayer*, J. Chem. Phys., **131**, 055101, 2009.
50. A. Savelyev and G. A. Papoian, *Molecular renormalization group coarse-graining of electrolyte solutions: application to aqueous NaCl and KCl*, J. Phys. Chem. B, **113**, 7785–7793, 2009.
51. A. Savelyev and G. A. Papoian, *Molecular Renormalization Group Coarse-Graining of Polymer Chains: Application to Double-Stranded DNA*, Biophys. J., **96**, 4044–4052, 2009.
52. A. Savelyev and G. A. Papoian, *Chemically accurate coarse graining of double-stranded DNA*, P. Natl. Acad. Sci., **107**, 20340–20345, 2010.
53. G. Megariotis, An Vyrkou, A. Leygue, and D. N. Theodorou, *Systematic Coarse Graining of 4-Cyano-4 '-pentylbiphenyl*, Ind. Eng. Chem. Res., **50**, 546–556, 2011.

54. B. Mukherje, Delle Site L., Kremer K., and C. Peter, *Derivation of a Coarse Grained model for Multiscale Simulation of Liquid Crystalline Phase Transitions*, J. Phys. Chem B submitted, 2012.
55. S. Izvekov and G. A. Voth, *A multiscale coarse-graining method for biomolecular systems*, J. Phys. Chem. B, **109**, 2469 – 2473, 2005.
56. J. Zhou, I. F. Thorpe, S. Izvekov, and G. A. Voth, *Coarse-grained peptide modeling using a systematic multiscale approach*, Biophys. J., **92**, 4289 – 4303, 2007.
57. R. D. Hills, L. Lu, and G. A. Voth, *Multiscale Coarse-Graining of the Protein Energy Landscape*, PLoS Comput. Biol., **6**, e1000827, 2010.
58. S. Izvekov, P. W. Chung, and B. M. Rice, *The multiscale coarse-graining method: Assessing its accuracy and introducing density dependent coarse-grain potentials*, J. Chem. Phys., **133**, 064109, 2010.
59. J. W. Mullinax and W. G. Noid, *Recovering physical potentials from a model protein databank*, P. Natl. Acad. Sci. Usa, **107**, 19867–19872, 2010.
60. M. S. Shell, *The relative entropy is fundamental to multiscale and inverse thermodynamic problems*, J. Chem. Phys., **129**, 144108, 2008.
61. A. Chaimovich and M. S. Shell, *Relative entropy as a universal metric for multiscale errors*, Phys. Rev. E, **81**, 060104, 2010.
62. A. Chaimovich and M. S. Shell, *Coarse-graining errors and numerical optimization using a relative entropy framework*, J .Chem. Phys., **134**, 094112, 2011.
63. J. W. Mullinax and W. G. Noid, *A Generalized-Yvon-Born-Green Theory for Determining Coarse-Grained Interaction Potentials*, J. Phys. Chem. C, **114**, 5661–5674, 2010.
64. C. R. Ellis, J. F. Rudzinski, and W. G. Noid, *Generalized-Yvon-Born-Green Model of Toluene*, Macromol. Theory Simul., **20**, 478–495, 2011.
65. L. Larini, L. Lu, and G. A. Voth, *The multiscale coarse-graining method. VI. Implementation of three-body coarse-grained potentials*, J. Chem. Phys., **132**, 164107, 2010.
66. C. F. Abrams and K. Kremer, *Combined coarse-grained and atomistic simulation of liquid bisphenol A-polycarbonate: Liquid packing and intramolecular structure*, Macromolecules, **36**, no. 1, 260 – 267, 2003.
67. D. Fritz, V. A. Harmandaris, K. Kremer, and N. F. A. van der Vegt, *Coarse-Grained Polymer Melts Based on Isolated Atomistic Chains: Simulation of Polystyrene of Different Tacticities*, Macromolecules, **42**, 7579–7588, 2009.
68. A. Villa, C. Peter, and N. F. A. van der Vegt, *Self-assembling dipeptides: conformational sampling in solvent-free coarse-grained simulation*, Phys. Chem. Chem. Phys., **11**, 2077–2086, 2009.
69. O. Bezkorovaynaya, A. Lukyanov, K. Kremer, and C. Peter, *Multiscale simulation of small peptides: Consistent conformational sampling in atomistic and coarse-grained models*, J. Comp. Chem, 2012, in press, DOI: 10.1002/jcc.22915.
70. Y. Wang, S. Izvekov, T. Yan, and G. A. Voth, *Multiscale coarse-graining of ionic liquids*, J. Phys. Chem. B, **110**, 3564–3575, 2006.
71. V. Rühle and C. Junghans, *Hybrid Approaches to Coarse Graining using the VOTCA Package: Liquid Hexane*, Macromol. Theory Simul., **20**, 472–477, 2011.
72. V. A. Harmandaris, N. P. Adhikari, N. F. A. van der Vegt, and K. Kremer, *Hierarchical modeling of polystyrene: From atomistic to coarse-grained simulations*, Macromolecules, **39**, no. 19, 6708 – 6719, 2006.

73. V. A. Harmandaris, D. Reith, N. F. A. van der Vegt, and K. Kremer, *Comparison between Coarse-Graining Models for Polymer Systems: Two Mapping Schemes for Polystyrene*, Macromol. Chem. Phys., **208**, 2109 – 2120, 2007.
74. A. Mukherjee and B. Bagchi, *Correlation between rate of folding, energy landscape, and topology in the folding of a model protein HP-36*, J. Chem. Phys., **118**, 4733–4747, 2003.
75. M. R. Betancourt and J. Skolnick, *Local propensities and statistical potentials of backbone dihedral angles in proteins*, J. Mol. Biol., **342**, 635–649, 2004.
76. V. Tozzini, W. Rocchia, and J. A. McCammon, *Mapping all-atom models onto one-bead coarse-grained models: General properties and applications to a minimal polypeptide model*, J. Chem. Theor. Comput., **2**, 667 – 673, 2006.
77. M. R. Betancourt, *Knowledge-based potential for the polypeptide backbone*, J. Phys. Chem. B, **112**, 5058–5069, 2008.
78. D. Alemani, F. Collu, M. Cascella, and M. Dal Peraro, *A Nonradial Coarse-Grained Potential for Proteins Produces Naturally Stable Secondary Structure Elements*, J. Chem. Theory Comput., **6**, 315–324, 2010.
79. R. L. Jernigan and I. Bahar, *Structure-derived potentials and protein simulations*, Curr. Opin. Struct. Biol., **6**, 195 – 209, 1996.
80. I. Bahar and R. L. Jernigan, *Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation*, J. Mol. Biol., **266**, 195 – 214, 1997.
81. R. L. C. Akkermans and W. J. Briels, *A structure-based coarse-grained model for polymer melts*, J. Chem. Phys., **114**, 1020–1031, 2001.
82. R. L. Henderson, *Uniqueness Theorem for Fluid Pair Correlation-Functions*, Phys. Lett. A, **A 49**, 197–198, 1974.
83. J. T. Chayes, L. Chayes, and E. H. Lieb, *The Inverse Problem in Classical Statistical-Mechanics*, Commun. Math. Phys., **93**, 57–121, 1984.
84. M. E. Johnson, T. Head-Gordon, and A. A. Louis, *Representability problems for coarse-grained water potentials*, J. Chem. Phys., **126**, no. 14, 144509, 2007.
85. M. D'Alessandro and F. Cilloco, *Information-theory-based solution of the inverse problem in classical statistical mechanics*, Phys. Rev. E, **82**, 021128, 2010.
86. W. Schommers, *A pair potential for liquid rubidium from the pair correlation function*, Phys. Lett., **43**, 157–158, 1973.
87. A. K. Soper, *Empirical potential Monte Carlo simulation of fluid structure*, Chem. Phys., **202**, 295–306, 1996.
88. V. Rühle, C. Junghans, A. Lukyanov, K. Kremer, and D. Andrienko, *Versatile Object-Oriented Toolkit for Coarse-Graining Applications*, J. Chem. Theory Comput., **5**, 3211–3223, 2009.
89. H. Wang, C. Junghans, and K. Kremer, *Comparative atomistic and coarse-grained study of water: What do we lose by coarse-graining?*, Eur. Phys. J. E, **28**, 221–229, 2009.
90. A. J. Rzepiela, M. Louhivuori, C. Peter, and S. J. Marrink, *Hybrid simulations: combining atomistic and coarse-grained force fields using virtual sites*, Phys. Chem. Chem. Phys., **13**, 10437–10448, 2011.
91. G. M. Torrie and J. P. Valleau, *Non-Physical Sampling Distributions in Monte-Carlo Free-Energy Estimation - Umbrella Sampling*, J. of Comp. Phys., **23**, 187–199, 1977.

92. W. K. Den Otter and W. J. Briels, *The calculation of free-energy differences by constrained molecular-dynamics simulations*, J. Chem. Phys., **109**, 4139, 1998.
93. B. Hess, C. Holm, and N. F. A. van der Vegt, *Osmotic coefficients of atomistic NaCl (aq) force fields*, J. Chem. Phys., **124**, 164509, 2006.
94. B. Hess, C. Holm, and N. F. A. van der Vegt, *Modeling multibody effects in ionic solutions with a concentration dependent dielectric permittivity*, Phys. Rev. Lett., **96**, no. 14, 147801, 2006.
95. J.-W. Shen, C. Li, N. F. A. van der Vegt, and C. Peter, *Transferability of Coarse Grained Potentials: Implicit Solvent Models for Hydrated Ions*, J. Chem. Theory Comput., **7**, 1916–1927, 2011.
96. R. Carr, J. Comer, M. D. Ginsberg, and A. Aksimentiev, *Atoms-to-microns model for small solute transport through sticky nanochannels*, Lab. Chip, **11**, 3766–3773, 2011.
97. A. Villa, N. F. A. van der Vegt, and C. Peter, *Self-assembling dipeptides: including solvent degrees of freedom in a coarse-grained model*, Phys. Chem. Chem. Phys., **11**, 2068–2076, 2009.
98. A. Villa, C. Peter, and N. F. A. van der Vegt, *Transferability of Nonbonded Interaction Potentials for Coarse-Grained Simulations: Benzene in Water*, J. Chem. Theory Comput., **6**, 2434–2444, 2010.
99. W. G. Noid, J.-W. Chu, G. S. Ayton, V. Krishna, S. Izvekov, G. A. Voth, A. Das, and H. C. Andersen, *The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models*, J. Chem. Phys., **128**, 244114, 2008.
100. J. F. Rudzinski and W. G. Noid, *Coarse-graining entropy, forces, and structures*, J. Chem. Phys., **135**, 214101, 2011.
101. A. Chaimovich and M. S. Shell, *Anomalous waterlike behavior in spherically-symmetric water models optimized with the relative entropy*, Phys. Chem. Chem. Phys., **11**, 1901–1915, 2009.
102. R. Baron, D. Trzesniak, A. H. De Vries, A. Elsener, S. J. Marrink, and W. F. van Gunsteren, *Comparison of thermodynamic properties of coarse-grained and atomic-level simulation models*, ChemPhysChem, **8**, 452–461, 2007.
103. M. R. Betancourt and S. J. Omovie, *Pairwise energies for polypeptide coarse-grained models derived from atomic force fields*, J. Chem. Phys., **130**, 195103, 2009.
104. J. W. Mullinax and W. G. Noid, *Extended ensemble approach for deriving transferable coarse-grained potentials*, J. Chem. Phys., **131**, 104110, 2009.
105. E. Brini, V. Marcon, and N. F. A. van der Vegt, *Conditional reversible work method for molecular coarse graining applications*, Phys. Chem. Chem. Phys., **13**, 10468–10474, 2011.
106. J. R. Silbermann, S. H. L. Klapp, M. Schoen, N. Chennamsetty, H. Bock, and K. E. Gubbins, *Mesoscale modeling of complex binary fluid mixtures: Towards an atomistic foundation of effective potentials*, J. Chem. Phys., **124**, 074105, 2006.
107. E. C. Allen and G. C. Rutledge, *A novel algorithm for creating coarse-grained, density dependent implicit solvent models*, J. Chem. Phys., **128**, 154115, 2008.
108. E. C. Allen and G. C. Rutledge, *Evaluating the transferability of coarse-grained, density-dependent implicit solvent models to mixtures and chains*, J. Chem. Phys., **130**, 034904, 2009.
109. J. Ghosh and R. Faller, *State point dependence of systematically coarse–grained potentials*, Mol. Simulat., **33**, 759–767, 2007.

110. V. Krishna, W. G. Noid, and G. A. Voth, *The multiscale coarse-graining method. IV. Transferring coarse-grained potentials between temperatures*, J. Chem. Phys., **131**, 024103, 2009.
111. A. Ben-Naim, *Solvation Thermodynamics*, Plenum Press, New York, 1987.

Particle-Based Dynamics Simulations of Multi-Protein Systems and Cellular Compartments

Volkhard Helms, Po-Hsien Lee, Tihamér Geyer

Center for Bioinformatics
Saarland University, D-66123 Saarbrücken, Germany
E-mail: {Volkhard.Helms, Tihamer.Geyer}@bioinformatik.uni-saarland.de

Coarse-grained representations allow sampling folding/unfolding transitions of proteins in dynamics simulations as well as formation of molecular complexes. We briefly introduce different types of interaction potentials used in the coarse-graining field as well as typical simulation algorithms used to propagate the particle dynamics. Then, two types of applications are described. First, we will discuss the permeation of bead-particles through nanopores. Secondly, we will discuss the formation of molecular assemblies. For this, we will introduce network measures that allow for an efficient description of a large number of association processes that occur in parallel.

1 Introduction

Proteins are biopolymers and consist of one or several chains of amino acids ranging from about 50 amino acids in length (ca. 500 non-hydrogen atoms) to many thousands of amino acids. Around 80% of all cellular proteins — on which we will focus here — adopt a stable, folded conformation of several nanometers in diameter that is encoded by their specific amino acid sequence. Tightly linked to this three-dimensional structure is typically the molecular and cellular function of each protein. One could assume that the complexity of an organism is a function of the number of different proteins encoded in its genome. However, this is only partly true: the simple bacterium *Escherichia coli* has about 4.200 genes, the already much more complex yeast *Saccharomyces cerevisiae* has around 6.000 genes, mammals have 20.000 to 25.000 genes, and the plant model organism *Arabidopsis thaliana* has even 30.000 genes coding for proteins. Thus, the complexity of an organism cannot simply be equated with the number of genes contained in their genome sequences. Rather it is related to the complicated architecture of the cellular circuits, so-called gene regulatory networks, that control the transcription machinery of genes. Moreover, complexity also arises from a myriad of biomolecular interactions occurring in biological cells since about half of all cellular proteins form multi-protein complexes. On average, every protein interacts permanently or transiently with six to eight other proteins as well as with nucleic acids and cell membranes.

Biological cells have typical dimensions of several micrometers. Taking into account that about 30% of the cellular volume is taken up by proteins, one cell roughly contains about 1 billion protein copies. In fact, a recent mass-spectroscopy study quantified the protein copy numbers in mouse fibroblast cells and found a total number of 1.15 billion proteins[1]. Obviously, studying the interactions of so many proteins by atomistic modelling approaches poses enormous problems because the number of interactions between atom pairs to be computed at every simulation time step is immense, and also because the time step required to sample the protein dynamics at atomistic details is in the order of a few

femtoseconds only. However, a vibrant new research area deals with describing proteins at coarser detail in order to reduce the number of interactions to be computed so that longer simulation times can be reached. As an example of cutting-edge research in this field of coarse-grained many-particle simulations we mention the work of McGuffee and Elcock who studied a model of the bacterial cytoplasm using multi-particle Brownian Dynamics simulations[2]. More precisely, they simulated the dynamic trajectories of 1000 cellular proteins whose shapes were modelled in atomistic detail in an implicit solvent environment over six microseconds. Comparison with experimental observables was made by following the trajectories of eight "reporter" green-fluorescent proteins. When suitably equilibrated, such simulations are nowadays able to study biomolecular processes in subcellular compartments that take place on microsecond time scales.

As an introduction to this field of modelling we will first introduce the simulation methodology and then illustrate it with three applications to protein systems.

2 Coarse-Grained Simulations of Proteins

Comprehensive introductions into the field of coarse-grained (CG) protein models are provided in Refs. 3-5. Some methods employ prior knowledge about the folded conformation of a protein. For example, elastic network models are bead-spring models that have been successfully employed to model harmonic geometric deformations of proteins about a reference structure[6]. Typically, this is the folded structure of a protein determined by X-ray crystallography. Each amino acid (residue) is represented by a single bead and spatially neighboring residues are connected by elastic springs. Another type of models requiring information about a reference configuration are Go-models[7]. These are often used for studying protein folding mechanisms, where during the simulation the protein chain is biased towards the native, folded conformation by means of simple attractive or repulsive non-bonded interactions between the beads. Go-type lattice simulations have allowed to derive fundamental principles governing the protein folding process[8].

2.1 Coarse-Grained Force Fields

In the following, we will focus on unbiased force fields that can be applied to model transitions between folded and unfolded conformations of proteins and to model association and dissociation processes. Force fields that refer to physical interactions express the total energy of the system, U, as a sum of various contributions.

$$U = \sum_i u_i^{local} + \sum_i \sum_{j>i} u_{ij} + \sum_{ijk} u_{ijk} + ... \quad (1)$$

Here, u_i^{local} denotes a local interaction term dependent on a single site, u_{ij} models the effective interaction between sites i and j, and u_{ijk} (and potentially higher order terms) denote multibody interactions between sites i, j, k etc. The term "site" stands here for the center of a bead that may represent either a part of an amino acid, an entire amino acid, or even larger structural units. Three methods are frequently employed for constructing such coarse grained potentials[4]: Boltzmann inversion of distribution functions, inverse Monte Carlo sampling, and force matching.

The force-matching method presented by Voth and coworkers[9] determines the coarse-grained potential function such that the derived mean forces acting on the CG sites show the least quadratic deviations from the average forces computed from an all-atom molecular dynamics (MD) simulation of the same system.

Boltzmann inversion of distribution functions can be either applied to distributions sampled by higher resolution models or to distributions characterized by experiments. One example of the latter category are knowledge based protein force fields that exploit the accumulating amount of data on the structures of folded proteins[5]. Here, one constructs an effective energy function based on the distributions of inter-residue distances, virtual-bond lengths, bond angles, dihedral angles, and other geometric parameters derived from structures deposited in the Protein Data Bank[10]. The basic equation used in deriving statistical potentials by Boltzmann inversion[5] is

$$W(\mathbf{x};\mathbf{s}) \;=\; -RT \ln \frac{N^{obs}(\mathbf{x};\mathbf{s})}{N^{ref}(\mathbf{x};\mathbf{s})}, \tag{2}$$

where $W(\mathbf{x};\mathbf{s})$ is the estimated potential of mean force of a fragment with geometry $\mathbf{x}$ and amino acid sequence or secondary-structure content $\mathbf{s}$, R is the universal gas constant, and T the absolute temperature. $N^{obs}(\mathbf{x};\mathbf{s})$ is the number of occurrences of fragments of this sequence and close to this geometry in the reference data set, and $N^{ref}(\mathbf{x};\mathbf{s})$ is the respective count in the absence of any interactions.

2.2 Integration Algorithms

When the atomistic details of the protein(s) are known, for example from crystal structures, molecular dynamics (MD) simulations can be performed to study the conformational dynamics of the protein(s). Given a many-particle interaction potential U, the MD method propagates the atomistic coordinates by a finite-difference propagation algorithm using time-scales on the order of the fastest degrees of freedom that are modelled explicitly. In atomistic simulations these are the vibrations of chemical bonds. For this, Newton's equations of motion are solved to obtain the changes of the coordinates and momenta of all N particles from the derivatives of U with respect to the coordinates of each of the particles:

$$m_i \frac{\mathrm{d}^2\mathbf{r}_i}{\mathrm{d}t^2} \;=\; -\nabla_{r_i} U \quad \text{with} \quad i = 1, 2, \ldots, N \tag{3}$$

Here, $\mathbf{r}_i$ is the three dimensional vector of the Cartesian coordinates of particle i, and m_i is its mass. The derivative $-\nabla_{r_i} U(\mathbf{r}_i)$ yields the force acting on particle i. MD is the most popular, but also most expensive method for studying the dynamics of complete proteins.

As long as the dynamics of individual proteins is investigated, the simulation box can be made relatively small so that only a part of the overall computational effort is spent on the explicitly modelled but otherwise uninteresting water molecules around the protein. For the diffusional association of two or more proteins, however, the simulation box has to be made much larger and most of the computational cost is spent just to propagate the water molecules. Thus, an efficient way to speed up the simulations is to replace the large number of explicitly modelled waters by an implicit solvent. Then, all the resources can be used to actually propagate the proteins.

If the solvent is removed from the simulation, its effects on the dynamics of the proteins, however, have to be added back explicitly via effective interactions. The two most import effects of the solvent are the friction due to the displacement of the solvent molecules when the proteins are moved, and the random kicks on the proteins resulting from the thermal motion of the solvent molecules. Furthermore, the interactions between the proteins have to be adapted to include, e.g., screening by counter ions, and so-called hydrodynamic interactions (HI) describe how the solvent displaced due to the motion of one protein pushes onto the neighboring proteins.

The friction, i.e., the conversion of the directed kinetic energy of a protein into undirected thermal motion of the surrounding solvent is modelled by a Stokes model, where the friction force $F_i^{(r)}$ is proportional to the velocity of the protein:

$$F_i^{(r)} = -\gamma_i v_i \tag{4}$$

The friction coefficient for a sphere scales with its radius a as $\gamma = 6\pi\eta a$, where η is the solvent viscosity. Thus, friction increases slower than the cross sectional area of a moving particle or its mass, so that that friction is only a minor effect for large, macroscopic particles, while it dominates the dynamics of small proteins or molecules.

In an atomistic simulation the collisions between the solvent molecules and the larger proteins can all be observed individually. However, for an implicit-solvent model that is meant to mimick effects on long time-scales these details do actually not matter. It is enough to know that, on average, the kicks do not lead to a directed displacement when the solvent near to the protein is isotropic. The next statistical moment is the variance, i.e., the strength of the kicks. It is related to the solvent temperature. When the mobility of the protein is expressed via its diffusion coefficient $D = k_B T/\gamma$, the displacements R_i resulting from the random kicks of the solvent molecules over a time interval Δt are conveniently expressed as

$$\langle R_i \rangle = 0 \quad \text{and} \quad \langle R_i R_k \rangle = 2D_{ik}\Delta t. \tag{5}$$

These findings, which relate the thermal energy $k_B T$ transferred from the solvent onto the protein to the dissipation of the protein's kinetic energy via friction are the core of Einstein's seminal explanation of Brownian motion[11]. They are enough to describe the diffusive motion of a single particle without any external fields.

For more interesting scenarios, however, multiple particles and the forces between them have to be considered, too. For a short derivation of a suitable implicit solvent propagation scheme we go back to the many-particle Newton equation (Eq. 3) of the protein(s) and the many solvent molecules, which can also be written as

$$m_i \frac{dv_i}{dt} = F_i = \sum_{k \neq i} F(r_{ik}), \tag{6}$$

where the changes of the particle velocities v_i are due to the sums of all external forces F_i. In the atomistic description, these are pairwise, distance-dependent, conservative forces $F(r_{ik})$. This many-particle system of Newton equations can be simplified into an implicit-solvent Langevin equation with the above explained friction term and the random kicks by the solvent f_i :

$$\frac{dv_i}{dt} = \frac{1}{m_i}(F_i + f_i - \gamma v_i) \tag{7}$$

The random forces f_i can be calculated from the random displacements (Eq. 5) as[12]

$$\langle f_i \rangle = 0 \quad \text{and} \quad \langle f_i f_k \rangle = \frac{2(k_B T)^2}{D_{ii} \Delta t} \delta_{ik}. \tag{8}$$

Assuming that the total force F_i acting on particle i is constant for a short time interval Δt, the above Langevin equation (Eq. 7) can be integrated analytically. For convenience we will drop the particle index i in the following and use F for the sum of the direct interactions between the proteins F_i plus the random kicks f_i. Then, the velocity $v(\Delta t)$ at the end of the timestep and the displacement $\Delta x(\Delta t)$ during Δt are given by

$$v(\Delta t) = \frac{F}{\gamma} + \left(v_0 - \frac{F}{\gamma} \right) \exp\left[-\frac{\gamma \Delta t}{m} \right] \tag{9}$$

and

$$\Delta x(\Delta t) = \frac{F}{\gamma} \Delta t - \frac{m}{\gamma} \left(\frac{F}{\gamma} - v_0 \right) \left(1 - \exp\left[-\frac{\gamma \Delta t}{m} \right] \right), \tag{10}$$

where v_0 is the velocity of the particle at the beginning of the timestep. These two Eqs. 9 and 10 can now be used directly to propagate the particles in an implicit solvent Langevin Dynamics (LD) scheme[12]. It assumes that all solvent molecules can be replaced by a velocity dependent Stokesian friction term and random kicks, and that the timestep is small enough so that the (configuration dependent) forces do not really change during Δt.

As can be seen from the LD equations (Eqs. 9 and 10), the contribution of the initial velocity v_0 decreases with a time constant $\tau = m/\gamma$, which is correspondingly called the velocity relaxation time. As mentioned above, τ becomes shorter than the relevant timescale Δt of the particle motion for larger particles. For colloidal particles or the pollen grains for which the stochastic motion had originally been observed by Robert Brown in 1827, τ is much smaller than a typical observation interval Δt and then the LD equations can be simplified to the well-known equations of Brownian dynamics (BD):

$$v(\Delta t) = \frac{F}{\gamma} \quad \text{and} \quad \Delta x(\Delta t) = \frac{F}{\gamma} \Delta t. \tag{11}$$

In this so-called overdamped regime the velocity instantaneously follows the force and can consequently be ignored during the simulation. This also means that in the Brownian regime no ballistic motion occurs, i.e., the particles grind to a halt as soon as the forces vanish. Remember, that in Eqs. 9, 10, and 11 the force F consists of external, inter-particle, and random contributions.

The most famous implementation of the BD equations is the algorithm by Ermak and McCammon[13]. In their algorithm the random forces are evaluated independently from the external forces and the displacements according to Eq. 5 are used directly. They were also the first ones who showed how hydrodynamic interactions (HI) that lead to correlations in the velocities of the particles via the displaced solvent can be considered, too. For more details and an efficient approximation to the many-body correlations see[14].

Other BD algorithms also exist in which the simple first-order integrator of Eq. 11 is replaced by more efficient schemes[15–17]. These, however, do not include HI — which have been shown recently to be important for protein-protein association processes[18] or flexible coarse-grained protein models[19,20].

3 Applications

3.1 Protein-Protein Association

When applied to a system of two proteins, the BD method just introduced describes the relative motion of the two proteins as a diffusive motion of two rigid bodies subject to external forces. The McCammon and Wade groups developed a computational scheme where one protein is kept fixed and the electrostatic potential around this protein is computed on a cubic grid by numerically solving the Poisson-Boltzmann equation[21,22]. This computation is done only once during initialization. In BD simulations, the second, moving protein may then be approximated by about 20 to 30 suitably placed electrostatic point charges[23]. At each BD step, the Coulombic interactions are computed from these effective charges and the electrostatic potential interpolated from the nearest grid points. The spatial shapes of the proteins are mapped onto grids, too, and a rejection algorithm prevents the spatial overlap of the proteins. Compared to atomistic MD simulations of the same system, the BD method is orders of magnitude (10^6 to 10^9 times) more efficient. By statistical averaging over a large number of BD trajectories, BD simulations have been shown to successfully reproduce experimental k_{on} rates for the association of electrostatically complementary protein-protein pairs[24]. Here, a successful binding event is counted as soon as two to four out of a given set of characteristic inter-protein contacts are established[24].

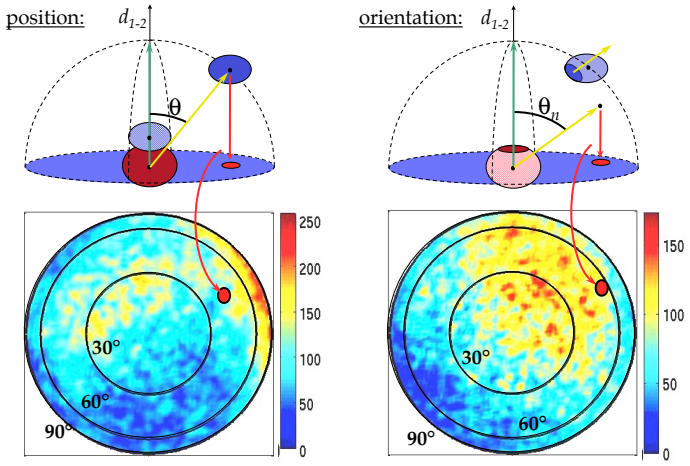


Figure 1. (Top) For visualizing the preferred association pathways, the relative position of the second protein during a BD trajectory with respect to the first protein kept fixed at the origin is projected onto the plane slicing through the center of mass of the fixed protein. In this particular orientation at a separation distance d_{1-2}, the second protein is located off-center at an angle θ from the line connecting the two proteins in the bound complex. (Bottom) Color-coded heat map showing the accumulated occupancies in the plane.

Dr. Alexander Spaar from our group has introduced the characterization of the free energy landscape for protein-protein interactions by on-the-fly-analysis of BD trajectories[25]. For this, he made use of the fact that BD trajectories mostly sample the low-energy valleys of the underlying free energy landscape. By storing the configurations visited during

the simulations in two occupancy maps for translations and rotations, one can deduce the shape of this unknown free energy surface. For this, the contributions of electrostatic and desolvation energies, as well as the translational and rotational entropy losses, are stored in matrices that represent the three-plus-three-dimensional configuration space. These matrices have the same grid sizes as the occupancy maps and they are computed separately for the positional and orientational coordinates. Fig. 1 illustrates the mapping of visited positions onto a two-dimensional hemisphere at a particular protein-protein distance d_{1-2}. To compute the local entropy loss at a given position and orientation with respect to the unbound state, the occupancy landscape is interpreted as a probability distribution. Boltzmann's definition of the entropy S then leads to

$$S = -k_B \sum P_n ln P_n \tag{12}$$

where the P_n are the probabilities for each state n.

With the energy and entropy contributions as functions of the translational and rotational coordinates, the free-energy landscape of the encounter process is given by the sum of the electrostatic energy ΔE_{el} computed during the BD step, the desolvation energy ΔE_{ds} estimated using an approximative formula, and the change of the translational/rotational entropy with respect to that of bulk solution:

$$\begin{aligned} \Delta G &= \Delta E_{el} + \Delta E_{ds} - T\Delta S_{tr} \\ \Delta S_{tr} &= \Delta S_{trans} + \Delta S_{rot} \end{aligned} \tag{13}$$

We applied this procedure to the association of the proteins barnase and barstar[26], see Fig. 2. The obtained ΔG profile along the lowest free energy pathway agrees well with that obtained by atomistic MD simulations in explicit solvent for the same protein:protein complex using an umbrella potential restraint[27] except for very close distances. There, the atomistic PMF continues to proceed downhill towards the bound complex due to an enhanced electrostatic attraction caused by a strongly ordered solvent[28] whereas the BD potential shows an upward swing. The MD simulation could, of course, only sample the approach along a particular one-dimensional pathway, whereas with BD the complete association funnel could be characterized. This example illustrates both the numerical efficiency of the BD method in sampling relative protein positions at medium to large distances and its limitations at very short separation distances.

In the same way, we have studied the interaction of cytochrome c with membrane-embedded cytochrome c oxidase[29]. Recent applications by other research groups that utilized the implementation of this sampling scheme in the SDA package of Rebecca Wade's group[22] include studying the association of TEM1-beta-lactamase and its inhibitor, beta-lactamase-inhibitor protein[30] and the assembly of icosahedral virus shells[31].

3.2 Diffusion of Bead Particles through Nanopores

A crucial process in biological cells is the translocation of newly synthesized proteins across cell membranes via integral membrane protein pores termed translocons. With recent techniques artificial porous membranes can be built with similar pore dimensions as the translocon system, i.e., with radii of a few nanometers. These artificial membranes then allow studying the behavior of the bare proteins without the inherent complications involved in modelling the translocon as well. To study the permeation of proteins through

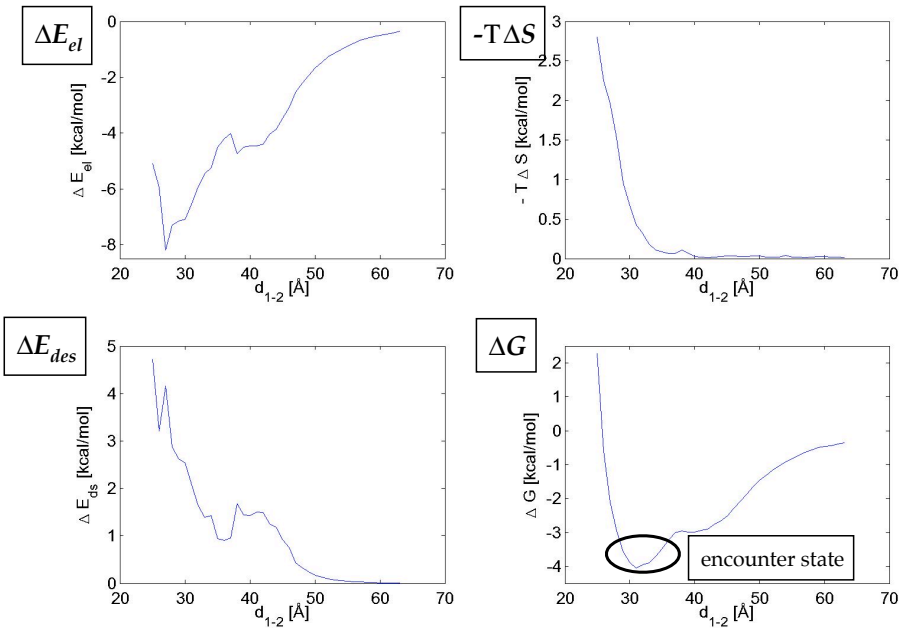


Figure 2. Energy profiles for the association of barnase and barstar as a function of the separation distance of the two protein centers: The electrostatic energy ΔE_{el} (upper left), the desolvation energy ΔE_{ds} (lower left), the translational/rotational entropy loss (upper right), and, as the sum of all, the free energy ΔG (lower right).

such porous membranes, we used coarse-grained BD simulations where proteins were modeled as spherical beads with a radius of 1.67 nm. This equals the hydrodynamic radius of the protein cytochrome c, i.e., its radius plus the thickness of its hydration layer. The free diffusion coefficient of the protein particles was set to $D = 1.48 \times 10^5 nm^2/ps$. Urea-unfolded cytochrome c was represented by bead-spring polymers with 2 to 10 beads such that the experimentally determined radius of gyration and diffusion coefficient were reproduced. The pores were represented by cylindrical openings in the membrane with varying radius r_{pore} and length L_{pore}. To reduce the complexity of the system, a simple repulsive Lennard-Jones potential was used between the different proteins and between the proteins and the membrane. Diffusion was driven by a concentration gradient created across the porous membrane. For this, a particle insertion/removal algorithm[32] allowed us to keep the particle concentrations at the walls of the simulation box above and below the membrane at fixed values ρ_{cis} and ρ_{trans}. This setup is sketched in Fig. 3.

In the simulations, pore radii of 4 to 16 nm and pore lengths between 5 and 40 nm were used to investigate how the pore geometry affects the diffusive flux Φ across the membrane.

An analytical continuum model of Brunn *et al.*[33] predicts that the diffusive flux Φ across a membrane with pores of length L_{pore} and radius r_{pore} is determined by the pore size and its aspect ratio r_{pore}/L_{pore} as

$$\Phi \;=\; C\frac{\rho_{cis}-\rho_{trans}}{\alpha\pi r_{pore}+L_{pore}} \;\propto\; \frac{1}{L_{pore}(1+\alpha' r_{pore}/L_{pore})} \tag{14}$$

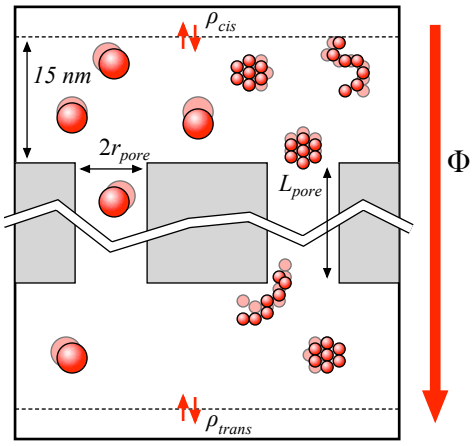


Figure 3. Sketch of the setup of the nanopore system: the *cis* and *trans* reservoirs of the experiment are modelled by constant density boundary conditions[32] that fix the protein densities to ρ_{cis} and ρ_{trans}, respectively. Unlike depicted in this sketch, in the simulations only one representation of the cytochrome c was used at a time. The models ranged from a single sphere for the folded variant up to bead-spring models with $N \leq 10$ subunits for the denatured, unfolded proteins. The resulting diffusive particle current Φ across the membrane was determined for various pore lengths L_{pore} and radii r_{pore}.

In this equation, C is a normalization constant that contains the number of pores, their cross sections, and the diffusion coefficient of the particles. In the analytical continuum model α is a constant close to unity.

This equation was compared to the simulation results via the parameters C and α. Fig. 4 shows for one example how the predicted decrease of Φ with increasing L_{pore} for a constant $r_{pore} = 8$ nm is reproduced by the simulations. Obviously, reproducing this

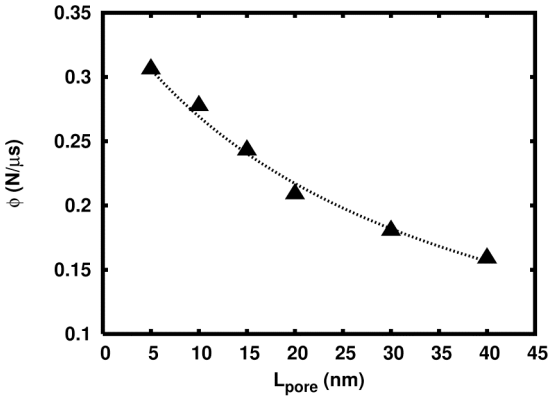


Figure 4. Decrease of the diffusive flux Φ of folded cytochrome c across a membrane with pores of length L_{pore} and radius $r_{pore} = 8$ nm. The triangles are simulation results, the dashed line is a fit using Eq. 14.

idealized case was only done as a control. The simulation model now allows to investigate the effects of varying interaction potentials between the membrane and the particles as well as to compare how unfolded proteins translocate through narrow pores. Also, the pore dimensions can be further decreased and the pore geometry can be made more complicated so that the pore may finally resemble a coarse-grained model of the cellular translocon.

3.3 Using Dynamic Interaction Graphs to Analyze Multi-Protein Association

Whereas simulations with only two or three particles can be easily analyzed by short scripts, it can become a tedious and computationally demanding task to monitor whether and when complexes with tens to hundreds of constituents are formed correctly. This problem becomes even more pronounced when more realistic simulations of large numbers of different proteins are considered, where more than one complete complex can be formed or where the complete complex is in a dynamic equilibrium with its components. In such a simulation, partial complexes of various sizes may be found together with complete ones and even with complexes which are assembled incorrectly. For situations such as the one described above, where one wants to identify a few target complexes in a sea of monomers and partially assembled intermediates, this task can be conveniently performed by mapping the spatial simulation onto a protein interaction graph, which can then be analyzed conveniently with efficient, well-known graph measures and algorithms[34]. This protein interaction network built from the simulation differs from the well-known protein-protein interaction networks because, here, each of its nodes denotes an individual copy of a protein. Additionally, the graph is generated dynamically. This means that links appear and disappear over time, as the proteins bind and unbind from each other during the spatial simulation.

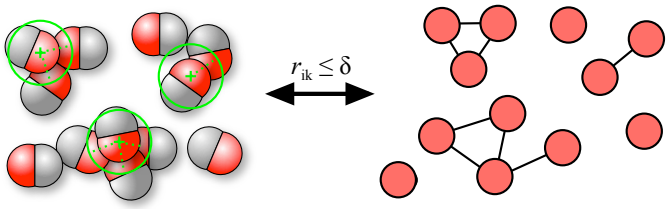


Figure 5. Mapping of a snapshot of a spatial simulation (left) onto a protein association graph (right) via a distance criterion. The red sides of the particles in the left panel are meant to attract each other. Correspondingly, a link is added between the corresponding nodes of the network when two of the red ends are closer together than the specified distance δ as indicated by the circles. To facilitate the comparison of the two plots, the network is arranged analogously to the spatial snapshot, even though the actual layout contains no information.

For converting the spatial snapshot into a network representation, each particle is associated with a node of the graph and their contacts or proximity are used to add links between the respective nodes. The most simple criterion for a link is a distance criterion as sketched in Fig. 5. The left-hand side depicts a spatial snapshot of simple dipolar particles with a "sticky" red half and a non-binding grey end. Whenever the red ends of two particles are closer than a specified minimal distance δ as indicated by the green circles, a link is

established between the respective nodes. The network resulting from this example snapshot is depicted on the right-hand side, here with a graph layout that resembles the spatial configuration. Note, however, that the information about the binding or association of the proteins is contained in the connectivity of the graph and not in its layout (which could have been drawn completely differently, too). The main advantage of doing the analysis on the graph is that any spatial configuration is represented as a set of binary connections, whereas in real-space also higher order terms as in Eq. 1 have to be considered. Whereas the number of possible binary distances grows quadratically with the particle number N, the number of angles, defined via three particle positions, already increases as $\mathcal{O}(N^3)$. Thus, when regular structures have to be identified in many-particle simulations, a lot of time is required to loop over all possible three- and maybe even four-body configurations. One of the examples below will show how icosahedral complexes of twelve particles can unambiguously be identified from their signatures in the connectivity of the graph which is based solely on binary interactions.

The most basic measure of a network is its size, i.e., the numbers of nodes, N, and links, L. Most often, the total number of particles in a simulation is constant. Thus, the number of links already gives an indication about how many contacts are formed in the spatial simulation. A connected component is a set of nodes that are all reachable from each other. Such a connected component of the graph representation corresponds to a cluster of particles in real space. Once the graph is set up (in $\mathcal{O}(N^2)$ time), checking for connected components is straightforward and their size distribution directly yields the sizes of the clusters in the simulation. Whereas the network and connected component sizes are rather global measures, a basic local quantifier is the so-called degree k of a node. It gives the number of links attached to this node. The degree distribution

$$P(k) \;=\; \frac{n_k}{N} \tag{15}$$

represents the normalized counts of how many nodes have a degree of k. A well-peaked $P(k)$ indicates a rather regular network structure, whereas a broad distribution of degrees shows that the underlying spatial configuration is inhomogeneous (also see the second example of Ref. 34). Further simple network measures are the clustering coefficient $C_i(k)$ that counts which fraction of the k neighbors of particle (node) i are themselves connected, or the distribution of shortest path lengths $D(l)$ which can be used to quantify the compactness of the connected components. For further graph measures see, e.g., the reviews by Albert and Barabási[35] or by Costa *et al.*[36].

As a first application of this graph analysis we review simulations that were inspired by the formation of icosahedral virus capsids. In these simple Monte-Carlo simulations particles with suitably arranged binding patches could form icosahedral complexes of twelve particles[34]. The simulations with up to 200 monomers were then mapped onto graphs and tested for the well-defined network signatures of the highly symmetric icosahedral complexes. Fig. 6 shows a few examples of observed complexes, both of the correctly formed icosahedron (second row) and of incorrectly assembled complexes of various sizes. Once the graph is set up, it is numerically much faster to identify the connected components and determine the degree distributions, the clustering coefficients, and the path length distributions for each of these typically small subgraphs than to perform a similar analysis in real space. And even setting up the graph was not expensive, as only two-body distances were required.

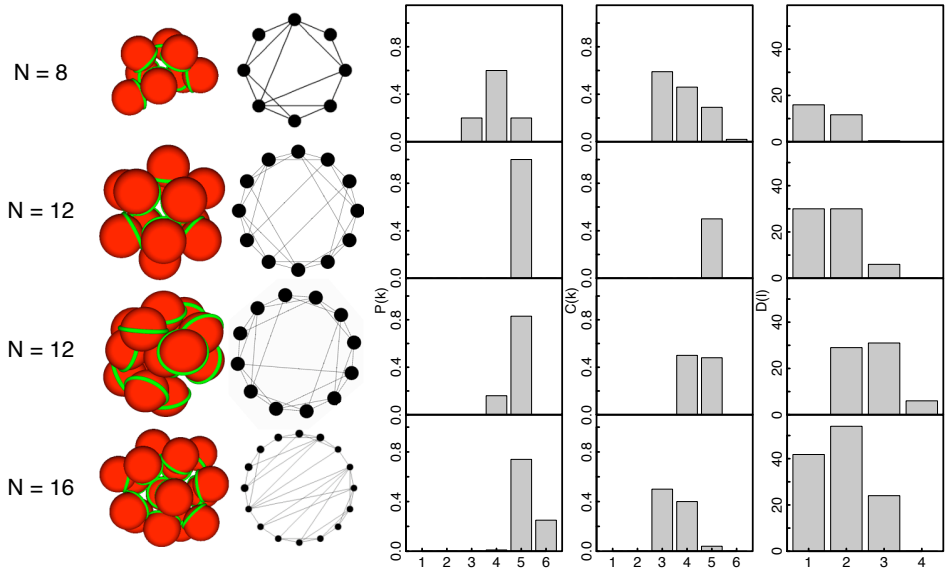


Figure 6. Network signatures are used to identify correctly assembled icosahedral complexes: the first column shows the spatial configurations of some example complexes, for which the next four columns give the corresponding association graphs, the degree distributions $P(k)$, the degree-dependent clustering coefficients $C(k)$, and the distribution of shortest all-neighbor path-lengths, $D(l)$. A correctly folded icosahedral complex is shown in the second row. Its high symmetry leads to well-defined signatures in its network measures. Any deviation from either the correct size or configuration, as observed in the three imperfectly assembled complexes, leads to a smearing out of the signature profile of contacts. This figure was adapted from Ref. 34.

Essentially, Fig. 6 shows details derived from single snapshots. It is straightforward to count the number of icosahedra per snapshot and thus to observe how fast these model virus capsids form during the course of the simulation. Similarly, the evolution of the connected component size distribution vs. time can be used to visualize *and* to quantify the dynamics of transient cluster formation and break-up as shown in the next example[20] (see Fig. 7). The upper row shows two snapshots from a simulation of 27 dipolar particles (see Fig. 5). The attraction was weak so that complexes could only form transiently. The lower panel shows the evolution of the observed cluster sizes over time with the two arrows indicating the two snapshots.

Especially for many-particle simulations the network analysis is an efficient and powerful tool to visualize the dynamics quantitatively, to identify specific (small) target complexes via their signatures, or to characterize the degree of ordering in large many-particle agglomerates. In principle, all these analyses could be done in real space, too, but a graph is the natural, and thus most appropriate data structure to store and to handle connectivity even for very many particles.

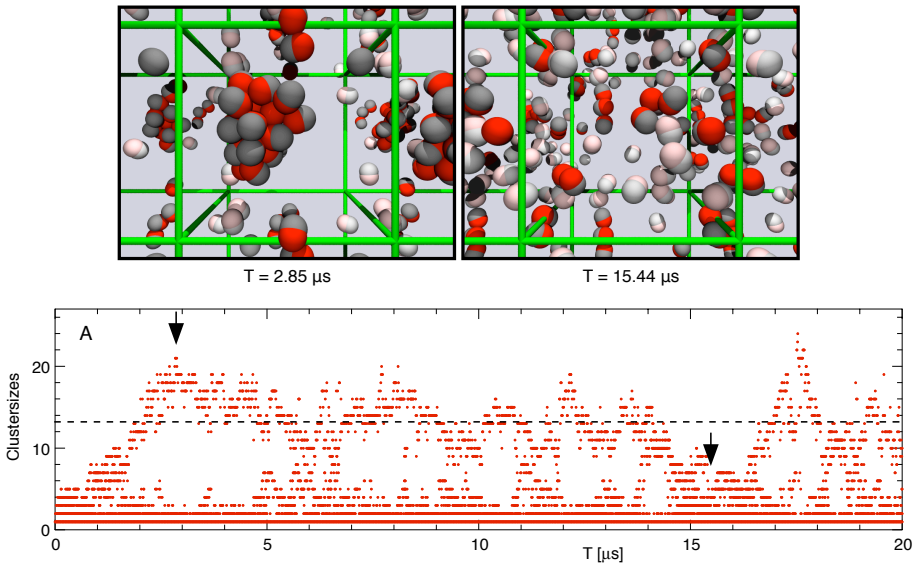


Figure 7. Visualization of the dynamics of contact formation and break-ups during a BD simulation of 27 dipolar particles with periodic boundary conditions. The top row illustrates two snapshots. The left one contains one of the largest complexes observed, the right one was taken from an interval where only small complexes with up to five particles existed. The lower panel shows the distribution of occurring cluster sizes vs. simulation time. Each dot denotes a cluster of a given size found in the snapshot. The large fluctuations in this plot indicate the highly dynamic nature of the simulation in which the particles had only a weak attraction with each other. The locations of the two snapshots shown on top are indicated by arrows. This figure was adapted from Ref. 20.

4 Summary and Outlook

On the example of several case studies, we showed how coarse-grained implicit solvent Brownian Dynamics (BD) simulations can be used efficiently to investigate the diffusional dynamics of protein association ranging from a pair of proteins up to many-particle systems. Compared to atomistic simulations in an explicit solvent, such BD schemes are many orders of magnitude faster because the number of particles is dramatically reduced. The basic idea is that for long-time trajectories it is not necessary to follow the path of each single water molecule. All what is needed from the internal structure of the solvent is the effective friction that is felt by the large proteins when they push away the solvent molecules and the random thermal kicks from the solvent onto the proteins. Plugging these effective interactions into the many-particle Newton equation leads to a few-body Langevin equation for the proteins alone, which can be used to propagate small proteins and peptides[12]. For larger proteins the observation (simulation) timesteps can be made longer and then the well-known BD propagator is recovered in the limit of a vanishing velocity relaxation time. This simple simulation method has become a workhorse technique since about three decades, and interest in such coarse-grained methods is increasing as larger and larger biological systems are being simulated.

One of the main applications of BD has been the estimation of association rates. Our

first example showed that beyond the mere rate constants the complete free energy landscape can be recovered[25,26] when the configurations visited during the trajectories are interpreted as occupation probabilities in the two-particle configuration space.

Our second example demonstrated how a non-equilibrium scenario can be used to investigate the translocation of folded and unfolded proteins through narrow nanopores with biologically relevant diameters. A most simple model with only repulsive interactions converges to the analytically derived dependency of the diffusive flux on the pore geometry. Based on this scaffold we can now investigate how the types and strengths of the interactions and the folding state of the protein affect the translocation process. Further, we can add details to the now still cylindrical pores to closer mimick real biological translocons.

Our third example demonstrated that association dynamics and complex formation in many-particle simulations can conveniently and efficiently be monitored and analyzed quantitatively with the help of dynamic association graphs. In these graphs, which can be set up with distance or energy criteria, the information about the sizes and the ordering of the complexes is encoded only using binary interactions. In contrast, spatial snapshots require already three particle positions for the definition of contact angles.

Other recent methodological developments in our group that further widen the applicability of coarse-grained implicit-solvent methods like BD are a fast approximation to evaluate the hydrodynamic many-body correlations[14] and the actual use of the Langevin dynamics (LD) propagation scheme which here only served to derive the BD equations of motion. In fact, the LD propagation allows to use coarse-grained techniques even for small particles such as parts of an amino acid for which the approximation of long observation times used for BD breaks down due to the small timesteps required for the fast dynamics[12]. HI, on the other hand, are crucial for flexible multi-bead models of proteins[20] that then allow to investigate the folding even of larger proteins. These recent developments aid in closing the resolution gap between the atomistic and united atom approaches for individual proteins, on one hand, and the simplifying many-body techniques like BD and LD, on the other hand. Now, for each combination of system size, resolution, and numerical costs there is an appropriate description at hand.

Acknowledgments

The authors thank the DFG for funding through GK 1279 and through HE3875/5-1.

References

1. B. Schwanhäuser, D. Busse, N. Li, G. Dittmar, J. Schuchardt, J. Wolf, W. Chen, and M. Selbach, *Global quantification of mammalian gene expression control*, Nature **473**, 337-342, 2011.
2. S. R. McGuffee, and A. H. Elcock, *Diffusion, crowding and protein stability in a dynamic molecular model of the bacterial cytoplasm*, PLoS Comput. Biol. **6**, e1000694, 2010.
3. V. Tozzini, *Coarse-grained models for proteins*, Curr. Opin. Struct. Biol. **15**, 144-150, 2005.
4. C. Peter, and K. Kremer, *Multiscale simulation of soft matter systems*, Farad. Discuss. **144**, 9-24, 2010.

5. C. Czaplewski, A. Liwo, M. Makowski, S. Oldziej, and H. A. Scheraga, *Coarse-Grained Models of Proteins: Theory and Applications* in Multiscale Approaches to Protein Modeling, A. Kolinski (Ed.), Springer, pp. 35-83, 2011.
6. I. Bahar, A. R. Atilgan, and B. Erman *Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential*, Fold. Des. **2**, 173-181, 1997.
7. Y. Ueda, H. Taketomi, and N. Go, *Studies on protein folding, unfolding and fluctuations by computer simulation. A three-dimensional lattice model of lysozyme*, Biopolymers **17**, 1531-1548, 1978.
8. J. D. Bryngelson, J. N. Onuchic, N. D. Socci and P. G. Wolynes, *Funnels, Pathways, and the Energy Landscape of Protein-folding — A Synthesis*, Proteins **21**, 167-195, 1995.
9. S. Izvekov, and G. A. Voth, *A Multiscale Coarse-Graining Method for Biomolecular Systems*, J. Phys. Chem. B **109**, 2469-2473, 2009.
10. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, *The Protein Data Bank*, Nucl. Acids Res. **28**, 235-242, 2000.
11. A. Einstein, *Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen*, Ann. Phys. **17**, 549-560, 1905.
12. U. Winter, and T. Geyer, *Coarse grained simulations of a small peptide: Effects of finite damping and hydrodynamic interactions*, J. Chem. Phys. **131**, 104102, 2009.
13. D. L. Ermak, and J. A. McCammon, *Brownian Dynamics with Hydrodynamic Interactions*, J. Chem. Phys. **69**, 1352-1360, 1978.
14. T. Geyer, and U. Winter, *An $\mathcal{O}(N^2)$ approximation for hydrodynamic interactions in Brownian dynamics simulations*, J. Chem. Phys. **130**, 114905, 2009.
15. D. A. Beard, and T. Schlick, *Inertial stochastic dynamics. I. Long-time-step methods for Langevin dynamics*, J. Chem. Phys. **112**, 7313-7322, 2000.
16. R. D. Skeel, and J. A. Izaguirre, *An impulse integrator for Langevin dynamics*, Molec. Phys. **100**, 3885-3891, 2002.
17. G. De Fabritiis, M. Serrano, P. Español, and P. V. Coveney, *Efficient numerical integrators for stochastic models*, Physica A **361**, 429-440, 2006.
18. T. Frembgen-Kesner, and A. H. Elcock, *Absolute protein-protein association rate constants from flexible, coarse-grained Brownian dynamics simulations: the role of intermolecular hydrodynamic interactions in barnase-barstar association*, Biophys. J. **99**, L75-L77, 2010.
19. T. Frembgen-Kesner, and A. H. Elcock, *Striking effects of hydrodynamic interactions of the simulated diffusion and folding of proteins*, J. Chem. Theory Comput. **5**, 242-256, 2009.
20. T. Geyer, *Many-particle Brownian and Langevin dynamics simulations with the Brownmove package*, BMC Biophysics **4**, 7, 2011.
21. J. D. Madura, J. M. Briggs, R. C. Wade, M. E. Wade, B. A. Luty, A. Ilin, J. Antosiewicz, M. K. Gilson, B. Bagheri, L. R. Scott, and J. A. McCammon, *Electrostatics and diffusion of molecules in solution: Simulations with the University of Houston Brownian Dynamics Program*, Comp. Phys. Commun. **91**, 57-96, 1995.
22. R. R. Gabdoulline, and R. C. Wade, *Brownian Dynamics Simulation of Protein-Protein Encounter*, Methods **3**, 329-341, 1998.

23. R. R. Gabdoulline, and R. C. Wade, *Effective Charges for Macromolecules in Solvent*, J. Phys. Chem. **100**, 3868-3878, 1996.
24. R. R. Gaboulline, and R. C. Wade, *Protein-protein association: Investigation of factors influencing association rates by Brownian dynamics simulations*, J. Mol. Biol. **306**, 1139-1155, 2001.
25. A. Spaar, and V. Helms, *Free Energy Landscape of Protein-Protein Encounter Resulting from Brownian Dynamics Simulations of Barnase: Barstar*, J. Chem. Theor. Comput. **1**, 723-736, 2005.
26. A. Spaar, C. Dammer, R. R. Gabdoulline, R. C. Wade, and V. Helms, *Diffusional Encounter of Barnase and Barstar*, Biophys. J. **90**, 1913-1924, 2006.
27. L. Wang, S. Siu, W. Gu and V. Helms, *Downhill binding energy surface of the barnase-barstar complex*, Biopol. **93**, 977-985, 2010.
28. M. Ahmad, W. Gu, T. Geyer, and V. Helms, *Adhesive water networks facilitate binding of protein interfaces*, Nature Commun. **2**, 261, 2011.
29. A. Spaar, D. Flöck, and V. Helms, *Association of cytochrome c with membrane-bound cytochrome c oxidase proceeds parallel to the membrane rather than in bulk solution*, Biophys. J. **96**, 1721-1732, 2009.
30. M. Harel, A. Spaar, and G. Schreiber, *Fruitful and Futile Encounters along the Association Reaction between Proteins*, Biophys. J. **96**, 4237-4248, 2009.
31. K. M. ElSawy, L. S. Caves, and R. Twarock, *The impact of viral RNA on the association rates of capsid protein assembly: bacteriophage MS2 as a case study*, J. Mol. Biol. **400**, 935-947, 2010.
32. T. Geyer, C. Gorba, and V. Helms, *Interfacing Brownian Dynamics Simulations*, J. Chem. Phys. **120**, 4573-4580, 2004.
33. P. O. Brunn, V. I. Fabrikant, and T. S. Sankar, *Diffusion through Membranes — Effect of a Nonzero Membrane Thickness*, Q. J. Mech. Appl. Math. **37**, 311-324, 1984.
34. F. Lauck, V. Helms, and T. Geyer, *Graph measures reveal fine structure of complexes forming in multiparticle simulations*, J. Chem. Theor. Comput. **5**, 641-648, 2009.
35. R. Albert, and A.-L. Barabási, *Statistical mechanics of complex networks*, Rev. Mod. Phys. **74**, 47-97, 2002.
36. L. da F. Costa, F. A. Rodrigues, G. Travieso, and P. R. Villas Boas, *Characterization of complex networks: A survey of measurements*, Adv. Phys. **56**, 167-242, 2007.

Algorithmic Rethinking and Code Reengineering for Truly Massively Parallel *ab initio* Molecular Dynamics Simulations

Costas Bekas and Alessandro Curioni

IBM Research - Zurich
Säumerstrasse 4, Rüschlikon, 8803, Switzerland
E-mail: {bek,cur}@zurich.ibm.com

Ab initio molecular dynamics simulations are indispensable tools in the hands of researchers and practitioners, that have pushed the frontiers of knowledge to a large extend. They have been very successful in diverse fields ranging from solid-state physics, to amorphous materials and liquids to biophysics and biochemistry. This power though comes at a staggering computational cost, thus *ab initio* molecular dynamics codes were always at the forefront of High Performance Computing. Indeed, simulations of this kind have benefitted tremendously from the advent of massively parallel machines. In this work we provide concrete examples of the systematic work in algorithmic rethinking and code re-engineering that is required to bring these codes to the next level and render them ready to use machines with millions of threads. We believe the lessons learned to serve as examples for future significant improvements.

1 Introduction

Ab initio molecular dynamics is the combination of first-principles electronic structure methods with molecular dynamics based on Newton's equation of motion. The use of electronic structure methods to calculate the interaction potential between atoms overcomes the main shortcomings of the otherwise highly successful pair potential approach. In particular, with *ab initio* methods many-body effects are included, they are parameter-free, and are able to adjust to new chemical situations that may be encountered during a simulation, for example when chemical reactions or structural phase transitions occur. In their seminal paper[1], Car and Parrinello introduced a new method that allows the efficient propagation of the electronic wave function together with the atomic cores. Although the method is very general, it is primarily used together with the Kohn-Sham approach to density-functional theory. The method has proved to be valuable in many fields. Recent applications include topics in traditional solid-state physics, surface science, interfaces, glasses and amorphous systems, liquids and solutions, catalysis and other chemical reactions, as well as problems from biophysics and biochemistry. For overviews of applications, see recent review papers[2–4]

The combination of a computationally demanding electronic structure method with molecular dynamics requiring thousands of force evaluations render *ab initio* molecular dynamics simulations highly dependent on high-performance computing resources. Many parallel implementations, following various strategies, have been reported in the literature[2,6–12] over the past decade. To be able to push the applications from originally a few atoms to now routinely several hundreds of atoms, it was instrumental to adapt algorithms and implementations to modern massively parallel architectures[2].

Achieving extreme scaleout to massively parallel architectures requires a systematic approach and work. The goal is to eliminate all bottlenecks that appear when we require to scale to millions of threads. It is crucial to say that often, for even moderate number of processors, these parts of the code may account for only a small fraction of the overall runtime. However, as Amdahl's law predicts, they tend to dominate the overall run time at massively parallel runs. On the other hand, it is equally important to design new methods that push the frontiers of scalability for the key kernels of the code as well. We describe perfect examples of this systematic approach in this lecture. In Sec. 2 we analyze a hierarchical parallelization approach for 3D Fast Fourier Transforms that is able to push scalability almost two orders of magnitude further than traditional parallel 3D FFT approaches. Sec. 3 illustrates how it is possible to completely rethink and re-engineer well known but not high performance, algorithms for wavefunction orthogonalization and render them into massively scalable and high performance kernels. In particular, we describe block Gram-Schmidt schemes that BLAS3 based and are particularly suited for the interconnects of modern supercomputers. Finally, in Sec. 4 we show how we can render fully scalable initialization from atomic wavefunctions. This is an example of non-scaling original kernel, to which little (if any) attention was paid to, which however becomes the bottleneck for large scale simulations with thousands of atoms.

A word about the computational platforms. Clearly, large scale simulations require massively parallel machines. We focused on the IBM BlueGene supercomputer series, that with its fast and multimodal interconnects and the overall balanced (and thus scalable) design allows for very fast large scale simulations to become possible. However, we stress that the lessons learned are not limited to this architecture only, but we claim them to be widely applicable. The main reason stems from the strong trends in modern supercomputers for localized networks of higher dimensions (i.e. toruses), multicore nodes and limited memory/network bandwidth per core.

2 Task Groups Strategy for 3D Parallel FFTs

A significant number of popular and highly successful electronic structure codes use a plane wave basis to discretize the Schrödinger equation and thus rely on heavy use of 3D Fast Fourier transforms. Thus, it is natural to spent a lot of efforts in developing highly efficient parallelization strategies for 3D `FFTs`. We describe here a scheme that exploits opportunities for hierarchical parallelism. In particular, the scheme is based on a Task Groups parallelization strategy that concurrently performs several parallel 3D `FFTs`, one per each group of processors. The approach was first implemented in `CPMD`[a],[5] and was later incorporated in Quantum-Espresso (starting from the `CPV` subtree[13]) (several recent similar implementations in other codes exist as well).

In plane wave codes, wavefunctions $\Psi_\rho = [\psi_1, \ldots, \psi_{occ}]$ are expanded in Fourier space, where occ is the total number of valence electrons, which in turn depends on the type and number of the involved atoms. In medium size simulations occ is in the order of several hundreds, while large simulations will push occ to thousands or even tens of

[a] `http://www.cpmd.org`

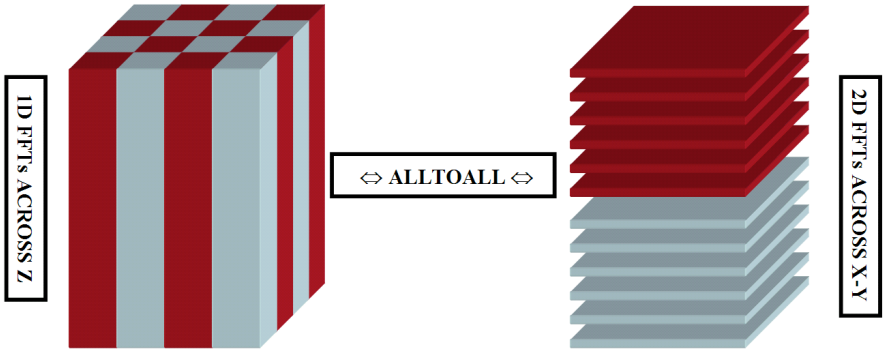


Figure 1. Structure of the standard parallel 3D FFT.

thousands. The charge density $\rho(r)$ at position r in real space is given as

$$\rho(r) = \sum_{i=1}^{occ} |\psi_i(r)|^2. \tag{1}$$

Observe that since the wavefunctions are expanded in Fourier space, computation of charge density in Fourier space would entail doubly nested summations. Instead, it is performed in real space. To this end, the wavefunctions are transformed back to real space by means of inverse 3D FFTs. In the case of P available processors, all of them can be devoted to a single parallel 3D FFT. On the other hand, we can do G, $(G \ll occ)$ parallel 3D FFTs concurrently. We define G groups of processors, each of which works on a single parallel 3D FFT. Thus, the number of loops in computing the charge density is $\lceil occ/G \rceil$ (special handling of the last loop takes care of the case that occ is not divided exactly by G).

Performing one parallel 3D FFT at a time, thus using all available processors, limits scalability. Here is why: The Fourier coefficients of the wavefunctions are organized in a $x - y - z$ 3D mesh in Fourier space. For all wavefunctions, each processor is assigned a number of pencils across the z (vertical) direction. Fig. 1 (left cube) illustrates the case for two processors and one wavefunction. The 3D inverse FFT is performed as follows:

1. 1D inverse FFTs across the z (vertical) direction are computed independently.
2. An all-to-all global communication distributes the results to all processors, so that each processor ends up with a number of complete $x - y$ planes (see right cube of Fig. 1).

3. Then, 2D inverse FFTs are performed independently by each processor without the need for further communication.

It is clear that if the number P of available processors is larger than the number of $x-y$ planes, which is the mesh dimension across the z direction, some processors will get no planes at all. In general, the scalability of this scheme is limited by the largest dimension of the FFT mesh. For parallel architectures with a moderate number of available processors this limitation is not severe as practical runs of *ab initio* codes use hundreds of $x-y$ planes. However, on massively parallel architectures we need to utilize thousands of processors and thus we need a different parallel 3D FFT scheme. Our solution is to exploit opportunities for hierarchical parallelism.

Observe that in order to calculate the charge density $\rho(r)$ by means of (1) we need to iterate through a loop of 3D FFTs, equal to the number *occ* of valence electrons. The Task Groups (TG) strategy will assign different groups of processors to different wavefunctions. Suppose that a processor p_e is empty, in the sense that no $x-y$ planes would be assigned to it if all P processors were to participate in an inverse 3D FFT. Then, since the number of its peers in the group will be P/G we can choose the number of groups G so that a processor will be never empty. We outline the TG scheme in Tab. 1.

The concurrent implementation of G 3D FFTs is organized on a 2D mesh of processors. Each processor belongs to its row group as well as to its column group. Global communications are restrained within these groups. Iteration k performs the 3D FFTs needed for wavefunctions (k-1)*G+i, i=1,...,G. Remember that each processor holds only part of the Fourier coefficients for each wavefunction. Thus, the all-to-all within the row group (line 2) brings to each column group all the Fourier coefficients for the wavefunction assigned to it. For example, at iteration k the $j-th$ processor of the first column group will send its parts of the (k-1)*G+i, i=2,...G wavefunctions, to its row group peers $\mathtt{i} = 2, \ldots, \mathtt{G}$, respectively, while it will receive from them all needed parts for the $(\mathtt{k}-1)*\mathtt{G}+1$ wavefunction. Then, all processors in each column group can perform a parallel 3D FFT (line 3). Finally, the charge density ρ can be accumulated by means of a global reduction across processors in each row group (line 4).

The Task Groups scheme requires additional memory. Remember that each processor holds a part of the wavefunctions coefficients for all eigenvalues. Thus, in order for a column group to work exclusively on a single eigenvalue each processor needs to receive additional wavefunction coefficients from its row group peers. The amount of the extra memory depends upon the number G of Task Groups. There is a tradeoff between the number of available processors P and the number of Task Groups. In order to exploit a large number of available processors we need many Task Groups. On the other hand, this will increase the amount of additional local memory as well as the traffic on the interconnect for the initial all-to-all. However, the 3D FFTs within each column group will also require less communication, since only P/G processors are involved in each column group.

Similar to the calculation of charge density, forces contribution to the orthogonality constraints for the wavefunctions requires a loop of forward 3D FFTs across the occupied states. A parallel 3D FFT is implemented following the same steps as in the inverse transformation in exactly the opposite order: i) Each processor holds a number of complete $x-y$ planes on which it performs 2D FFTs, ii) a global all-to-all assigns to each processor a number of z sticks on which independent 1D FFTs are performed. The

Define a 2D processor array
The number of columns is equal to the number G of Task Groups
The number of rows is equal to the number of processors in each Task Group

1. **DO** $k = 1,\ occ/G$
   2. `all-to-all` communication in row group: brings all needed Fourier coefficients for 3D `FFT`
   3. parallel 3D `FFT` within column group
   4. `allreduce` to accumulate charge density within row group
5. **ENDDO**

Table 1. The TG parallel 3D `FFT` scheme for the calculation of charge density $\rho(r)$.

Task Groups strategy is analogously adopted, so that each column group works on different wavefunctions.

We note that very good scaling for parallel 3D `FFTs` has been achieved by means of the Volumetric `FFT` algorithm[14], which employs distribution of the `FFT` mesh across all three $x - y - z$ directions. However, employing this scheme in out testbed codes would require a major redesigning of the data organization and the corresponding data structures. The Task Groups scheme allows for very good scalability while requiring only minimal changes to the underlying electronic structures code.

Customizing for machines with localized interconnects (i.e. toruses) The decisive parameters in order to select the optimal G involve i) the amount of memory available to each processor core ii) the latency and bandwidth of the dedicated collective communication tree interconnect. For example, on the BlueGene /L machine (which was the first to test the Task Groups strategy), latency of tree traversal was 2.5 μs with 2.5 GB/s bandwidth per link, thus leading to a 23TB/s total binary tree bandwidth (64k machine). It is typical in our practical applications to use 8-32 Task Groups.

2.1 Scalability Experiments

We experimented with a molecular system comprised of 80 water molecules, that represents a problem of intermediate size (240 atoms). The size of the FFT mesh used was 128^3. The number of occupied electrons of the system is $occ = 320$.

The left plot of Fig. 2 illustrates scalability results (total run time) for the calculation of the charge density and the forces contribution to the orthogonality constraints with and without the TG strategy. There are 128 $x - y$ planes across the z direction. Thus, the standard parallel 3D `FFT` implementation scales only up to 128 computing nodes. On the other hand, the TG implementation continues to scale, where we have used 2 Task Groups in the case of 256 computing nodes and 4 Task Groups in the case of 512 computing nodes.

The right plot of Fig. 2 illustrates the percentage, in terms of run time, of the FFT related computation (with TG) compared with the percentage of the orthogonalization related computations. We stress that the latter is dominated by the diagonalization of a dense

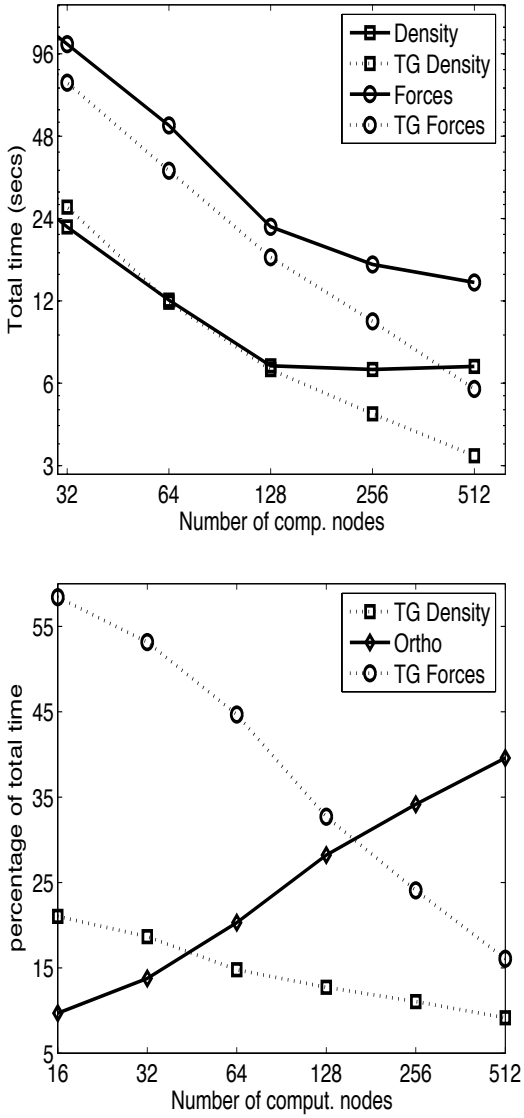


Figure 2. Left: Scalability experiments for charge density and force contribution to the orthogonality constraints. Right: Comparison of total percentages for charge density and force contribution against orthogonalization. The test case is a system of 80 water molecules.

matrix of $occ \times occ$ at each Molecular dynamics step. These constitute the main computational kernels of the application. What remains involves input-output operations and other tasks whose relative load reduces drastically as the size of the simulation increases. It is evident that the improved scaling of the 3D `FFTs` causes the diagonalization to become

dominant in terms of cost: 40% for the orthogonalization while 25% for the 3D FFTs in the case of 512 computing nodes. It is important to note that diagonalization is based on a BLAS 3 implementation that uses a high performance DGEMM library available for the compute nodes.

3 Large Scale Wavefunction Orthogonalization

In large scale electronic structure calculations, that involve thousands of valence electrons, keeping an orthogonal set of wavefunctions starts to dominate the overall cost. In this section we describe recent developments in high performance orthogonalization methods, that allow extreme scalability.

3.1 Orthogonalization by Means of the Cholesky Factorization

Consider the matrix $Q = [q_1, q_2, \ldots, q_k] \in \mathbb{R}^{n\times k}$ the columns q_j of which we wish to orthonormalize. Thus, consider matrix $X \in \mathbb{R}^{k\times k}$ such that:

$$Y = QX, \quad \text{and} \tag{2}$$

$$Y^\top Y = I_k, \tag{3}$$

where I_k is the identity matrix of dimension $k \times k$. From now on we will omit the subscript k and simply use I when the dimensions are clear from the context. Observe that the columns of matrix Y will span the same linear subspace as the columns of X, since the latter are linear combinations of the columns of matrix X. Substituting Eq. 2 into Eq. 3 leads to

$$X^\top Q^\top Q X = I.$$

Thus, if we set $S = Q^\top Q$ to be the "overlap" matrix it is straightforward to see that

$$S = X^{-\top} X^{-1}. \tag{4}$$

Since the overlap matrix S is symmetric positive definite (SPD) we can choose matrix X to be the inverse of the Cholesky factor of S. In other words, let the upper triangular matrix $R \in \mathbb{R}^{k\times k}$ be the Cholesky factor of matrix S

$$S = R^\top R \quad \text{and set} \tag{5}$$

$$X = R^{-1}. \tag{6}$$

Then, we have

$$X^{-\top} X^{-1} = (R^{-1})^{-T} (R^{-1})^{-1} = R^\top R = S. \tag{7}$$

The overall cost of the scheme is straightforward to analyze. The computation of the overlap matrix S induces a cost of $O(nk^2)$, since symmetry allows us to calculate only the upper (or lower) triangular part and every entry requires a dot product calculation that costs $O(2n)$. The Cholesky factorization of matrix S induces a cost $O(k^3/3)$. In order to calculate the final orthonormal matrix Y we need to perform the computation $Y \equiv XR^{-1}$. Inverting matrix R will induce a cubic cost $O(k^3)$ and the final matrix-matrix multiplication XR^{-1} will require $O(2nk^2)$. The total computational cost sums to $O(3nk^2 + k^3)$.

3.1.1 Practical Implementation

In the practical situation of Density Functional Theory (DFT) electronic structure calculations, the columns of matrix $Q = [q_1, q_2, \ldots, q_k]$ hold the coefficients of the expansion of the k occupied wavefunctions on a suitable basis. In typical parallel implementations of DFT codes, such as `CPMD` which is our target code platform, matrix Q is distributed row-wise to the available processors. We denote this by writing

$$Q = \begin{bmatrix} Q_0 \\ Q_1 \\ \vdots \\ Q_{P-1} \end{bmatrix}, \tag{8}$$

where P is the number of the available processors and Q_i are $\lceil n/P \rceil \times k$ blocks. That is, every processor holds a number of consecutive rows of matrix Q. Thus, the computation of the overlap matrix $S = Q^\top Q$ is accomplished as

$$S = \sum_{i=0}^{P-1} Q_i^\top Q_i. \tag{9}$$

Observe that the matrix-matrix $Q_i^\top Q_i$ is local to each processor, and that the summation of the overlap matrix S requires a global reduction operation (`MPI_ALLREDUCE`). Furthermore, the local matrix-matrix multiplications (in particular Rank-k updates) are of `BLAS 3` type (routine `xSYRK`), which insures close to peak processor performance.

In the case that the number of occupied wavefunctions k is small, then the Cholesky factorization $S = R^\top R$ of the overlap matrix S can be performed on each processor independently. In the sequel, the calculation of the orthonormalized wavefunctions Y by means of solving the linear system $R^\top Y^\top = Q^\top$ can also be performed completely independently,

$$R^\top [Y_1^\top, Y_2^\top, \ldots, Y_{P-1}^\top] = [Q_1^\top, Q_2^\top, \ldots, Q_{P-1}^\top], \tag{10}$$

where each processor solves its local linear system $R^\top Y_i^\top = Q_i^\top$. The result is the orthonormal wavefunction matrix Y which is distributed row-wise to the available processors.

However, in the case of large systems that involve thousands of electrons, the number of occupied states k becomes so large that the Cholesky factorization $S = R^\top R$ has to be done in parallel. In fact, the overlap matrix $S = Q^\top Q$ cannot be replicated to all processors, but rather matrix S has to be distributed as well. This is especially crucial in massively parallel platforms, that are equipped with tens of thousands of processing elements, each of which has limited physical memory available. For example, for a system with $k = 10000$, the overlap matrix S alone will easily consume at least 800 `MBytes` of main memory. Furthermore, the computational cost of independently calculating the Cholesky factorization will rise to a staggering 0.33 `TFlop`[b] at each processor.

The choice of distribution of the overlap matrix S to the available processors will be affected by a number of software design and parallel platform parameters.

[b] `TFlop` $= 10^{12}$ floating point operations.

Massively parallel deployment. In the present work we are interested in massively parallel computational platforms, such as the IBM Blue Gene supercomputer series[c]. In this case, the number of available processors P is directly comparable to the number k of occupied states or even significantly larger. Let us proceed with an example from planewave codes such as our target software platform `CPMD`. The number of planewaves n will typically reach several tens of millions for large systems, while the number of occupied states k will be in the order of some thousands. Thus, while the planewave matrices Q and Y can be efficiently distributed to tens of thousands of processors, the same is not meaningful for the overlap matrix S, since we will end up with a distribution where each processor will hold a rather small number of elements of the matrix, or none whatsoever. It is well known that dense linear algebra kernels, such as the Cholesky factorization, are practically impossible to efficiently scale to thousands of processors without the involved matrices becoming adequately large. Thus, one is led to choose a distribution scheme where only a subset of processors will actually hold a part of the overlap matrix S. The parallel Cholesky factorization will have to be computed on this subset of processors.

Software design In implementing a parallel Cholesky factorization one may utilize several different approaches. In the popular `SCALAPACK` library[15], a 2D cyclic block distribution is used. In particular, the active subset of processors is organized into a two dimensional grid and the overlap matrix S is mapped on this grid. This means that each processor in the grid will take a number of consecutive rows and columns. The block size has to be carefully selected so as to maximize both processor performance as well as maximum utilization of the interconnection network of the parallel machine. In all our implementation decisions we have to keep in mind the existing data distribution on all processors. Remember that the wavefunctions (columns of matrix Q) are distributed row-wise. This directly implies that a block column distribution of the overlap matrix S is natural and can be easily implemented. Indeed, the summation (9) can be implemented by blocks. Consider the block partitions

$$S = [S_1, S_2, \ldots, S_p] \tag{11}$$

$$Q_i = [Q_{i,1}, Q_{i,2}, \ldots, Q_{i,p}], \tag{12}$$

where each of the blocks $S_j, Q_{i,j}$ has $\lceil k/p \rceil$ columns and $p < P$ is the number of active processors in the subset that will take part in the Cholesky factorization. Then, each of the blocks S_j is calculated as

$$S_j = \sum_{i=0}^{P-1} Q_{i,j}^{\top} Q_{i,j}, \quad j = 1, \ldots, p$$

and stored only by the $j-$th processor of the active set of p processors. Typically the number of messages required in parallel Cholesky implementations grows as $O(p^2)$, which is one of the main problems in achieving high scalability. Once the Cholesky factorization is computed we then need to solve the linear system $R^{\top} Y^{\top} = X^{\top}$ and since the Cholesky factor R is distributed as well, this will require additional communication.

We have shown that the simple and effective scheme of orthonormalization by means of the Cholesky factorization needs special attention when we consider the study of very

[c] `http://www.research.ibm.com/bluegene`

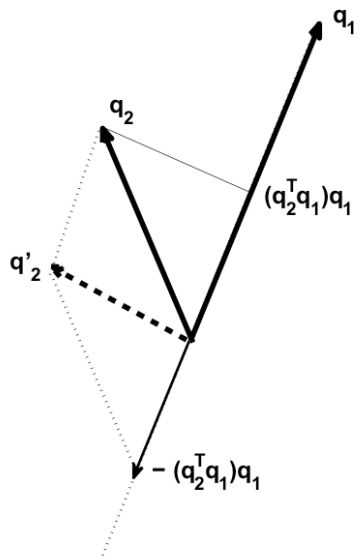


Figure 3. Graphical (geometric) illustration of (standard) Gram-Schmidt.

large systems, with thousands of occupied states, on massively parallel computing platforms. In the sections that follow we describe a method that is based on the well known Gram-Schmidt orthonormalization method that aims to combine the appealing `BLAS 3` characteristics of the Cholesky approach with the simplicity and scalability of the Gram-Schmidt algorithm.

3.2 The Scalar Gram-Schmidt Algorithm

The Gram-Schmidt method is one of the oldest and most popular methods to orthonormalize a set of vectors[16]. The principle behind the method is simply explained in geometrical terms. Consider the vectors q_1, q_2 in $\mathbb{R}^2$ (standard Euclidean 2 dimensional space, see Fig. 3). Let us assume that q_1 is normalized such that $\|q_1\|_2 = 1$. Then, we calculate the projection of vector q_2 on vector q_1. It is clear that the vector $(q_2^\top q_1)q_1$ is the coefficient of vector q_2 towards the direction (or span) of vector q_1. Thus, if we subtract this contribution from vector q_2 we will get a new vector that is clear of any directions towards the direction of vector q_1, i.e. it will be orthogonal to it. The argument readily generalizes for more vectors in higher dimensions. It is also not difficult to see that the new set of vectors spans the same linear subspace, as these new vectors are just linear combinations of the original vectors, with the first one (q_1) being the same. The upper part of Tab. 2 illustrates an algorithmic description of the standard Gram-Schmidt algorithm. Note that the orthonormalization is performed in place, meaning that no additional memory space is required to store the resulting orthonormal vectors.

The Scalar Gram-Schmidt Algorithm

(Input *)* Matrix $Q = [q_1, q_2, \ldots q_k] \in \mathbb{R}^{n \times k}$ the columns q_j of which we would like to orthonormalize

(Output *)* Matrix $Q \in \mathbb{R}^{n \times k}$ such that $Q^\top Q = I$

Standard Gram-Schmidt

1. Set $r_{11} = \|q_1\|_2,\ q_1 = \frac{q_1}{r_{11}}$
2. **for** j=2, ..., k
3. Calculate $r_{ij} =< q_j, q_i >$ for $i = 1, 2, \ldots j-1$
4. $\hat{q}_j = q_j - \sum_{i=1}^{j-1} r_{ij} q_i$
5. $r_{jj} = \|\hat{q}_j\|_2$
6. **if** $r_{jj} == 0$ **then** `stop`, **else** $q_j = \hat{q}_j / r_{ij}$
7. **end**

Modified Gram-Schmidt

1. Set $r_{11} = \|q_1\|_2,\ q_1 = \frac{q_1}{r_{11}}$
2. **for** j=2, ..., k
3. Set $\hat{q} = q_j$
4. **for** i=1 ... j-1
5. $r_{ij} =< \hat{q}, q_i >$
6. $\hat{q} = \hat{q} - r_{ij} q_i$
7. **end**
8. $r_{jj} = \|\hat{q}\|_2$
9. **if** $r_{jj} == 0$ **then** `stop`, **else** $q_j = \hat{q}_j / r_{ij}$
10. **end**

Table 2. The Gram-Schmidt orthogonalization algorithm. Top: the standard version. Bottom: the modified version. With $<, >$ we denote the vector scalar product.

The Modified Gram-Schmidt algorithm Note that in the standard Gram-Schmidt algorithm, the order of calculations in lines 3-4 (Tab. 2) can be interchanged. In particular, line 3 calculates the projection coefficients of the current vector to be orthogonalized against all previous (already orthogonal vectors) and line 4 performs the subtractions of these projections. In the modified Gram-Schmidt algorithm (bottom of Tab. 2) the order of operations just described is changed. In particular, when one projection coefficient has been calculated (i.e. $r_{ij} =< \hat{q}, q_i >$), then the corresponding projected vector is immediately subtracted from the current approximation $\hat{q} = \hat{q} - r_{ij} q_i$. It is not difficult to see that this is mathematically completely equivalent to what is done in standard Gram-Schmidt. However, it is well known than in the environment of floating point calculations modified Gram-Schmidt can yield better numerical accuracy, especially in the situation that two vectors are almost parallel with each other (see for example Ref. 16).

Observe that the computational cost of both variants of Gram-Schmidt is in the order of $O(2nk^2)$ since,

- The cost of computing the projection coefficients is

$$\sum_{j=2}^{k} 2n(j-1) = O(nk^2) \tag{13}$$

- the cost of subtracting the projections is

$$\sum_{j=2}^{k} (n(j-1)+n) = O(nk^2) \tag{14}$$

A careful look in the algorithmic description of Gram-Schmidt (see Tab. 2) shows that if we denote by $R = \{r_{i,j}\}$ the matrix projection coefficients, then matrix R is upper triangular. Furthermore, the matrix product QR, where Q is the orthonormalized set of the original vectors, yields the original set of non-orthonormal columns. This is the well known QR factorization and Gram-Schmidt is one of the several possible ways to obtain it (see Ref. 16).

3.2.1 Practical Implementation Issues

In standard Gram-Schmidt the calculation of the projection coefficients r_{ij} (line 3 of Standard Gram-Schmidt in Tab. 2) can be organized as a `BLAS 2` matrix-vector product. Indeed, we can write

$$\text{Calculate } r_{ij} =< q_j, q_i >,\ i = 1, \ldots, j-1$$

equivalently as

$$r_{1:j-1,j} = Q_{:,1:j-1}^{\top} q_j, \tag{15}$$

where we have adopted `MATLAB` notation in which, : denotes either a complete row or column (depending on its position at the subscript). On the other hand, observe that for modified Gram-Schmidt we are not able to organize calculations as matrix-vector products, since the projection coefficient $r_{i,j}$ depends on the value of the immediately previous projection coefficient $r_{(i-1),j}$. As we will see in the sequel, this property has more perplexed implications when we consider the parallel implementation of the Gram-Schmidt algorithm.

3.2.2 Parallel Implementation and Scalability

Remember that the wavefunction matrix $Q = [q_1, q_2, \ldots, q_k]$ is distributed row-wise to the P available processors. This has an immediate implication on the parallelization scheme of choice. In particular, for standard Gram-Schmidt observe that the matrix-vector product for the calculation of the projection coefficients $r_{i,j}$ can be performed by means of a global reduction. In particular, let $Q_{i,j}, i = 0, \ldots, P-1$ be the part of the $j-$th column of matrix Q that resides on processor i. Then, the calculation of the projection coefficients according to (8) is

$$r_{1:j-1,j} = \sum_{i=0}^{P-1} Q_i^{\top} Q_{i,j}. \tag{16}$$

Method	Comp. per Proc.	Messages	max. size of message
St. G-S	$O(2nk^2/P)$	$O(2k)$	$k-1$
Mod. G-S	$O(2nk^2/P)$	$O(k^2)$	1
Bl. St. G-S	$O(2nk^2/P)$	$O(2k/b)$	$O(b(k-b))$
Bl. Mod. G-S	$O(2nk^2/P)$	$O((k/b)^2)$	$O(b^2)$

Table 3. Computation/communication characteristics of standard, modified and block Gram-Schmidt.

Observe that the matrix vector product $Q_i^\top Q_{i,j}$ is local to each processor. Thus a global reduction (i.e. `MPI_ALLREDUCE`) will ensure that all processors have the final projection coefficients. Then, the subtraction of the projection vectors (line 4 of standard Gram-Schmidt, Tab. 2) can proceed without any communication. The remaining required communication is needed for the calculation of the normalization factor $r_{j,j}$ (line 5), for which a global reduction can be utilized again. Thus, the total number of global reductions required to orthonormalize the k columns of matrix Q is $2k-1$ and the largest size of the message required will be $k-1$ floating point numbers, which corresponds to the vector $r_{:,k}$ of projection coefficients at the last step of the algorithm. Line 2 of Tab. 3 summarizes the computation and communication requirements of parallel standard Gram-Schmidt.

The case of modified Gram-Schmidt is quite different in terms of the communication pattern. The projection coefficient $r_{i,j}$ requires again a global summation since vectors $\hat{q}, q_j$ are distributed row-wise. However, since the current vector $\hat{q}$ is updated at each step of the inner loop (lines 4-7, bottom of Tab. 2), the total number of global reductions is quadratic ($O(k^2)$) in terms of the number of vectors to orthonormalize. On the other hand, the size of the messages is minimal, namely 1 floating point number. The third line in Tab. 3 summarizes the computation/communication characteristics of parallel modified Gram-Schmidt.

In contrast to the parallel implementation of the Cholesky based orthogonalization, both variants of parallel Gram-Schmidt, are able to utilize all of the available processors at all stages of the algorithms. On the other hand, in contrast to Gram-Schmidt, as we saw in Sec. 3.1, the Cholesky approach in orthogonalization, can be implemented entirely in highly optimized `BLAS 3` matrix operations. This has a profound effect in overall processor performance in favor of the Cholesky approach against the Gram-Schmidt algorithm (see Sec. 3.6.

3.3 Block Gram-Schmidt

A natural question arises whether a `BLAS 3` variant of Gram-Schmidt is possible. Consider again the matrix $Q = [q_1, q_2, \ldots, q_k]$ that we wish to orthonormalize. Now, let us partition this matrix column-wise defining a set of block submatrices $B_l \in \mathbb{R}^{n \times b}$, where $l = 1, \ldots, k/b$ and b is the block size, such that

$$Q = [B_1, B_2, \ldots, B_{k/b}]. \tag{17}$$

For simplicity of the discussion we have asserted that the block size b exactly divides the number k of columns of matrix Q. Let us temporarily assume that the columns of the first block B_1 are mutually orthonormal, i.e. $B_1^\top B_1 = I_b$. Then, we seek to orthogonalize all vectors of the next block B_2 against the columns of the block B_1.

The Block Gram-Schmidt Algorithm

(Input *)* Matrix $Q = [B_1, B_2, \ldots B_{k/b}] \in \mathbb{R}^{n \times k}$ that we would like to orthonormalize, block size b

(Output *)* Matrix $Q \in \mathbb{R}^{n \times k}$ such that $Q^\top Q = I$

Standard Gram-Schmidt

1. Compute Cholesky factorization $B_1^\top B_1 = R^\top R$
   if R is singular **then** `stop`, **else** orthonormalize
   $B_1 \equiv B_1 R^{-1}$
2. **for** j=2, …, k/b
3. Calculate $R_{i,j} = B_i^\top B_j$ for $i = 1, 2, \ldots j-1$
4. $\hat{B}_j = B_j - \sum_{i=1}^{j-1} B_i R_{i,j}$
5. Compute Cholesky factorization $\hat{B}_j^\top \hat{B}_j = R^\top R$
   if R is singular **then** `stop`, **else** orthonormalize
   $B_j \equiv \hat{B}_j R^{-1}$
6. **end**

Modified Gram-Schmidt

1. Compute Cholesky factorization $B_1^\top B_1 = R^\top R$
   if R is singular **then** `stop`, **else** orthonormalize
   $B_1 \equiv B_1 R^{-1}$
2. **for** j=2, …, k/b
3. Set $\hat{B} = B_j$
4. **for** i=1 …j-1
5. $R_{i,j} = B_i^\top \hat{B}$
6. $\hat{B} = \hat{B} - B_i R_{ij}$
7. **end**
8. Compute Cholesky factorization $\hat{B}^\top \hat{B} = R^\top R$
   if R is singular **then** `stop`, **else** orthonormalize
   $B_j \equiv \hat{B} R^{-1}$
9. **end**

Table 4. The Block Gram-Schmidt orthogonalization algorithm. Top: the standard version. Bottom: the modified version.

Consider the following block generalization of the scalar Gram-Schmidt projection process

$$\hat{B}_2 = B_2 - B_1(B_1^\top B_2), \tag{18}$$

where the projection matrix $B_1^\top B_2$ is of dimension $b \times b$. Then, it is not difficult to see that

the columns of matrix $\hat{B}_2$ are all orthogonal to the columns of the first block B_1. Indeed,

$$\begin{aligned} B_1^\top \hat{B}_2 &= B_1^\top (B_2 - B_1(B_1^\top B_2)) \\ &= B_1^\top B_2 - (B_1^\top B_1)(B_1^\top B_2) \\ &= B_1^\top B_2 - B_1^\top B_2 \\ &= 0, \end{aligned} \tag{19}$$

where we have exploited the assumption that the fist block B_1 is orthonormal.

In order to orthonormalize this first block we can either use the scalar Gram-Schmidt algorithm, or we can exploit the Cholesky approach as we saw in Sec. 3.1. The same applies for matrix $\hat{B}_2$ that will yield a new orthonormal block B_2. Note that after the orthonormalization of block $\hat{B}_2$ against itself, the block remains orthonormal to the previous block B_1, since the new block B_2 spans the same subspace $\hat{\mathcal{B}}_2$ which is by construction orthogonal to the subspace $\mathcal{B}_1$ spanned by the columns of block B_1. The process is repeated until the last block $B_{k/b}$ is orthonormalized against all previous blocks $B_1, B_2, \ldots$.

In order to obtain a fully `BLAS 3` Gram-Schmidt variant we opt to use the Cholesky approach in orthogonalizing the blocks $\hat{B}_l, l = 1, \ldots, k/b$. Tab. 4 contains an algorithmic description of the new block Gram-Schmidt algorithm. The last two rows of Tab. 3 summarize the computation and communication characteristics of block Gram-Schmidt.

The cost of the block Gram-Schmidt scheme is summarized as follows.

- The cost of calculating the projection coefficients and subtracting the projections (lines 3-4, upper part of Tab. 4) runs in the order $O(2nk^2 - 2nkb)$.
- The cost of orthogonalizing the latest block using the Cholesky approach is $O(3nkb - \frac{7}{3}kb^2)$.

Thus the total cost is $O(2nk^2 + nkb)$. It is interesting to note that this cost approaches the cost of the scalar Gram-Schmidt algorithms for small and medium block sizes b, while is converges to the cost of the Cholesky approach at the extreme case that $b = k$. A similar cost analysis holds for the modified block Gram-Schmidt algorithm.

3.3.1 Practical Implementation Issues

It is clear that we designed both block Gram-Schmidt variants in order to allow their efficient implementation using `BLAS 3` matrix-matrix operations. In particular, following the practice for the scalar standard Gram-Schmidt, the computation of the projection matrices $R_{i,j}$ (see line 3, upper part of Tab. 4), is more efficiently implemented by grouping them in a coarser matrix-matrix multiplication

$$R = B_{1:j-1}^\top B_j. \tag{20}$$

In this case the resulting projection matrix R is of size $(j-1)b \times b$, attaining a maximum size of $(k-b) \times b$ when $j = k/b$. Thus, the significant factors in selecting a proper blocksize b are the following two

- BLAS 3 performance. The complexity of the matrix-matrix multiplication for the calculation of the projection matrix R (see 20) ranges from the minimum $O(nb^2)$ to

the maximum $O(nkb)$. The respective volume of data traffic to main memory would be $O(nb)$ and $O(kn)$, which leads to a constant $O(1/b)$ ratio of memory traffic over computation. Thus, it is clear that as the block size b increases we are getting a better ratio which translates into increased processor performance. Observe that the same analysis holds in the parallel implementation of block standard Gram-Schmidt, since as we will see both the totality of computations and data traffic/storage is equally distributed among all available processors.

- Message size. We analyzed that increasing the block size b benefits processor performance. On the other hand, in a parallel implementation the calculation of the projection matrix R will require a global summation of $O(kb)$ floating point numbers in the worst case. Thus, increasing the block size b too much can potentially over-stress the interconnection network of the machine and thus potentially cause a significant loss of the maximum communication performance. Notice however, that the dependence on the block size b is linear. A well designed balance is required.

For the case of block modified Gram-Schmidt the same line of analysis holds. The main difference is that the projection blocks $R_{i,j}$ (see line 5, bottom of Tab. 4) computed and communicated are of size $b \times b$. Although the dependence of communication load depends quadratically on the block size b, this load is still smaller than the respective on the block standard Gram-Schmidt ($O(kb)$ and typically $k \gg b$). On the other hand, as we will see in the following section, introducing a block in modified Gram-Schmidt is beneficial in terms of number of required messages.

3.4 Parallel Implementation

The parallelization strategy follows the same design lines of the scalar case (see Sec. 3.2.1). Tab. 5 provides an algorithmic description for the parallel block standard Gram-Schmidt algorithm. In particular, the columns of matrix Q are distributed row-wise to the available processors. Thus, the computation of the projection matrices local projection matrices $R_{i,j}$ (line 5) can be performed local to each processor, using a high performance `xGEMM` matrix-matrix multiplication routine, followed by a global reduction (`MPI_ALLREDUCE`) to accumulate the projection coefficients to all processors (line 6). Note that we use a single reduction operation (one call to `MPI_ALLREDUCE`) in order to minimize the number of messages and thus the latency. Then, the subtraction of the projected vectors (line 7, Tab. 5) is performed completely locally, requiring no communication whatsoever. We perform this by using a single call to highly optimized matrix-matrix multiplication routine (`xGEMM`) in order to maximize performance.

Concerning the number of messages (global reductions in this case), it is not difficult to see that since the outer loop does not have k but rather k/b iterations, the total number of messages is reduced accordingly. Tab. 3 depicts the computation/communication profile of all proposed variants.

For the orthonormalization of the current block B_j we opt to employ the `BLAS 3` scheme that is based on the Cholesky factorization (see Sec. 3.1). First the overlap matrix $S_j = \hat{B}^\top \hat{B}$ is computed in parallel (lines 9-10) Observe that in contrast with the original Cholesky based scheme, the size of the overlap matrix S_j is $b \times b$ and does not grow with the number k of the columns to be orthogonalized. Thus, since b will be typically much

The Parallel Block Gram-Schmidt Algorithm

(Input *)*	Matrix $Q = [B_1, B_2, \dots B_{k/b}] \in \mathbb{R}^{n/P \times k}$ that we would like to orthonormalize (distributed to P processors), block size b
(Output *)*	Matrix $Q \in \mathbb{R}^{n/P \times k}$ such that $Q^\top Q = I$ (distributed to P processors)
1.	Compute local overlap matrix $S_l = B_1^\top B_1$
2.	Compute global overlap matrix S_1 be means of `MPI_ALLREDUCE` on local matrices S_l
3.	Compute Cholesky factorization $S_1 = R^\top R$ (local comp.) **if** R is singular **then** `stop`, **else** orthonormalize $B_1 \equiv B_1 R^{-1}$ (local comp.)
4.	**for** j=2, …, k/b
5.	Calculate local projections $R_{i,j}^l = B_i^\top B_j$ for $i = 1, 2, \dots j-1$ as $[B_1, \dots, B_{j-1}]^\top B_j$ by means of `xGEMM`
6.	Calculate global projection matrix $R^{(g)} = [R_{1,j}^\top, \dots, R_{j-1,j}^\top]^\top$ by means of a global reduction `MPI_ALLREDUCE` on $[R_{1,j}^{l^\top}, \dots, R_{j-1,j}^{l^\top}]^\top$
7.	Compute $\hat{B}_j = B_j - [B_1, \dots, B_{j-1}] R_{i,j}$ by means of `xGEMM`
8.	Compute local overlap matrix $S_j^l = \hat{B}_j^\top \hat{B}_j$
9.	Compute global overlap matrix S_j be means of `MPI_ALLREDUCE` on local matrices S_j^l
10.	Compute Cholesky factorization $S_j = R^\top R$ (local comp.)
11.	**if** R is singular **then** `stop`, **else** orthonormalize $B_j \equiv \hat{B}_j R^{-1}$ (local comp.)
12.	**end**

Table 5. Parallel Block Gram-Schmidt orthogonalization algorithm.

smaller than k (i.e. a few hundreds at most) there is no need to distribute it to the processors, but rather the global overlap matrix S_j is replicated. Then, the Cholesky factorization is computed locally inducing a negligible cost $O(b^3/3)$. Finally, the local part of the final orthonormal block B_j is again performed completely locally at each processor, without any required communication (line 11).

We note at this point that the check for the overlap matrix S_j being singular (i.e. lines 3 and 11) is essentially performed by means of the Cholesky factorization. In particular, the Cholesky factorization (for example routine `xPOTRF` from `LAPACK`) will return with an error message, since the Cholesky factor R will have a zero (or very small) entry on its main diagonal.

The parallelization of the block modified Gram-Schmidt method follows the same lines as described above, and we omit its detailed description here for the economy of the paper.

It is interesting to point that unlike the scalar versions of the Gram-Schmidt methods which can be performed almost entirely in place, the block variants require a modest ad-

ditional memory space. Indeed, observe that scalar standard Gram-Schmidt requires $O(k)$ memory to store the projection coefficients $r_{i,j}$ (see Tab. 2) while the modified version requires $O(1)$ additional memory. On the other hand, the block versions of Gram-Schmidt require $O(kb)$ additional memory space ($O(b^2)$ for the modified version) for the projection coefficients, and an additional $O(nb)$ for the solution of the linear systems $\hat{B}_j R^{-1}$ (and the $\hat{B}_j R^{-1}$ for the modified version). Observe that the space for the projection matrices can be reused to store the local Cholesky factors of the overlap matrix $\hat{B}_j^\top \hat{B}_j$ (respectively $\hat{B}^\top \hat{B}$ for the block modified Gram-Schmidt).

It is obvious that the parallelization of the new block Gram-Schmidt schemes relies heavily on the use of collective communication primitives. This is a crucially favorable property that can take great advantage of the high bandwidth, low latency TREE network that is available on the Blue Gene/P Supercomputer.

Finally, it is important to stress that, unlike the parallel Cholesky based approach, in the parallel block Gram-Schmidt algorithm we utilize all of the P available processors. In fact, the the serial part of the new algorithm has complexity in the order of $1/3b^3$. Thus, its percentage over the total cost of the algorithm per processor is:

$$C_s = \frac{1}{6}\frac{Pb^3}{nk^2}. \tag{21}$$

Since the block size b is kept constant and small $(O(100))$, this ratio is small and tends to be even smaller as the size of the problem (n, k) increases, even when we utilize many thousands of processors. For example, setting $P = 10^5, b = 150$ and $k = 60 \times 10^6, k = 4 \times 10^3$ will give a ratio $C_s < 10^{-5}$. Thus, we can expect the algorithm to scale very well to massively parallel platforms.

3.5 Numerical Experiments

In this section we provide several numerical experiments that illustrate the performance of all pre-existing schemes as well as the new block schemes. We measured performance profiles both in serial as well as in massively parallel mode. The computational platform was a Blue Gene/P Supercomputer. Each compute node of this architecture is equipped with a quad core PPC 450 processor at 850 MHz, with 4 `GBytes` of main memory. The theoretical peak performance of each core reaches 3.4 GFLOPS. The largest configuration at our disposal consisted of 8 Blue Gene/P racks, with a total of 32768 compute cores[d]. The top performance we achieved on this system using our new block Gram-Schmidt schemes was 73 TFLOPS which corresponds to 67% of peak performance on 8 Blue Gene/P racks.

3.6 Comparison of Cholesky Orthogonalization with Scalar Gram-Schmidt

We start with a comparison of the scalar Gram-Schmidt schemes against the `BLAS 3` Cholesky based orthogonalization method. Fig. 4 clearly illustrates the superiority of the Cholesky based scheme in terms of performance. We describe two experimental settings. The length of the vectors to be orthonormalized was set to $n = 10000$ and 20000, while

[d] WatsonShaheen system at IBM T. J. Watson Research Center.

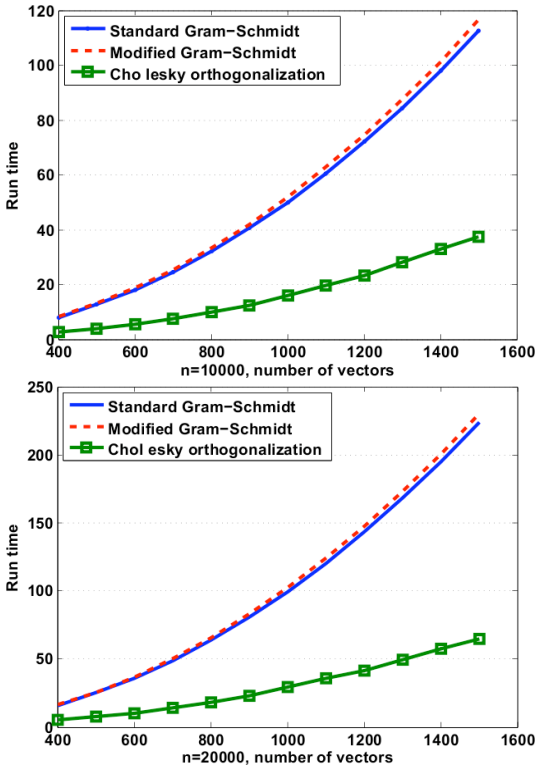


Figure 4. Run time comparison of serial (1 thread) scalar standard and modified Gram-Schmidt with Cholesky orthogonalization for increasing number of vectors $k = 400 : 100 : 1500$. Left: $n = 10000$, Right: $n = 20000$. All times are in seconds.

the number of vectors to be orthonormalized covered the range $k = 400 : 100 : 1500$. The best performance achieved by the scalar Gram-Schmidt schemes was 0.4 GFLOPS while the Cholesky based scheme exceeded 2.2 GFLOPS. In Fig. 5 we analyze the run-time breakdown of the various stages of the Cholesky based method. In particular, we tested with $n = 20000$ and $n = 40000$ while $k = 500 : 100 : 4000$. The left column in Fig. 5 illustrates performance (GLOPFS) while the right column illustrates percentage of total run-time. It is evident that the Cholesky factorization part achieves the lowest performance, while it also is responsible for a small part of the overall run-time. It is exactly this behavior that significantly limits the massively parallel scalability of the Cholesky based scheme.

3.6.1 Serial Block Gram-Schmidt

We now compare the scalar (not parallel) performance of the new block Gram-Schmidt algorithms. In particular, we analyze in detail the performance profile of their various stages and we compare them with the original orthogonalization scheme based on the Cholesky factorization.

Since the block size b is a parameter of crucial importance, we first illustrate its effect

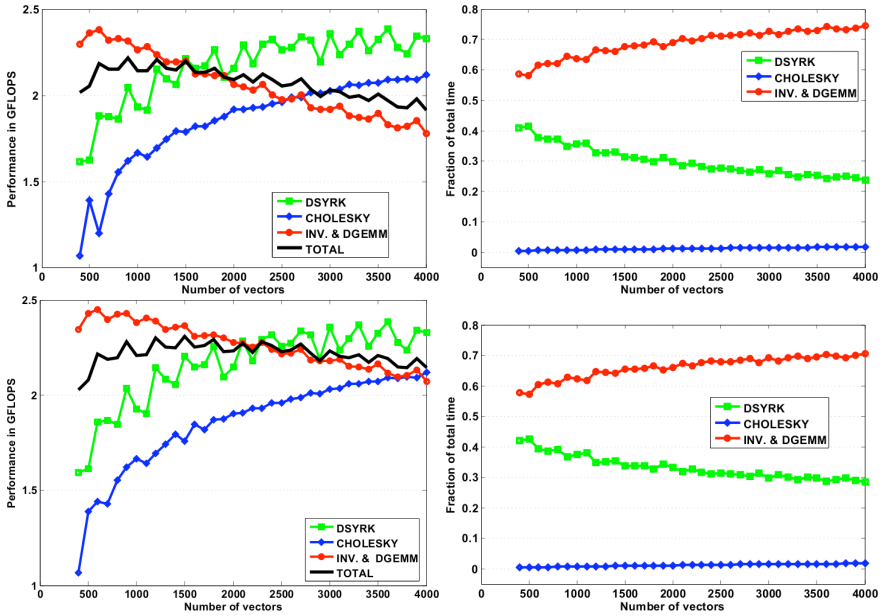


Figure 5. Left column: Performance break up (in GFLOPS) for the various stages of Cholesky orthogonalization (top: $n = 20000$, bottom: $n = 40000$). Right column: respective break up of run time fractions of the corresponding stages.

on the performance of the new scheme. Fig. 6 holds a performance breakup (in GLFOPS) for the various stages of standard block Gram-Schmidt, for $n = 20000$ (top 4) and $n = 40000$ (bottom 4), $k = 400, 1000, 2000$ and 4000, with varying block sizes b. Our first observation is that the matrix-matrix multiplication (`xGEMM`) dominates, as expected, the overall performance of the scheme. Furthermore, the performance profile (its saw like nature being typical in `BLAS 3` performance studies), shows that a selection of block size between 110-160 brings the best performance. It is important to note that best block size b does not depend upon the length n of vectors or the number k of vectors to be orthogonalized. Indeed, the optimal block size b depends upon the underlying processor architecture as well as on the memory hierarchy characteristics, and it can safely be pre-computed before any useful computations take place.

The second important observation that we can draw out of these runs is that the best performance achieved by the block Gram-Schmidt scheme is directly comparable and even better than the performance achieved by the Cholesky based scheme. This is an important finding, that clearly indicates that the new scheme should achieve smaller run times than the Cholesky based scheme in order to orthonormalize the same set of vectors. This is so since the computational complexity of the latter is $O(3nk^2)$ while the complexity of the new scheme is $O(2nk^2)$. Indeed, Fig. 7 illustrates a direct comparison of the standard block Gram-Schmidt scheme against the Cholesky based scheme for $n = 20000$ (top-left) and $n = 40000$ (top-right), using two different block sizes $b = 120$ and $b = 160$. The bottom part illustrates a percentage of run times breakup for the same sizes n and block

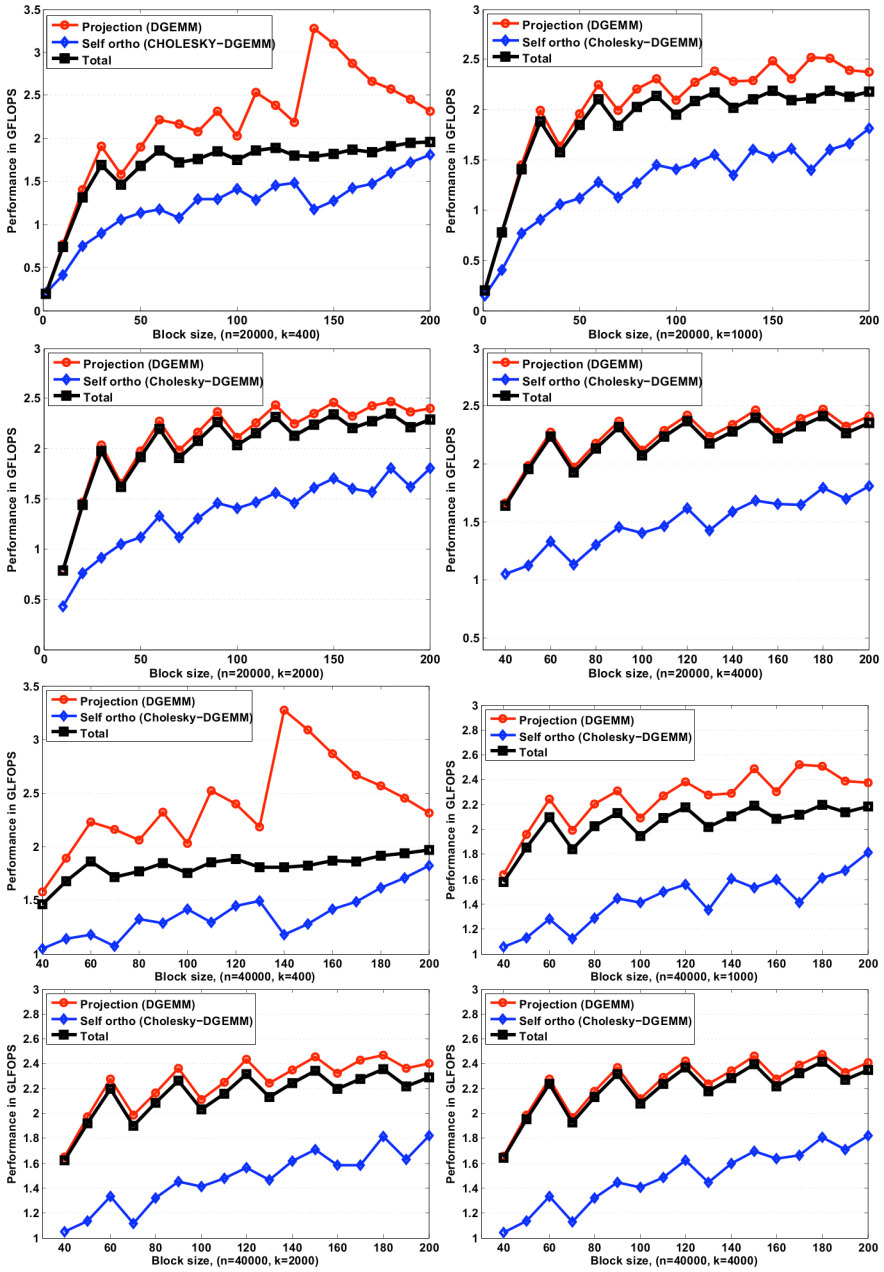


Figure 6. Performance breakup (in GFLOPS) for the various stages of standard block Gram-Schmidt, $k = 400, 1000, 2000, 4000$ and varying block sizes. Top 4: $n = 20000$. Bottom 4: $n = 40000$.

size $b = 120$. It is important to stress the marked difference on the size of the Cholesky part for the cases. In the block Gram-Schmidt scheme, the relevant matrix is always kept

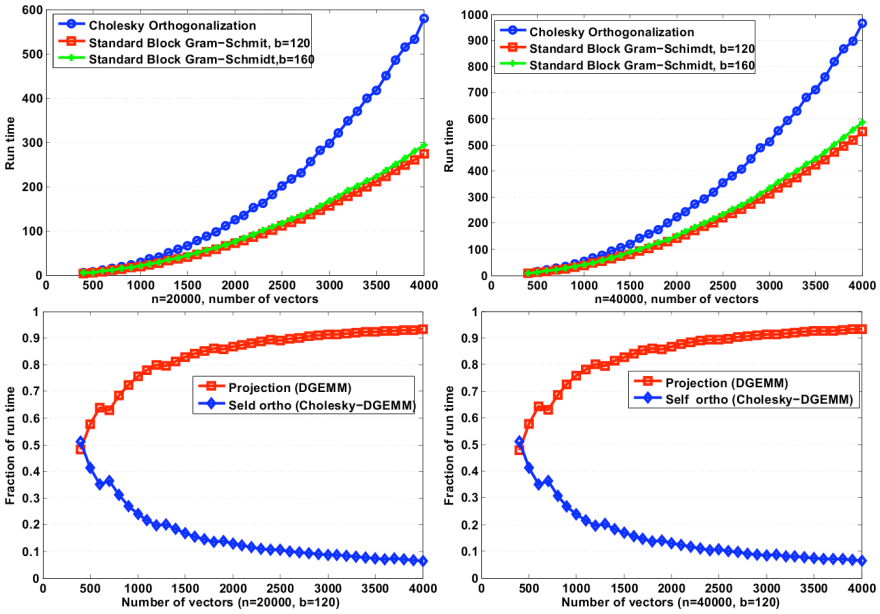


Figure 7. Top: Comparison in run times between Cholesky orthogonalization and standard block Gram-Schmidt (left $n = 20000$, right $n = 40000$) for varying number of vectors to orthogonalize and block sizes $b = 120, 160$. Bottom: Respective breakup of run time fractions for block Gram-Schmidt ($b = 120$) for $n = 20000$ (left) and $n = 40000$ (right).

small $b \times b$, while in the Cholesky based approach it increases with the number of vectors k as $k \times k$.

3.6.2 Symmetric Multiprocessor Scaling on Each Node

The 4 cores of the PPC 450 microprocessor can be used in a symmetric multiprocessor mode. In fact, the Blue Gene/P supercomputer allows for three different modes of parallel execution. The SMP mode, where each compute node hosts an MPI process and each such process can spawn 4 threads of execution. The DUAL mode, where we have 2 MPI processes per node and each process can spawn 2 threads. The Virtual Node (VN) mode where each one of the 4 cores of the PPC 450 chip hosts one MPI process. The top of Fig. 8 illustrates the performance of the block Gram-Schmidt scheme, using the multithreaded version of the ESSL library for the Blue Gene/P. We give both the total performance as well as the performance of the `DGEMM` part. We note that the peak performance of each compute node is 13.6 GFLOPS. It is evident that the block Gram-Schmidt scheme exhibits excellent multithreaded parallel scaling on each node. On the other hand that Cholesky based approach does not achieve such a good scaling, mainly because the Cholesky decomposition is hard to parallelize. In particular, the bottom part of Fig. 8 shows the achieved speedup. Observe that while the scaling of the block Gram-Schmidt scheme is very good, the scaling of the Cholesky approach is poor.

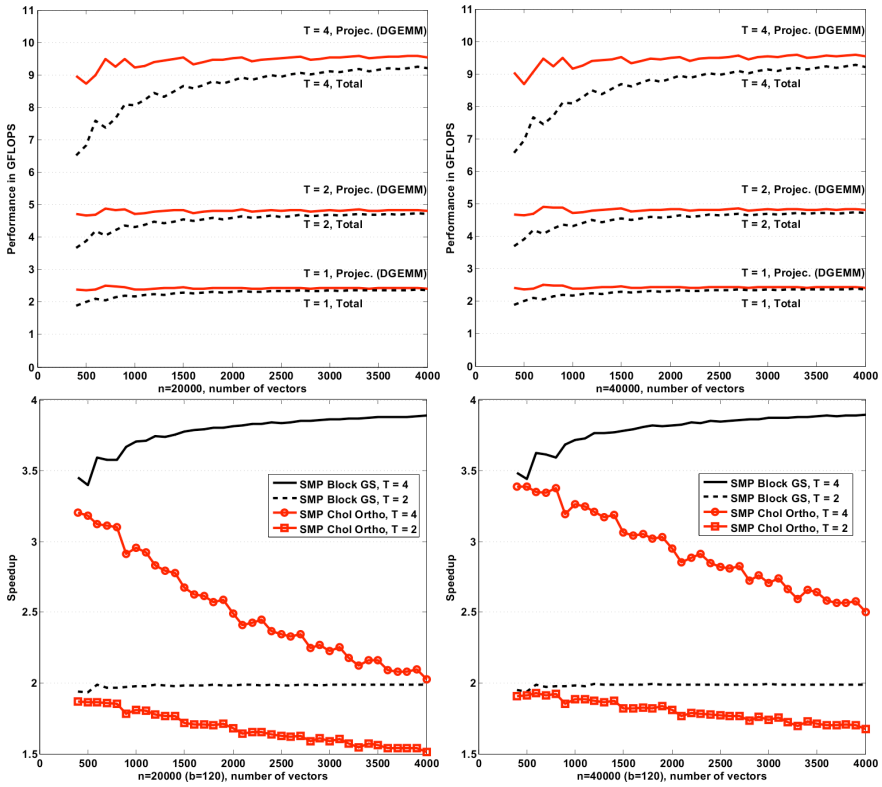


Figure 8. Top: Performance breakdown of SMP block standard Gram-Schmidt using 1,2 and 4 threads (left $n = 20000$, right $n = 40000$) block size $b = 120$. Bottom: Respective speedup comparison for SMP block standard Gram-Schmidt and SMP Cholesky orthogonalization.

3.6.3 Massively Parallel Runs

We now turn our attention to very large parallel runs with the new block Gram-Schmidt schemes. We have chosen to illustrate results representing a smaller ($n = 4 \times 10^6$), an intermediate ($n = 10 \times 10^6$) and a large ($n = 60 \times 10^6$) example. In all cases we experimented with $k = 2000$ and $k = 4000$ number of vectors that we wish to orthonormalize. We utilized up to 8 Blue Gene/P racks which correspond to 111 TFLOPS of theoretical peak performance. All of our runs were performed in SMP mode, where we utilized the multithreaded version of the ESSL library for the Blue Gene/ P compute nodes. Fig. 9 illustrates the run times, using logarithmic scales in both axes. Note that the horizontal axis corresponds to the total number of compute cores used. Thus 32768 cores correspond to 8 BG/P racks. We provide scaling results for both the projection phase of the block Gram-Schmidt algorithm as well as for the orthonormalization of the current block and the overall scaling.

Our first observation concerns the run-time break down of the various phases of the algorithm. As expected we verify in practice the minimal impact of the orthonormalization of the current block (lines 8-11, Tab. 5). It's overall share is in general one order of mag-

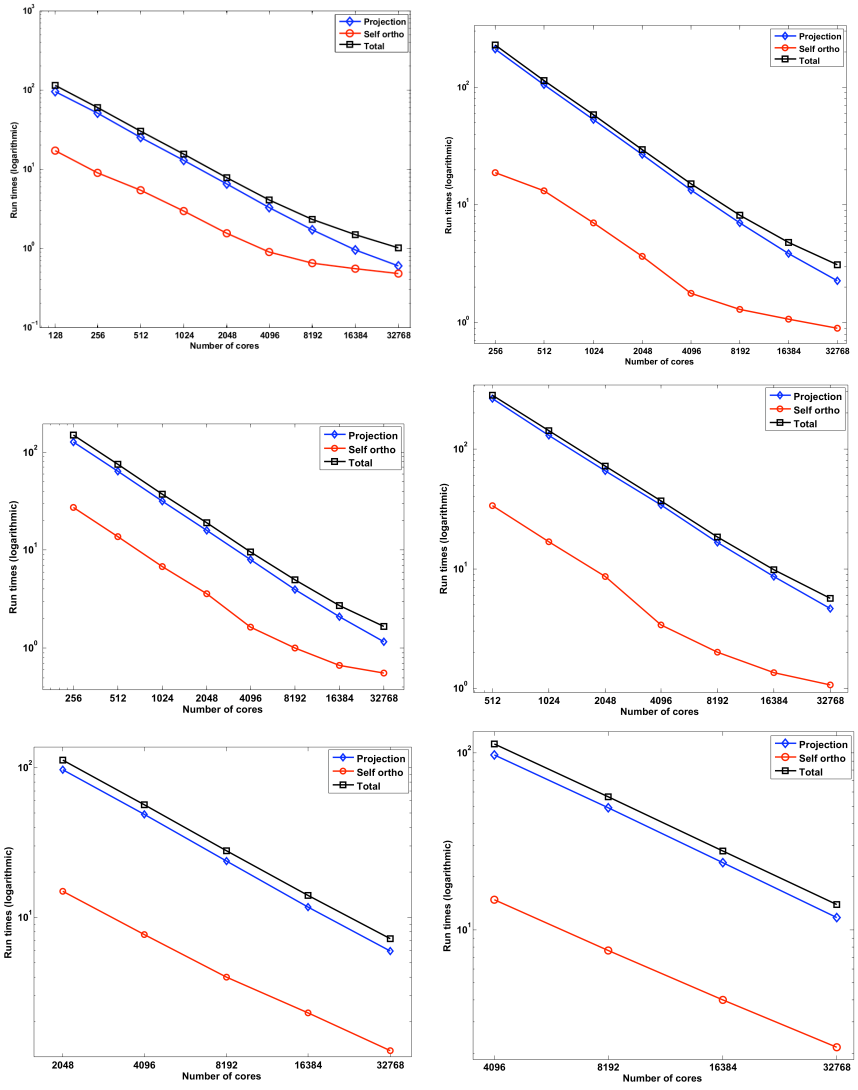


Figure 9. Massively parallel runs with block standard Gram-Schmidt ($b = 150$). Top-left: $n = 4M, k = 2000$, top-right: $n = 4M, k = 4000$. Middle-left $n = 10M, k = 2000$, middle-right: $n = 10M, k = 4000$. Bottom-left: $n = 10M, k = 2000$, bottom-right: $n = 60M, k = 4000$.

nitude less than the projection phase. Observe however, that as the problem size increases even the former exhibits excellent scaling.

Our main observation is that it is clear that the method exhibits very good scaling and performance even for the smaller case ($n = 4 \times 10^6$), which improves with the size of the problem and becomes excellent for the largest case. Indeed, for the case of $n = 60 \times 10^6$, $k = 4000$ and using all 8 Blue Gene/P racks we achieved a performance of 73 TFLOPS which corresponds to 67% of peak performance.

4 Initialization from Atomic Orbitals

Consider a molecular system with N atoms and M valence electrons. Obviously, it is far easier to solve the Kohn-Sham equations for each atom separately, i.e. to define a separate Hamiltonian for each atom of the system and thus producing a set of "atomic" wavefunctions for each atom. Then, we can approximate the solution to the complete problem by superimposing these single atom wavefunctions. Indeed, electronic structure codes make use of precalculated "atomic" wavefunctions for different types of atoms. Thus, from a linear algebra point of view, atomic wavefunctions initialization consists of restricting the full system Hamiltonian operator on a large enough basis of atomic wavefunctions of dimension $k > M$ and then solving for the M smallest eigenvectors of the restricted Hamiltonian.

Formally put, let matrix $W_k \in \mathbb{C}^{n\times k}$ be the expansion of the atomic wave functions on the basis of n plane-waves, where each column of W_k corresponds to a single atom wavefunction. The standard methodology for initialization from atomic wavefunctions proceeds as follows:

1. Compute the restricted Hamiltonian: $\tilde{H}_k = W_k^* H W_k$ and the overlap matrix $O_k = W_k^* W_k$. Matrices $\tilde{H}_k \in \mathbb{C}^{k\times k}$ and $O_k \in \mathbb{C}^{k\times k}$ are both Hermitian.

2. Calculate the eigendecomposition of the restricted generalized Hermitian eigenproblem

$$\tilde{H}_k x = \lambda O_k x. \tag{22}$$

3. Approximate the $M < k$ desired initial wavefunctions as $U_m = W_k X_m$, where the columns of X_m hold eigenvectors that correspond to the m smallest eigenvalues of the restricted generalized eigenproblem (22).

The calculation of the restricted Hamiltonian $\tilde{H}_k$ and of the overlap matrix O_k is performed in parallel since plane-waves are distributed among processors:

1. Each processor: Calculate the application of the Hamiltonian H to its set of plane-waves: HW_k.

2. Each processor: Calculate the overlap $W_k^*(HW_k)$.

3. All processors: Calculate matrix $\tilde{H}_k$ using global summation of $W^*(HW_k)$ among all processors.

In `CPMD` the solution of the generalized eigenproblem (22) was initially not distributed across available processors, but rather solved exclusively on a single processor. We next illustrate that this practice, although perfectly adequate for conventional simulations, is absolutely impractical for next generation target simulations that involve tens of thousands of atoms. Instead, we propose a fully parallel initialization from atomic wavefunctions that is based on the parallel Lanczos algorithm.

4.1 Parallel Initialization Using Lanczos

The dimension k of the restricted Hamiltonian $\tilde{H}_k$ is immediately linked to the number of valence electrons M, and thus with the total number of atoms N. Thus, in the context of large simulations that involve several thousands of atoms, the dimension of the restricted Hamiltonian $\tilde{H}_k$, which is a dense matrix, will be in the order of tens of thousands. Clearly, i) storage requirements, in the order of $\mathcal{O}(k^2)$ ii) as well as computational complexity of the generalized eigenproblem (22), in the order of $\mathcal{O}(k^3)$, render the calculation intractable on a single processor. The following observations are key in the design characteristics of a fully parallel approach:

- Matrices $\tilde{H}_k, O_k$ are dense. Solution of the eigenproblem (22) will require the transformation of these matrices to simpler form. Namely, since this is a Hermitian generalized eigenproblem, it can be transformed to a simple eigenproblem, by means of a Cholesky factorization of the overlap matrix O_k, and then a reduction to tridiagonal form of matrix $O_k^{\dagger}\tilde{H}_k$ is needed, where $O_k^{\dagger}$ is the pseudoinverse of O_k, using the Cholesky factors.
- The calculation of the eigendecomposition of the resulting tridiagonal matrix will require $\mathcal{O}(k^2)$ storage. Thus, it must be done in parallel.
- The new approach should exploit the current distribution of all involved matrices in terms of the distribution of plane-waves across processors.

In light of the above we propose to use the Lanczos algorithm for iterative partial tridiagonalization of a modified restricted Hamiltonian.

4.2 The Lanczos Algorithm

Consider a symmetric matrix A and a starting vector v_1 such that $\|v_1\|_2 = 1$, where $\|.\|_2$ denotes the standard Euclidian norm. The Lanczos algorithm computes an orthonormal basis for the Krylov subspace

$$\mathcal{K}_l(A, v_1) = \mathtt{span}\{v_1, Av_1, A^2v_1, \ldots, A^{l-1}v_1\}. \tag{23}$$

In particular, after l steps of the Lanczos algorithm for matrix A and starting vector v_1, the following Lanczos factorization holds:

$$AV_l = V_lT_l + \beta_{l+1}v_{l+1}e_l^*, \tag{24}$$

where V_l is the orthonormal basis for $\mathcal{K}_l(A, v_1)$ and T_l is a symmetric tridiagonal matrix with structure

$$T_l = \begin{bmatrix} \alpha_1 & \beta_2 & & & \\ \beta_2 & \alpha_2 & \beta_3 & & \\ & \ddots & \ddots & \ddots & \\ & & \beta_{l-1} & \alpha_{l-1} & \beta_l \\ & & & \beta_l & \alpha_l \end{bmatrix}. \tag{25}$$

When l is taken to be equal to the dimension n of matrix A, then matrix T_n will have the same eigenvalues as A. Thus, for $l < n$, the Lanczos algorithm can be viewed as a means for partial reduction to tridiagonal form for matrix A.

4.3 A Preprocessing Step

A preprocessing step that will significantly simplify the whole process is orthogonalization of matrix W_k. Then, the overlap matrix O_k reduces to the identity matrix $O_k = W_k^* W_k = I_k$ and the new restricted eigenproblem becomes a standard one: $H_k x = \lambda x$.

The cost of orthogonalizing W_k is in the order of $\mathcal{O}(nk^2)$. We have adopted the new approach used in `CPMD` (see previous section). For what follows we consider W_k to have orthonormal columns.

4.4 The Distributed Initialization

The working matrix is the new restricted Hamiltonian $W_k^* \tilde{H}_k W_k$ (see the previous section). Then, the proposed fully distributed initialization method proceeds as follows:

1. Calculate a Lanczos factorization for matrix $H_k = W_k^* \tilde{H}_k W_k$ (see Fig. 6)

$$(W_k^* \tilde{H}_k W_k) V_l = V_l T_l + \beta_{l+1} v_{l+1} e_l^*, \tag{26}$$

   where $M < l \leq k$, matrix $T_l \in \mathbb{R}^{l \times l}$ is symmetric tridiagonal and matrix $V_l \in \mathbb{C}^{k \times l}$ has orthonormal columns. The eigenvalues of the restricted Hamiltonian $H_k = W_k^* \tilde{H} W_k$ are approximated by the eigenvalues of matrix T_l. Factorization (26) is calculated by means of the Lanczos algorithm (see Fig. 6), which requires only a matrix-vector product with matrix $(W_k^* H W_k)$. It it clear that this matrix need not be formed, but rather the product $(W_k^* H W_k)y$ with a vector y can be calculated as a series of matrix-vector products (with matrices that are already distributed across processors).

2. Notice however, that if we let the length l of the Lanczos basis V_l approach the size k of the restricted Hamiltonian H_k, then it is preferable to form the restricted Hamiltonian H_k explicitly and distribute it row-wise across the involved processors. This choice greatly simplifies the implementation of the parallel Lanczos algorithm, since the formation of H_k can already be done in parallel in `CPMD`. Furthermore, note that in `CPMD` the application of the Hamiltonian on a vector is fully distributed among all available processors. In this case, the restricted Hamiltonian H_k is distributed row-wise.

3. The basis V_l can be easily distributed row-wise across the available processors, as is the standard approach followed in parallel implementations of the Lanczos algorithm. Note then that, the `xAXPY` operations at lines 3 and 5 of the Lanczos algorithm (see Fig. 6) can be accomplished with no communication whatsoever. The only synchronization points are in line 4 and in line 6, which require global reduction operations.

4. Monitoring the convergence of eigenvalues can be cheaply calculated at every step of Lanczos. The Lanczos iteration is a variational process: approximations to eigenvalues at step $i + 1$ will always be better than those of the previous step. Extremal eigenvalues tend to converge first: thus, one can monitor convergence of the smallest eigenvalue of the tridiagonal matrix T_l (at every step l) and when this has converged

move to monitoring convergence of the next one. Cheap algorithms for the calculation of a few selected eigenvalues of symmetric tridiagonal matrices are available in LAPACK. The cost of computing only the eigenvalues (not the eigenvectors) will be at most a modest $\mathcal{O}(l^2)$ with storage requirements $\mathcal{O}(l)$, since T_l is tridiagonal. Thus, it can be easily achieved on a single processor.

5. When convergence of all M desired (leftmost) eigenvalues has been achieved, these are distributed to the available processors. Then, each processor will do a small number of inverse iteration steps with the exact eigenvalue as shift: $(T_l - \lambda_i I)^{-1} q_i$ on a random starting vector $q_i^{(0)}$. Two to three iterations should be enough to achieve very good convergence to targeted eigenvectors q_i of T_l. Observe that $q_i \in \mathbb{R}^l, l \ll n$, thus storage requirements per processor are kept very small. However, a potential problem with this approach can arise when the converged eigenvalues are closely clustered. Alternatively, we can utilize "divide and conquer" techniques also available through LAPACK expert drivers such as xSYEVX, in which the user can choose to compute particular consecutive eigenvalues/eigenvectors. Thus, we can easily distribute the calculation of wanted eigenpairs on the involved processors. Each processor calculates only a number of consecutive eigenpairs by suitably calling routine xSYEVX. We adopted this latter approach in our current implementation the method in CPMD.

6. The approximate eigenvectors for the modified restricted Hamiltonian are computed as: $\tilde{q}_i = V_l q_i$. Since the basis V_l is distributed row-wise this calculation is performed in parallel. Note that, because a processor holds complete eigenvectors q_i, the calculation of $\tilde{q}_i$ will require a loop of broadcast collectives. Each processor will again end up with complete (consecutive) eigenvectors $\tilde{q}_i$.

7. Finally, approximations to wave functions are similarly (see above) computed in parallel as $x_i = W_k \tilde{q}_i$. Note that matrix W_k is distributed row-wise while processors hold complete consecutive eigenvectors $\tilde{q}_i$.

In Tab. 6 we give an algorithmic outline of the Lanczos method.

4.5 Practical Application of the Parallel Lanczos Algorithm

The plane-wave code CPMD, with single processor initialization, has been demonstrated to achieve excellent scalability on massively parallel systems, consisting of hundreds of thousands of cores[5]. However, it is straightforward to see that in order for the new parallel initialization to scale analogously, a huge number atomic wavefunctions k would be required (i.e. 1 million), which is far beyond our target. Thus, the implementation has to able to utilize only a subset of the available processors. For instance, while all processors contribute in the calculation of the restricted Hamiltonian H_k, only a subset of them will actually be employed in the Lanczos iteration. To this end, matrix H_k is distributed row-wise to these processors. This is facilitated by means of a new MPI communicator for these processors. An additional benefit is that the collective communications will be restricted to only a subset of the machine, thus reducing the overall communication latency.

The matrix-vector operation (line 3) as well as the DAXPY operations (line 3, 5) require no communication. On the other hand, the calculation of scalars α_i, β_i require global reductions (MPI_ALLREDUCE), for which very efficient implementations are available on modern supercomputer interconnection networks.

Lanczos

*(*Input*)* Hamiltonian $\tilde{H}$, orthonormal matrix W_k,
starting vector v_1, $\|v_1\|_2 = 1$, scalar $l \leq k$
*(*Output*)* Orthogonal basis $V_l \in \mathbb{R}^{k \times l}$
unit norm vector v_{l+1} such that $V_l^\top v_{l+1} = 0$

1. Set $\beta_1 = 0$, $v_0 = 0$
2. **for** $i = 1, \ldots, l$
3. $\quad r_i = W_k^*(\tilde{H}(W_k v_i)) - \beta_i v_{i-1}$
4. $\quad \alpha_i =< r_i, v_i >$
5. $\quad r_i = r_i - \alpha_i v_i$
6. $\quad \beta_{i+1} = \|r_i\|_2$
7. $\quad$ **if** $(\beta_{i+1} == 0)$ **then** `stop`
8. $\quad v_{i+1} = r_i / \beta_{i+1}$
9. **end**

Table 6. The Lanczos algorithm for matrix $W_k^* \tilde{H} W_k$. The inner product for vectors is denoted by $< ., . >$.

Reorthogonalization of Lanczos vectors It is well known that although the Lanczos iteration theoretically ensures orthogonality among the basis vectors, in practice the basis vectors quickly loose orthogonality due to roundoff. To remedy this we employ reorthogonalization at each step. We stress though that future versions of `CPMD` will utilize techniques for partial reorthogonalization[17–19], that perform orthogonalization only when it is deemed necessary. Reorthogonalization is performed by means of standard Gram-Schmidt[16]. At each step i, and before calculation of scalar β_{i+1} (see also previous section):

- Compute the local projection coefficient $w_i^l = V_{i-1}^* r_i$. No communication required.
- Compute the global projection coefficients w_i^g by global reduction on the local projection coefficients w_i^l (`MPI_ALLREDUCE`).
- Compute the reorthogonalized $r_i^r = r_i - \sum_{j=1}^{i-1} w_j^g V_j$. No communication required.

Observe that standard Gram-Schmidt (GS) reorthogonalization induces only an additional collective operation per Lanczos step. We note that although modified GS is known to be more stable than standard GS (see Ref. 16), it requires $i - 1$ additional collectives at each step i, which includes $O(l^2)$ additional collectives for a total of l Lanczos steps. Since we are only interested in initial guesses for the wavefunctions we opt to use standard GS. In practical applications so far we have not encountered a problem of unrecoverable severe loss of orthogonality. However, if this happens, we can always temporarily switch to modified GS.

4.6 Numerical Examples

We now illustrate the scalability performance of the parallel Lanczos algorithm for distributed atomic wavefunctions initialization. We experimented with a family of super

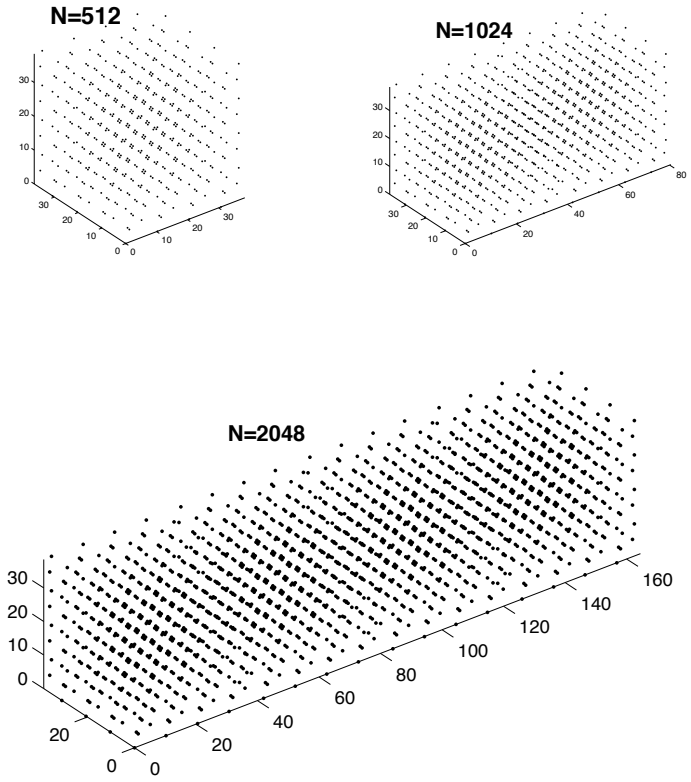


Figure 10. Silicon clusters with N=512, 1024 and 2048 atoms. The last two clusters were generated by replication of the smallest cluster along the X axis.

cells of silicon bulk, ranging $N = 512, 1024$ and 2048 atoms (see Fig. 10). For the first case (N=512) we used a cutoff energy of 20 Rydbergs while for the larger cases of $N = 1024, 2048$ atoms the cutoff energy was set to 12 and 8 Rydbergs respectively. The cutoff energy controls the dimension of the plane-wave basis on which the Hamiltonian and the wavefunctions are expanded (a larger value for the cutoff energy results into more plane-waves).

The smallest case (N=512) is a cubic mesh (with cube edge equal to 41.0449) and the larger two cells were generated by replication of the smallest cell along the X axis. The sizes of the restricted Hamiltonians were $4 \times N =$ 2048, 4096, 8192. `CPMD` utilized in all runs 128 compute nodes of a BG/L system, while for the atomic wavefunctions initialization we utilized a subset of $2^k, k = 1, \ldots, 7$ nodes. We stress that the ability to utilize only a subset of the available compute nodes is crucial in achieving good scaling for the initialization, while other parts of `CPMD` can take advantage of the full set of available processors.

N=512		N=1024	N=2048
#CPUs	time (secs)	time (secs)	time (secs)
2	90	-	-
4	46	381	-
8	23	191	-
16	13	99	738
32	9	54	382
64	8.6	36	211
128	8.7	30	114

Table 7. Run times for distributed initialization using the parallel Lanczos algorithm.

Tab. 7 illustrates run times for the distributed initialization. The dashes in the columns of Tab. 7 indicate that no run was possible for the corresponding number of CPUs because of node memory was not enough to hold the data. Remember that we have implemented the parallel Lanczos algorithm in `CPMD` in such a way that only a subset of the available processors are actually employed for the initialization. This versatility is crucial when we utilize thousands of processors with `CPMD`. It is clear that as the dimension of the restricted Hamiltonian increases, the distributed Lanczos algorithm scales better. For example, while for $N = 512$ scaling stops at 32 processors, for $N = 1024$ it stops at 64 and for $N = 2048$ scaling continues up to 128 processors.

In Tab. 8 we provide detailed run times (in seconds) for the various computational stages of the distributed initialization. `MATVEC` is the time spent multiplying the restricted Hamiltonian by the current Lanczos vector (line 3). `REORTH` is the time spent for re-orthogonalizing the new Lanczos vector. `COLL` is the time for collective communications in the Lanczos loop (excluding the `REORT`). `DAXPY` is the time spent in `BLAS daxpy` operations within the Lanczos loop. `L. EIGS` is the time spent to calculate the Lanczos eigenvectors and eigenvectors (immediately after the Lanczos loop has run). `VECS` is the time spent to multiply the Lanczos eigenvalues with the atomic wavefunctions W_k to get the final estimation of the initial wavefunctions. Remember that matrix W_k is row-wise distributed across all available processors, while the eigenvectors of the restricted Hamiltonian are distributed column-wise across the group of processors that participate in the Lanczos run.

Clearly, the matrix-vector operation (`MATVEC`) scales very well in all cases and the `DAXPY` operations contribute only minimally to the overall cost. The computation of the eigenvalues and eigenvectors of the Lanczos matrix (`L. EIGS`) also scales every well, especially for the larger problems. Reorthogonalization exhibits satisfactory scaling which is attributed to the modest additional communication cost of standard GS and to the very fast collective communication available on the BlueGene machines. Finally, we see that the part that does not scale is the final calculation of the approximate initial wavefunctions. This is expected, since the total number of messages in this part increases as we increase the number of processors in the Lanczos group. To see this, remember that the Lanczos eigenvectors are distributed column-wise to the group of processors that takes part in the Lanczos loop, i.e. each processor in this group holds a number of consecutive Lanczos eigenvectors. On the other hand, remember that matrix W_k is distributed row-wise across

N=512						
#CPUs	MATVEC	REORT	COLL	DAXPY	L. EIGS	VECS
2	36	49.2	1.2	0.07	4.9	1.0
4	18	24.3	1.4	0.05	2.6	1.1
8	7.2	11.9	1.3	0.03	1.4	1.3
16	3	6.2	1.6	0.03	0.87	1.7
32	1.5	3.5	1.8	0.02	0.51	2.2
64	0.8	2.2	1.7	0.02	0.41	3.7
128	0.4	1.3	0.3	0.03	0.5	6.3

N=1024						
#CPUs	MATVEC	REORT	COLL	DAXPY	L. EIGS	VECS
4	141	201	20	0.12	15.1	4
8	7.0	97	9	0.09	9.7	4.4
16	35	48	6.5	0.07	4.8	5.1
32	14	24	5	0.06	2.7	6.9
64	6.5	13	4.8	0.06	1.7	9.6
128	3.2	7.7	1	0.05	1.2	16.8

N=2048						
#CPUs	MATVEC	REORT	COLL	DAXPY	L. EIGS	VECS
16	277	384	33	0.16	35	11.1
32	137	191	25	0.15	16	12.3
64	69	97	19	0.12	9	17.4
128	29	53	5	0.12	5	21.6

Table 8. Detailed run times, in seconds, for the various stages of distributed initialization using the parallel Lanczos algorithm.

all of the available processors. Thus, this part of the calculation involves collectives that span the full breadth of the machine. Indeed, the number of these broadcast collective operations increases as the number of processors in the Lanczos group increases. However, we also note that for the larger case (N=2048) the VECS part is only a fraction of the total cost (roughly %20), which explains the very good scaling of the scheme for this case.

4.7 Discussion

Initialization from atomic wavefunctions in *ab initio* molecular dynamics codes is crucial in order to facilitate large scale next generation simulations with thousands of atoms. This initialization leads to very large dense eigenproblems that are impossible to solve on a single processor, both in terms of computational complexity as well as of memory requirements. In this paper we are reporting a new scheme for distributed initialization that is based on a distributed version of the Lanczos algorithm. Our decision to use Lanczos instead of parallel dense methods such as the ones in SCALAPACK[20] is based on the following key observations:

- We needed to respect as much as possible the existing data structures in `CPMD`, which is the host molecular dynamics code. The distribution of the matrices involved, in terms of plane-waves, significantly favors matrix vector operations, rather than matrix transformations-factorizations that are inherit in parallel dense linear algebra.
- We can safely use in practice standard Gram-Schmidt reorthogonalization, instead of modified Gram-Schmidt, which induces only one additional collective operation at each step of the parallel Lanczos algorithm.
- Our target computational platform is the BlueGene supercomputer series which is equipped with an excellent separate network for collective communications.

We point out that although good scalability of the new scheme is of course a desired property it is not of crucial importance. This is due to the fact that the total run time of large simulations is by far dominated by the minimization of the Kohn-Sham equations (after atomic wavefunction initialization has run) and the subsequent molecular dynamics simulation thereafter. For example, for the case of 1024 Silicon atoms (see previous section) the new distributed initialization scheme on 8 BlueGene /L nodes required 193 seconds while the total run time for minimization was 2400 seconds. Then, each step, out of the thousands, of the molecular dynamics run costs itself roughly the same as the distributed initialization. However, if it was not for this successful initialization from atomic wavefunctions, the first minimization would require an enormous number of iterations in order to converge.

5 Discussion

We have described our efforts towards massively parallel electronic structure calculations. We have focused in efficient parallel 3D FFTs, wavefunction orthogonalization and initialization from atomic wavefunctions. We demonstrate that extreme scaleout is indeed possible, allowing for routine simulations with tens of thousands of atoms in very reasonable time frames. We followed an approach of algorithmic redesign and extensive software reengineering. This means that we adopted algorithms that, although at first looked unsuitable, in fact allowed extreme scaleout. This is proof that we have much to gain by rethinking and adopting basic computational kernels, especially in view of many core nodes, with lower memory and bandwidth per core, that Exascale machines are projected to be composed of.

References

1. R. Car and M. Parrinello. Unified approach for molecular dynamics and density functional theory. Phys. Rev. Lett. 55 (22) 2471-2474, 1985.
2. D. Marx and J. Hutter. Ab-initio molecular dynamics: Theory and implementation. Modern Methods and Algorithms of Quantum Chemistry, J. Grotendorst (Ed.), NIC Series, vol. 1, Forschungszentrum Jülich, Germany, 2000, see also `http://www.fz-juelich.de/nic-series/Volume1`.
3. W. Andreoni and A. Curioni. New advances in chemistry and materials science with CPMD and parallel computing. Parallel Computing 26 (2000) 819-842.

4. J. S. Tse. Ab initio molecular dynamics with density functional theory. Annual Review of Physical Chemistry 53 (2002) 249-290.
5. J. Hutter and A. Curioni. Car-parinello molecular dynamics on massively parallel computers. *Chem. Phys. Chem.*, 6:1788–1793, 2005.
6. K. D. Brommer, B. E. Larson, M. Needels and J. D. Joannopoulos. Implementation of the Car-Parrinello algorithm for ab initio total energy calculations on a massively parallel computer. Computers in Physics 7 (3) (1993) 350-362.
7. L. J. Clarke, I. Štich and M. C. Payne. Large-scale ab initio total energy calculations on parallel computers. Computer Physics Communications 72 (1993) 14-28.
8. J. Wiggs and H. Jónsson. A parallel implementation of the Car-Parrinello method by orbital decomposition. Computer Physics Communications 81 (1994) 1-18.
9. J. Wiggs and H. Jónsson. A hybrid decomposition parallel implementation of the Car-Parrinello method. Computer Physics Communications 87 (1995) 319-340.
10. C. Cavazzoni and G.L. Chiarotti. A parallel and modular deformable cell Car-Parrinello code. Computer Physics Communications 123 (1999) 56-76.
11. R. Gruber, P. Volgers, A. De Vita, M. Stengel and T. M. Tran. Parametrization to tailor commodity clusters to applications. Future Generation Computer Systems 19 (2003) 111-120.
12. J Hutter and A. Curioni. Dual-level parallelism for ab initio molecular dynamics: Reaching teraflop performance with the CPMD code Parallel Computing 31 (2005) 1-17.
13. C. Bekas, A. Curioni and W. Andreoni. New scalability frontiers in ab initio electronic structure calculations using the BG/L supercomputer. In proc., PARA 06, Umea, Sweden, June 2006.
14. M. Eleftheriou, J. E. Moreira, B. G. Fitch and R. S. Germain. A Volumetric FFT for BlueGene/L. *HiPC 2003*, LNCS 2913, 194-203, 2003.
15. L.S. Blackford, J. Choi, A. Cleary, E. D'Azevedo, J. Demmel, I. Dhillon, J. Dongarra, S. Hammarling, G. Henry, A. Petitet, K. Stanley, D. Walker, and R.C. Whaley. *ScaLAPACK User's Guide*. SIAM, Philadelphia, 1997. See also `www.netlib.org/scalapack`.
16. G. H. Golub and C. van Loan. *Matrix Computations*. 3rd Edit., John Hopkins, 1996.
17. C. Bekas, Y. Saad, M. L. Tiago, and J. R. Chelikowsky. Computing charge densities with partially reorthogonalized Lanczos. *Comp. Ph. Com.*, 171(3):175–186, 2005.
18. R. M. Larsen. PROPACK: A software package for the symmetric eigenvalue problem and singular value problems on Lanczos and Lanczos bidiagonalization with partial reorthogonalization, SCCM, Stanford University,
URL: `http://sun.stanford.edu/~rmunk/PROPACK/`.
19. H. D. Simon. Analysis of the symmetric lanczos algorithm with reorthogonalization methods. *Linear Algebra Appl.*, 61:101–132, 1984.
20. J. Demmel, J. Dongarra, R. van der Geijn, and D. Walker. SCALAPACK: Linear algebra software for distributed memory architectures. In T.L. Casavant, P. Tvrdík, and F. Pl'ašil, editors, *Parallel Computers: Theory and practice*, pages 267–282, Los Alamitos, CA, 1995. IEEE Computer Society Press.

Non-Equilibrium Molecular Dynamics for Biomolecular Systems Using Fluctuation Theorems

Gerhard Hummer

Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases
National Institutes of Health, Bethesda, MD 20892-0520, USA
E-mail: gerhard.hummer@nih.gov

Many important molecular processes occur on time scales too long to be sampled efficiently with conventional molecular simulations, or even within the finite time windows accessible to single-molecule experiments. Non-equilibrium methods provide powerful tools to overcome this time scale problem. Remarkably, on the basis of recent advances in non-equilibrium statistical mechanics, many of the important equilibrium properties can be recovered rigorously. In particular, one can obtain accurate estimates of the free energy of states and of molecular free energy surfaces. The physical foundation for these non-equilibrium methods and their practical implementation in computer simulations will be discussed.

1 Introduction

Among the many challenges faced in simulations of complex molecular systems, in particular those encountered in biology, is the problem that the slowest relaxation times are often on the seconds scale and beyond, yet the longest times that one can simulate barely exceed the microsecond range. As a result, many important processes, such as the folding of all but the smallest proteins or the dissociation of a tightly bound ligand, are not accessible to direct molecular simulations. It should not surprise that many single-molecule spectroscopies suffer from related problems. In experiments, the observation time is limited, for instance, by drifts in the mechanical stage or by bleaching of fluorescent dyes.

Just as mechanical force can induce rare transitions at the macroscale, for instance the fracture of a material probe, one can use mechanical perturbations to accelerate transitions at the molecular scale. With the development of atomic force microscopes or laser optical tweezers, it has become possible to manipulate individual molecules while performing precision force and distance measurements on them[1–4]. The mechanical tension caused by the pulling has been used not only to induce molecular transitions such as protein unfolding or the dissociation of a molecular complex, but also to manipulate molecular machines such as molecular motors moving directionally along a track. Initially developed as mimics of these experiment, but immediately recognized as powerful probes of molecular systems in their own right, analogous computer simulation techniques have been used to study molecular processes[5,6]. Both experiments and simulations provide an atomistically detailed picture of the molecular dynamics under force.

The quantitative use of these methods requires careful analyses techniques to recover results that are meaningful and relevant for the dynamics and energetics at equilibrium conditions. By applying external time-dependent perturbations, the speed of conformational changes in a molecular system can be greatly accelerated. However, these external perturbations evidently create non-equilibrium conditions. In the following, I will describe how one can extract free energies of states and molecular free energy surfaces in a for-

mally rigorous manner despite the non-equilibrium sampling. These procedures exploit the Jarzynski identity[7] and its extensions[8] and generalizations, in particular the Crooks fluctuation theorem[9]. After reviewing the underlying theory I will briefly discuss issues of implementation in practical applications, and highlight connections to other simulation methods.

2 Equilibrium Thermodynamics from Non-Equilibrium Simulations and Experiments

2.1 Background: Free Energy Perturbation

A classic problem in molecular free energy calculations, dating back to the seminal work of Born[10], is the charging of an ion in solution. Following Ref. 11, we consider a molecular system of a neutral solute in a solvent, described by a (classical) Hamiltonian energy function $H_0(\mathbf{x})$ with $\mathbf{x}$ a point in phase space. For a charge λe on the solute particle, the Hamiltonian is $H_0(\mathbf{x}) + \lambda V(\mathbf{x})$. The difference in the Helmholtz free energies of the system with charged and uncharged solutes is

$$G(\lambda) - G(0) = -\beta^{-1} \ln \frac{\int e^{-\beta(H_0+\lambda V)} d\mathbf{x}}{\int e^{-\beta H_0} d\mathbf{x}}. \tag{1}$$

where $\beta^{-1} = k_{\mathrm{B}}T$, with k_{B} Boltzmann's constant and T the absolute temperature. The numerator and denominator on the right are the canonical partition functions for the charged and uncharged solute-solvent systems, respectively.

In molecular simulations such free energy differences can be determined in multiple ways. As the basis of the thermodynamic integration formalism, we use Eq. 1 to write the derivative of the free energy difference as $\partial G/\partial\lambda = \langle V\rangle_\lambda$, where the angular brackets indicate a Boltzmann average of the potential V for a system with Hamiltonian $H_0 + \lambda V$. We can then integrate this derivative from $\lambda = 0$ to $\lambda = 1$ to obtain the free energy difference as

$$G(1) - G(0) = \Delta G = \int_0^1 \langle V\rangle_\lambda d\lambda \,. \tag{2}$$

This relation is the basis for the famous Born formula[10] for ion solvation. Alternatively, we can rewrite Eq. 1 as a Boltzmann average over the ensemble of the uncharged solute,

$$G(\lambda) - G(0) = -\beta^{-1} \ln \left\langle e^{-\beta\lambda V} \right\rangle_0 \,, \tag{3}$$

which is Zwanzig's famous free-energy perturbation formula[12].

We can think of Eqs. 2 and 3 as two extreme ways of performing the perturbation. In thermodynamic integration, the perturbation is introduced infinitely slowly, exploiting the relation between the change in free energy and the reversible work, and requiring that a full ensemble average of $\langle V\rangle_\lambda$ is performed at each λ. In free-energy perturbation, the perturbation is applied infinitely rapidly, such that the system has no time to relax and the phase-space average over $\mathbf{x}$ can be performed in place, albeit with an exponential weight factor.

2.2 Free Energies from Non-Equilibrium Dynamics

What happens if we turn on the perturbation at a finite rate? This is the question answered by Jarzynski[7]. He considered perturbations that follow a given path $\lambda(t)$. Systems evolving under the influence of the resulting time-dependent energy function $H_0(\mathbf{x})+\lambda(t)V(\mathbf{x})$ will be driven out of equilibrium, even if they started out perfectly equilibrated at time $t = 0$. Nevertheless, it turns out that there is a rigorous way to determine the free energies from such nonequilibrium trajectories through the Jarzynski identity[7],

$$e^{-\beta[G(\lambda(t))-G(0)]} = \left\langle e^{-\beta W(t)} \right\rangle \,, \tag{4}$$

where $W(t) = \int_0^t \lambda'(\tau)V[\mathbf{x}(\tau)]d\tau$ is the work performed in the charging process and $\lambda'(\tau) = d\lambda/d\tau$. Here, the average $\langle \cdots \rangle$ is over trajectories starting from a Boltzmann equilibrium distribution corresponding to H_0 and evolving in time according to the time-dependent Hamiltonian $H(t) = H_0+\lambda(t)V$. The free energy calculation (or measurement) thus requires an average over paths, i.e., the evaluation of a path integral.

Initially, Jarzynski's identity was greeted with skepticism by many scientists, even expert statistical mechanicians. How could one possibly be able to recover equilibrium thermodynamic properties from non-equilibrium trajectories? In effect, the Second Law of Thermodynamics, which in the present context states that

$$\langle W(t) \rangle \geq 0 \,, \tag{5}$$

has been turned into an equality. But the preceding discussion should have raised some hopes that in Born's and Zwanzig's methods, we have free energy calculation methods at the two extremes of the the kind of process underlying Jarzynski's identity, infinitely slow and infinitely fast. Moreover, we should keep in mind that Jarzynski's identity does not inform us about the nature of nonequilibrium states. Instead, we can use it to recover the equilibrium ensemble from infinitely many fast transformations, instead of one infinitely slow one.

In the following, I will sketch two proofs, one establishing a connection to quantum mechanics and the other leading to an important generalization, the Crooks fluctuation theorem.

2.3 Relation of Jarzynski's Identity to the Feynman-Kac Theorem for Path Integrals

The average in Eq. 4 involves a path integral. One of the most famous relations for path integrals is the Feynman-Kac theorem, providing a foundation for the path-integral formulation of quantum mechanics. It turns out that Jarzynski's identity is closely related to, and in some sense a direct consequence of, the Feynman-Kac theorem for path integrals.

To motivate the Feynman-Kac theorem, we use the case of chemical kinetics. Consider a system with two states, say "folded" and "unfolded," that interconvert according to first-order chemical kinetics,

$$U \underset{k_u}{\overset{k_f}{\rightleftharpoons}} F \tag{6}$$

such that the relative populations evolve in time according to

$$\dot{F}(t) = -k_u F(t) + k_f U(t) \tag{7}$$

$$\dot{U}(t) = k_u F(t) - k_f U(t) \tag{8}$$

with $F(t)+U(t) = 1$. Now consider further that a probe is attached to our molecules. This probe can be excited by a laser pulse at time zero, and the excitation quenches at a rate k_q in the folded state, but not in the unfolded state. Such experiments are indeed performed, for instance, to measure the rate of contact formation in protein folding. We then have the modified rate scheme

$$U \underset{k_u}{\overset{k_f}{\rightleftharpoons}} F \overset{k_q}{\rightarrow} Q \; . \tag{9}$$

In this trivial example, we could of course simply solve the modified kinetic equations, with Eq. 7 changed to

$$\dot{F}(t) = -k_u F(t) + k_f U(t) - k_q F(t) \; , \tag{10}$$

to account for the irreversible quenching and the growth of the quenched population, $\dot{Q}(t) = k_q F(t)$.

However, we could also simulate this kinetic system to estimate the population $S(t) = F(t) + U(t) = 1 - Q(t)$ that have not yet quenched. There are two extreme forms of such simulations. In the direct non-equilibrium simulation, we would start out by choosing initial states F or U according to the prescribed initial condition, evolve them in time according to the rules of chemical kinetics by making transitions between the connected states F, U, and Q, and then stop individual trajectories as soon as state Q is reached for the first time. From the statistics of the times of repeated simulations of quenching events, we would extract $S(t)$ and $Q(t)$.

Alternatively, we can also perform an equilibrium simulation of our original system, Eq. 9, without any irreversible quenching. That is, starting from an initial state F or U, the trajectory would simply hop back and forth between the two states, with waiting times drawn from the appropriate exponential distributions. In this equilibrium sampling, we can incorporate the quenching simply by reweighting. We ask ourselves: given that we had excited our molecule at time zero, what is the probability that the excitation has not been quenched up to time τ? For an individual trajectory, this weight is simply $\exp\left[-\int_0^\tau k(t)dt\right]$, where $k(t) = 0$ if the system is in state U at time t and $k(t) = k_q$ if it is in state F. The quantity in the exponent is the product of the time spent in state F up to time t, multiplied by k_q. We now need to average over many such equilibrium trajectories, reweighted to describe the required non-equilibrium quenching kinetics,

$$S(t) = \left\langle \exp\left[-\int_0^t k(\tau)d\tau\right]\right\rangle \; , \tag{11}$$

where the average is over trajectories evolving according to Eqs. 7 and 8. This reweighting of a set of trajectories evolving under one evolution operator to account for the dynamics under a different evolution operator is one application of the Feynman-Kac theorem for path integrals. This procedure is the basis not only of path integrals in quantum mechanics, but also many other important relations, including the theory of spectral line shapes in condensed phase.

To return to free energy calculation, we assume that the dynamics of our system is described by a time-dependent Liouville-type operator $\mathcal{L}_t$ that preserves the Boltzmann

distribution at equal time,

$$\mathcal{L}_t e^{-\beta H(\mathbf{x},t)} = 0 \,, \tag{12}$$

where H is a Hamiltonian that depends explicitly on time because of an external perturbation, such as the application of a time-varying mechanical force. Examples of $\mathcal{L}_t$ include the Liouville operator of classical mechanics or the the Fokker-Planck operator for diffusion. We now define a phase-space density normalized at time 0 but not at later times,

$$f(\mathbf{x},t) = \frac{e^{-\beta H(\mathbf{x},t)}}{\int e^{-\beta H(\mathbf{x}',0)} d\mathbf{x}'} \,. \tag{13}$$

If we differentiate f with respect to time, we obtain

$$\frac{\partial f(\mathbf{x},t)}{\partial t} = \mathcal{L}_t f(\mathbf{x},t) - \beta \frac{\partial H(\mathbf{x},t)}{\partial t} f(\mathbf{x},t) \tag{14}$$

which has a form similar to Eq. 10. Here we used Eq. 12. We now apply the Feynman-Kac theorem in this more general case to express f in terms of a trajectory average[8],

$$f(\mathbf{x},t) = \frac{e^{-\beta H(\mathbf{x},t)}}{\int e^{-\beta H(\mathbf{x}',0)} d\mathbf{x}'} = \left\langle \delta(\mathbf{x} - \mathbf{x}(t)) e^{-\beta \int_0^t \frac{\partial H}{\partial \tau}[\mathbf{x}(\tau),\tau] d\tau} \right\rangle \tag{15}$$

where the $\delta(\mathbf{x})$ is Dirac's delta function, and the expectation value is over trajectories starting from an equilibrium Boltzmann distribution at time 0 and evolving according to $\mathcal{L}_t$. By integration over phase space $\mathbf{x}$, the left hand side becomes the ratio of partition functions, and the right hand side becomes an expectation value of the Boltzmann-weighted work. We thus recover Jarzynski's identity[7]

$$e^{-\beta \Delta G(t)} = \left\langle e^{-\beta \int_0^t \frac{\partial H}{\partial \tau}[\mathbf{x}(\tau),\tau] d\tau} \right\rangle = \left\langle e^{-\beta W(t)} \right\rangle \,, \tag{16}$$

where $\Delta G(t) = G(t) - G(0)$ is the difference in free energy between the states corresponding to the Hamiltonians at times t and 0. This celebrated theorem relates the difference between the free energies corresponding to the Hamiltonians at two different times to the average of the Boltzmann factor $e^{-\beta W(t)}$ of the external work done on the system.

2.4 Crooks Fluctuation Theorem

It should not surprise that by combining both forward and reverse perturbations, one obtains better estimates of the free energy. In our example of charging an ion in solution, we could also perform the reverse processes of uncharging. Even before the development of fluctuation theorems, the work values $W(t)$ and $\underline{W}(t)$ accumulated on the forward and reverse processes provided us with upper and lower bounds on the free energy through the Second Law,

$$-\langle \underline{W}(t) \rangle \leq \Delta G \leq \langle W(t) \rangle \,, \tag{17}$$

where the reverse protocol samples trajectories according to $\underline{\lambda}(\tau) = \lambda(t - \tau)$.

The Crooks fluctuation theorem[13,9] provides us with a remarkable and powerful exact result going beyond these bounds. Crooks showed that the normalized probability densities of work values from forward and reverse paths are related to each other

$$\frac{p_f[W = W(t)]}{p_r[W = -\underline{W}(t)]} = e^{\beta[W - \Delta G(t)]} \,. \tag{18}$$

Jarzynski's identity follows from the Crooks fluctuation theorem by integration over W, noting that the probability densities are normalized. As it turns out, this relation is useful for both simulations and experiment[14]. In particular, it allows us to use Bennett's optimal estimator[15,16] for the free energy.

In the following, a derivation of the Crooks fluctuation theorem is sketched[17] that is based on the path-sampling approach used originally[13]. We consider a discrete trajectory $\mathbf{x}_0 \overset{H_1}{\to} \mathbf{x}_1 \overset{H_2}{\to} \dots \overset{H_N}{\to} \mathbf{x}_N$ in phase space. At each step $\mathbf{x}_{i-1} \overset{H_i}{\to} \mathbf{x}_i$, the new phase point is chosen according to the new Hamiltonian $H_i = H(\mathbf{x}, t_i)$, for instance by using Metropolis Monte Carlo sampling (but good Newtonian or Langevin dynamics integrators in discrete time steps would follow the same rules). Each of the steps is Markovian in full phase space (i.e., it does not depend explicitly on the preceding path). As a result, the probability P_f of the entire trajectory can be factorized:

$$P_f(\mathbf{x}_0 \overset{H_1}{\to} \mathbf{x}_1 \overset{H_2}{\to} \dots \overset{H_N}{\to} \mathbf{x}_N) = p_0(\mathbf{x}_0) \prod_{i=1}^{N} p_i(\mathbf{x}_i|\mathbf{x}_{i-1}) \; . \tag{19}$$

The initial probability $p_0(\mathbf{x}_0) = \exp[-\beta[H(\mathbf{x}_0, t_0) - G(t_0)]$ is the normalized equilibrium Boltzmann distribution at time t_0. The $p_i(\mathbf{x}_i|\mathbf{x}_{i-1})$ are transition probabilities from $\mathbf{x}_{i-1}$ to $\mathbf{x}_i$ under the influence of the Hamiltonian H_i. If the transition probabilities satisfy detailed balance (as they do, say, in Metropolis Monte Carlo dynamics),

$$\frac{p_i(\mathbf{x}_i|\mathbf{x}_{i-1})}{p_i(\mathbf{x}_{i-1}|\mathbf{x}_i)} = e^{-\beta[H_i(\mathbf{x}_i) - H_i(\mathbf{x}_{i-1})]} \tag{20}$$

then we end up with a relation between the the probabilities P_f and P_r of the forward and time-reversed paths,

$$\begin{aligned} \frac{P_f}{P_r} &= \frac{P(\mathbf{x}_0 \overset{H_1}{\to} \mathbf{x}_1 \overset{H_2}{\to} \dots \overset{H_N}{\to} \mathbf{x}_N)}{P(\mathbf{x}_N \overset{H_N}{\to} \mathbf{x}_{N-1} \overset{H_{N-1}}{\to} \dots \overset{H_1}{\to} \mathbf{x}_0)} = \frac{p_0(\mathbf{x}_0) \prod_{i=1}^{N} p_i(\mathbf{x}_i|\mathbf{x}_{i-1})}{p_N(\mathbf{x}_N) \prod_{i=1}^{N} p_i(\mathbf{x}_{i-1}|\mathbf{x}_i)} \\ &= \exp\left(\beta \sum_{i=0}^{N-1} [H_{i+1}(\mathbf{x}_i) - H_i(\mathbf{x}_i)] - \beta[G(t_N) - G(0)] \right) = e^{\beta[W(t_N) - \Delta G]} \end{aligned} \tag{21}$$

Here $\Delta G = G(t_N) - G(t_0)$ and the work $W(t)$ is the accumulated change in the energy. In essence, Eq. 21 generalizes detailed balance to non-equilibrium trajectories by establishing an exact relation of the path probabilities (or the action) of forward path and reverse paths. The Crooks fluctuation theorem follows from Eq. 21 by averaging over the ensembles of forward and reverse paths, and collecting histograms of the work values from these averages.

2.5 Free Energy Surfaces

In molecular simulations, and also in single-molecule experiments, one is often interested less in the free energy of the system as a whole, and more in the free energy as function of a particular coordinate, say, $q = q(\mathbf{x})$. These free energy surfaces $G(q)$ are often referred to as potentials of mean force (PMFs) because their negative gradient is the mean force,

$-\partial G/\partial q = -\langle \partial H/\partial q \rangle$. Many efficient techniques have been developed to determine the PMF defined as

$$e^{-\beta G_0(q)} = \langle \delta[q - q(\mathbf{x})] \rangle_0 = \frac{\int \delta[q - q(\mathbf{x})] e^{-\beta H_0(\mathbf{x})} d\mathbf{x}}{\int e^{-\beta H_0(\mathbf{x})} d\mathbf{x}} \tag{22}$$

up to a constant. In particular, umbrella sampling employs harmonic biasing functions to enhance the local sampling along q and combines the results to obtain a global surface. Umbrella sampling can be generalized to the kinds of non-equilibrium protocols employed in experiments and simulations[8]. If a harmonic spring acts on q, with a potential $V(q,t) = k_s(q - z(t))^2/2$ whose center is moved according to a prescribed protocol $z(t)$, we can use Eq. 15 above, integrate out all the degrees of freedom except q, and obtain[8]

$$e^{-\beta G_0(q)} = \left\langle \delta\left[q - q\left(\mathbf{x}(t)\right)\right] e^{-\beta\left(\int_0^t \frac{\partial V}{\partial \tau}[q(\mathbf{x}(\tau)),\tau] d\tau - V(q[\mathbf{x}(t)],t)\right)} \right\rangle \tag{23}$$

where $\langle \cdots \rangle$ is again an average over all trajectories started from a Boltzmann distribution corresponding to the initial Hamiltonian $H(\mathbf{x},0) = H_0(\mathbf{x}) + V[q(\mathbf{x}),0]$ and evolving according to $H(\mathbf{x},t)$.

2.6 Use and Implementation

Force-probe molecular dynamics[5] and steered molecular dynamics[6] are powerful simulation analogs of single-molecule pulling experiments. In these and related approaches, forces are applied to one or several atoms to induce a molecular transition. To implement the above procedures for the determination of system free energies and free energy surfaces, it is important that these forces arise from a well-defined potential function, which may be time dependent. Many simulation packages already contain code to perform simulations in the presence of such time-dependent perturbations.

In planning the simulations, one should, whenever possible, attempt to sample both the forward and the reverse processes. Indeed, the more challenging of these processes (vaguely speaking, the one with the wider work distribution) tends to be more informative with respect to the free energy[18]. In combination, more accurate results can be obtained. In addition, the problem of large systematic biases in the exponential estimator in Eq. 4 can be avoided.

Whereas the construction of free energies is more or less straightforward for states defined by a control parameter, such as the location of the pulling spring $z(t)$, it is more challenging to obtain free energy surfaces. As in the equilibrium analogs, one can use histogram reweighting techniques both for one-sided perturbations[8] and for forward-and-reverse perturbations[19].

Critical for the success is the choice of the coordinate q. Whereas in an experiment, one is limited by the ways one can attach the molecules of interest to the pulling apparatus, one has considerable choice in simulations. For instance, if one is interested in the release of a ligand from a buried binding site, the distance between the centers of mass of the two binding partners may be far from an optimal pulling coordinate. With the apparatus built up above we can understand why: the amount of dissipated work, $\langle W(t) \rangle - \Delta G(t)$, depends on the pulling protocol and on the pulling coordinate. We would like to find protocols and coordinates that minimize the amount of dissipation, such that the dynamics remains as close as possible to the equilibrium dynamics. In our example of ligand dissociation, we may thus want to include the motion of a lid occluding the binding site or any other

obstacles explicitly in a generalized pulling coordinate. In general, we can adopt methods of assessing and optimizing reaction coordinates (see, e.g., Ref. 20) for this process.

As a final word of caution, it is important to realize that the one-sided exponential estimators of the free energies are biased (e.g., Eq. 4). In situations where the distribution of work values is broad as compared to the thermal energy k_BT, this bias can significantly distort the results. In essence, a few trajectories with low work values dominate the sample. These problems can be minimized by using two-sided estimators that combine results from forward and reverse pulling, in particular Bennett's acceptance ratio[15,16] in combination with the Crooks fluctuation theorem[9] and its extension to free energy surfaces[19].

Additional practical issues and implementation questions are discussed, for instance, in Ref. 17. Theoretical questions are discussed in Ref. 11, and in particular in Ref. 21. A broader perspective on fluctuation theorems can be found, for instance, in Refs. 22,23. The connection to experiment is explored in Ref. 24. An overview of applications can be found, for instance, in Ref. 25. Recent experimental applications can be found in Refs. 14, 26, 27.

3 Concluding Remarks

Jarzynski's identity[7], the Crooks fluctuation theorem[9], and their extensions to free energy surfaces[8,19] provide powerful tools to determine free energies from non-equilibrium simulations and experiments alike.

Additional non-equilibrium methods have been developed and used with considerable success. In the construction of diffusion, master-equation, or Markov state models[28–30] it is possible to restrict the non-equilibrium to the choice of initial condition, but use an otherwise unbiased time evolution. With detailed balance and time-reversal symmetry of the time propagation, one can in this way build up a coarse-grained representation of the dynamics that retains the exact equilibrium populations as a steady state. In addition, adaptive free energy sampling methods fall into the broader category of non-equilibrium methods, such as Wang-Landau sampling[31], metadynamics[32], and adaptive biasing force[33].

An important question is: are non-equilibrium pulling methods more efficient than fully optimized equilibrium sampling methods? Based on theoretical analyses[34,35], the answer is likely no. But, going back to the opening argument, that may not be entirely discouraging, not only because non-equilibrium methods have already proved particularly powerful in an initial search for possible reaction mechanisms. On the one hand, the non-equilibrium methods allow us to induce reactions over the time scales accessible to a simulation or an experiment. On the other hand, they have opened up a window into non-equilibrium ensembles and their poorly understood properties.

Acknowledgments

Dr. Attila Szabo is gratefully acknowledged for many discussions and collaborations. This work was supported by the Intramural Programs of the NIDDK, NIH.

References

1. E. L. Florin, V. T. Moy, and H. E. Gaub, *Adhesion Forces Between Individual Ligand-Receptor Pairs*, Science, **264**, no. 5157, 415–417, 1994.

2. T. T. Perkins, D. E. Smith, and S. Chu, *Direct Observation of Tube-Like Motion of a Single Polymer Chain*, Science, **264**, no. 5160, 819–822, 1994.
3. S. B. Smith, Y. J. Cui, and C. Bustamante, *Overstretching B-DNA. The Elastic Response of Individual Double-Stranded and Single-Stranded DNA Molecules*, Science, **271**, no. 5250, 795–799, 1996.
4. M. Rief, M. Gautel, F. Oesterhelt, J. M. Fernandez, and H. E. Gaub, *Reversible Unfolding of Individual Titin Immunoglobulin Domains by AFM*, Science, **276**, no. 5315, 1109–1112, 1997.
5. H. Grubmüller, B. Heymann, and P. Tavan, *Ligand Binding Molecular Mechanics Calculation of the Streptavidin Biotin Rupture Force*, Science, **271**, no. 5251, 997–999, 1996.
6. B. Isralewitz, S. Izrailev, and K. Schulten, *Binding Pathway of Retinal to Bacterioopsin. A Prediction by Molecular Dynamics Simulations*, Biophys. J., **73**, no. 6, 2972–2979, 1997.
7. C. Jarzynski, *Nonequilibrium Equality for Free Energy Differences*, Phys. Rev. Lett., **78**, no. 14, 2690–2693, 1997.
8. G. Hummer and A. Szabo, *Free Energy Reconstruction from Nonequilibrium Single-Molecule Pulling Experiments*, Proc. Natl. Acad. Sci. USA, **98**, no. 7, 3658–3661, 2001.
9. G. E. Crooks, *Path-Ensemble Averages in Systems Driven Far from Equilibrium*, Phys. Rev. E, **61**, no. 3, 2361–2366, 2000.
10. M. Born, *Volumen und Hydratationswärme der Ionen*, Z. Phys., **1**, 45–48, 1920.
11. G. Hummer and A. Szabo, *Free Energy Surfaces from Single-Molecule Force Spectroscopy*, Acc. Chem. Res., **38**, 504–513, 2005.
12. R. W. Zwanzig, *High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases*, J. Chem. Phys., **22**, 1420–1426, 1954.
13. G. E. Crooks, *Nonequilibrium Measurements of Free Energy Differences for Microscopically Reversible Markovian Systems*, J. Stat. Phys., **90**, no. 5-6, 1481–1487, 1998.
14. D. Collin, F. Ritort, C. Jarzynski, S. B. Smith, I. Tinoco, and C. Bustamante, *Verification of the Crooks Fluctuation Theorem and Recovery of RNA Folding Free Energies*, Nature, **437**, no. 7056, 231–234, 2005.
15. C. H. Bennett, *Efficient Estimation of Free Energy Differences from Monte Carlo Data*, J. Comput. Phys., **22**, 245–268, 1976.
16. M. R. Shirts, E. Bair, G. Hooker, and V. S. Pande, *Equilibrium Free Energies from Nonequilibrium Measurements Using Maximum-Likelihood Methods*, Phys. Rev. Lett., **91**, no. 14, 140601, 2003.
17. G. Hummer, "Nonequilibrium methods for equilibrium free energy calculations", in: Free Energy Calculations. Theory and Applications in Chemistry and Biology, C. Chipot and A. Pohorille, (Eds.), chapter 5, pp. 171–198. Springer, New York, 2007.
18. C. Jarzynski, *Rare Events and the Convergence of Exponentially Averaged Work Values*, Phys. Rev. E, **73**, no. 4, 046105, 2006.
19. D. D. L. Minh and A. B. Adib, *Optimized Free Energies from Bidirectional Single-Molecule Force Spectroscopy*, Phys. Rev. Lett., **100**, no. 18, 180602, 2008.
20. R. B. Best and G. Hummer, *Reaction Coordinates and Rates from Transition Paths*, Proc. Natl. Acad. Sci. U.S.A., **102**, 6732–6737, 2005.

21. T. Lelièvre, M. Rousset, and G. Stoltz, *Free Energy Computations. A Mathematical Perspective*, Imperial College Press, London, 2010.
22. C. Jarzynski, *Equalities and Inequalities Irreversibility and the Second Law of Thermodynamics at the Nanoscale*, Annu. Rev. Cond. Matt. Phys., **2**, 329–351, 2011.
23. D. J. Evans and D. J. Searles, *The Fluctuation Theorem*, Advances Phys., **51**, no. 7, 1529–1585, 2002.
24. G. Hummer and A. Szabo, "Thermodynamics and kinetics from single-molecule force spectroscopy", in: Theory and Evaluation of Single-Molecule Signals, E. Barkai, F.L.H. Brown, M. Orrit, and H. Yang, (Eds.), chapter 5, pp. 139–180. World Scientific, Singapore, 2008.
25. M. Sotomayor and K. Schulten, *Single-Molecule Experiments in vitro and in silico*, Science, **316**, no. 5828, 1144–1148, 2007.
26. J. Liphardt, S. Dumont, S. B. Smith, I. Tinoco, and C. Bustamante, *Equilibrium Information from Nonequilibrium Measurements in an Experimental Test of Jarzynski's Equality*, Science, **296**, no. 5574, 1832–1835, 2002.
27. A. N. Gupta, A. Vincent, K. Neupane, H. Yu, F. Wang, and M. T. Woodside, *Experimental Validation of Free-Energy-Landscape Reconstruction from Non-Equilibrium Single-Molecule Force Spectroscopy Measurements*, Nature Phys., **7**, no. 8, 631–634, 2011.
28. G. Hummer and I. G. Kevrekidis, *Coarse Molecular Dynamics of a Peptide Fragment Free Energy Kinetics and Long-Time Dynamics Computations*, J. Chem. Phys., **118**, no. 23, 10762–10773, 2003.
29. S. Sriraman, L. G. Kevrekidis, and G. Hummer, *Coarse Master Equation from Bayesian Analysis of Replica Molecular Dynamics Simulations*, J. Phys. Chem. B, **109**, no. 14, 6479–6484, 2005.
30. X. H. Huang, G. R. Bowman, S. Bacallado, and V. S. Pande, *Rapid Equilibrium Sampling Initiated from Nonequilibrium Data*, Proc. Natl. Acad. Sci. U.S.A., **106**, no. 47, 19765–19769, 2009.
31. F. G. Wang and D. P. Landau, *Efficient Multiple-Range Random Walk Algorithm to Calculate the Density of States*, Phys. Rev. Lett., **86**, no. 10, 2050–2053, 2001.
32. A. Laio and M. Parrinello, *Escaping Free-Energy Minima*, Proc. Natl. Acad. Sci. U.S.A., **99**, no. 20, 12562–12566, 2002.
33. E. Darve and A. Pohorille, *Calculating Free Energies Using Average Force*, J. Chem. Phys., **115**, no. 20, 9169–9183, 2001.
34. G. Hummer, *Fast-Growth Thermodynamic Integration Error and Efficiency Analysis*, J. Chem. Phys., **114**, no. 17, 7330–7337, 2001.
35. H. Oberhofer, C. Dellago, and P. L. Geissler, *Biased Sampling of Nonequilibrium Trajectories. Can Fast Switching Simulations Outperform Conventional Free Energy Calculation Methods?*, J. Phys. Chem. B, **109**, no. 14, 6902–6915, 2005.

Multigrid QM/MM Approaches in *ab initio* Molecular Dynamics

Teodoro Laino

Mathematical and Computational Sciences
IBM Zurich Research Laboratory, CH-8803 Rüschlikon, Switzerland
E-mail: teo@zurich.ibm.com

In this manuscript I will review the problem of optimizing the algorithm for the construction of the QM/MM electrostatic potential. In fact, while the cost of solving the Schrödinger equation in the QM part is the bottleneck of these calculations, evaluating the Coulomb interaction between the QM and the MM part can be surprisingly expensive: in some cases, it can be just as time-consuming as solving the QM part. The result of this investigation, performed during my PhD thesis, is a new real space multi-grid approach which handles Coulomb interactions very effectively. This algorithm has been implemented in the CP2K code[1]. This novel scheme cuts the cost of this part of the calculation by two orders of magnitude with respect to other schemes[2], at similar accuracies. The method does not need fine tuning or adjustable parameters and it is quite accurate, leading to a dynamics with very good energy conservation. Moreover it provides a natural extension to QM/MM periodic boundary conditions, treating properly the long-range electrostatic interactions in hybrid QM/MM simulations. In the last part of this manuscript, I will show the importance of a correct treatment of the long–range interactions in QM/MM calculations with periodic boundary conditions (PBC), when studying highly ordered crystal structures.

In summary, the present multigrid method allows for fast, accurate and both periodic and non-periodic simulations in molecular simulations of biological and material science systems.

1 Introduction

The study of chemical reactions in condensed phases is computationally demanding, owing not only to the size of the simulating system but also to the large degree of configurational sampling necessary to characterize a chemical reaction. This places severe demands on the efficiency of the implementation of any QM/MM scheme. Two main bottlenecks can be identified in such calculations: one concerns the evaluation of the QM energy and derivatives while the other is associated with the evaluation of the electrostatic interaction between the QM and the MM part. In this respect we can identify two classes of codes, those based on Gaussian-type orbitals (GTOs) to represent both the wave-function and the charge density[3,4] and those using grids in real space to represent the charge density[5,1,6]. The latter encompasses both codes fully based on plane waves (PWs) and the more recent mixed approaches based on Gaussian plane waves (GPWs).

For localized basis sets (GTOs), the use of an efficient pre-screening technique is imperative in order to avoid the quadratic construction of the one-electron QM/MM Hamiltonian matrix. For non-local basis sets (PWs), if the interaction is evaluated analytically, the computational price is proportional to the number of grid points times the number of MM atoms. Surprisingly the evaluation of the QM/MM electrostatic interaction, for the latter scheme, requires between 20% and 100% of the time needed by the QM calculation, this in spite of the use of sophisticated hierarchical multipole (HMP) methods[7] or of clever

implementations based on electrostatic cutoffs[2]. Furthermore these techniques require a fine tuning of parameters to yield optimal performance, and lead to a loss of accuracy that makes error control difficult.

I will describe in the next sections a recent implementation of the QM/MM coupling term that avoids the use of any hierarchical method or multipole technique, based on the use of multi-grid techniques in conjunction with the representation of the Coulomb potential through a sum of functions with different cutoffs, derived from the new Gaussian expansion of the electrostatic potential (GEEP for short). The overall speedup is of 1-2 orders of magnitude with respect to other PW-based implementations of the QM/MM coupling Hamiltonian[8,2]. The lack of tuning parameters and electrostatic cutoffs makes this implementation a totally free parameter scheme, without any significant loss of accuracy. Consequently, very stable simulations can be obtained with optimal energy conservation properties.

Another important issue is the treatment of long–range Coulomb interactions, much less well established for hybrid quantum mechanics/molecular mechanics (QM/MM) than for classical simulations. So far, most of the QM/MM implementations have relied on a spherical truncation scheme, in which the solute(QM)-solvent(MM) electrostatic interactions are neglected beyond a certain cutoff distance R_c. There are only a few exceptions. A very popular and inexpensive approach is the reaction field method, which couples the spherical truncation scheme with a polarizable continuum medium that extends beyond a cutoff distance R_c[9–14] . Ewald's lattice summation technique were also investigated to treat the long-range QM/MM electrostatic interactions. Within a semi-empirical framework, the first implementation is due to Gao and Alhambra[15]. In their scheme only the long-range QM/MM interactions are evaluated, while the QM/QM ones are omitted. For the particular set of applications tested by these authors, namely solvation phenomena, the solute-solvent (QM/MM) interactions were considered as the determining ones. Recent implementations of Ewald techniques extended to the full QM/MM long-range interactions[16,17] show indeed that even for solvation cases long-range QM/QM electrostatic interactions play a significant role. Within a self-consistent DFT scheme, to the best of my knowledge there is only one QM/MM scheme that allows PBC[18] to be used. This approach is conceptually similar to the one we present here and it relies on the use of splines in reciprocal space (k-space), optimally designed for use within plane wave (PW) codes. The present multigrid approach is on the other hand, based on real space techniques, and is designed to be used with Gaussian basis codes, as is CP2K[1].

Afterwards, I will propose an extension of the QM/MM algorithm to applications where the use of PBC is mandatory. It relies on the most efficient methods for calculating long-range electrostatic interactions of point charges within PBC and scales linearly with respect to the number of MM atoms. Moreover the evaluation of the MM electrostatic potential using PBC is independent of the number of QM atoms, depending only on the dimension of the coarsest grid used in the multi-grid approach. Both schemes (periodic and non–periodic) will be shown on a realistic system: the modeling of silica.

Nonetheless, before explaining in details the multi-grid framework, it is necessary a little digression about the renormalization of the QM/MM Hamiltonian. In fact, the development of an appropriate coupling Hamiltonian between the two subsystems is the biggest challenge in such hybrid methods[2,19,20]. The point-like description of the MM-atom charges and their interactions with the QM electrons at short ranges can cause an

artificial and non physical polarization of the QM electron density[2,19,21]. Such an artificial polarization can influence the outcome of a chemical reaction study, the dipole moment, and other properties based on electronic charge density[2,19,21].

In fact, the point-charge description for the MM atom cannot provide a compatible picture for the QM/MM Coulomb interaction at distances close to zero and this can be a source of non physical polarization of QM electrons and divergent forces on the MM atoms. To remove this divergence, arising from a point-like charge description of the MM atom, an empirical description of a renormalized Coulomb potential was introduced[2,21]. Recently[22] a theoretical derivation based on a localized partial-wave expansion of the MM charge was proposed, adjusting the extension of the charge distribution in order to normalize the Coulomb potential near interatomic separations of the order of twice the covalent radius.

In the following sections, I will first review the renormalization of the QM/MM interaction potential[22]. Then, I will present the multigrid approach both for periodic[42] and non–periodic[41] systems. Finally, I will report on the modeling of Silica with a QM/MM Hamiltonian[72]. Latin letters a, b will be used to index the MM atoms, while Greek letters α, β will be used for QM atoms.

2 Renormalization of the QM/MM Hamiltonian

The central issue of a QM/MM hybridization scheme is the definition of the QM/MM coupling part $\mathcal{H}_{QM/MM}$[19,23,24,20]. $\mathcal{H}_{QM/MM}$ accounts for the interaction between the quantum system and the MM atoms. In general, $\mathcal{H}_{QM/MM}$ contains Coulomb (long-range) and short-range interactions (non–bonded) and is taken as[19,23,24,20]

$$\mathcal{H}_{QM/MM} = \sum_{a\in MM} \int d\mathbf{r} \frac{\rho_e(\mathbf{r},\mathbf{r}_\alpha) q_a}{\mathbf{r}-\mathbf{r}_a} + \sum_{a\in MM}\sum_{\alpha\in QM} \frac{Z_\alpha q_a}{|\mathbf{r}_\alpha - \mathbf{r}_a|} + \sum_{a\in MM}\sum_{\alpha\in QM} V_{\text{NB}}^{a\alpha}(\mathbf{r}_\alpha,\mathbf{r}_a) \quad (1)$$

where $\mathbf{r}$, $\mathbf{r}_\alpha$ and $\mathbf{r}_a$ represent the position vector for electrons, QM nuclei with charge Z_α and MM nuclei with atom partial charge q_a, respectively. ρ_e represents the electron density. The short-range repulsion and attractive mutual average polarization is modeled by a general non-bonded term ($V_{\text{NB}}^{a\alpha}(\mathbf{r}_\alpha,\mathbf{r}_a)$), like for instance a Lennard-Jones (LJ) potential[25] or a BKS interaction[26].

If an interaction Hamiltonian like Eq. 1 is used, artifacts may arise due to the presence of unscreened Coulomb charges of the MM atoms. This effect is ultimately due to the absence of Pauli exclusion repulsion for the QM electrons by the MM atoms. The atom included in the MM subsystem should exert Pauli repulsion due to its own electrons (which are replaced together with the nuclear charge by an effective point charge) and would deter the QM electrons to penetrate the atom valence shell. In a purely classical force field calculation, the $\frac{1}{|\mathbf{r}_\alpha - \mathbf{r}_a|}$ term of the Lennard-Jones potential[25] takes into account this effect and provides sufficient repulsion between atoms at short range, thus keeping the attractively interacting MM atoms at appropriate separations. For QM theories, the Pauli exclusion repulsion is incorporated either properly antisymmetrizing the electronic wave function or by

employing an exclusion hole concept (for methods with DFT origin). Nevertheless, incorporating the Pauli exclusion repulsion between the QM electronic charge distribution and the MM point charges in a hybrid QM/MM calculation remains a formidable challenge. One idea is to seek a comprehensive description of the QM/MM Coulomb interaction considering a localized expansion of the charges which regularizes the potential at short range while reduces to the Coulomb potential for larger distances ($\mathbf{r} >> 2\mathrm{r}_c$). As the charge reflects the overall electrostatic potential acting at a point in the configuration space, it accounts for the Pauli exclusion effect too. However, this conjecture is valid only beyond a certain radius and not at short distances where the notion of point charge looses its validity. Thus it is customary to regularize the potential at these short distances without affecting its value for distances grater than r_c.

Earlier Eichinger *et al.*[27] and recently Das *et al.*[21] proposed to replace the MM point charge with a Gaussian delocalized charge density to remedy the short-range artifact. They used a multistep approach to evaluate the Coulomb interaction between an MM atom and the quantum system. The Coulomb part of their hybrid QM/MM Hamiltonian is given by

$$\mathcal{H}_{\rho_e,q_a} = \int d^3r \rho_e(\mathbf{r},\mathbf{r}_\alpha) q_a \frac{\mathrm{Erf}(|\mathbf{r}-\mathbf{r}_a|/\sigma)}{|\mathbf{r}-\mathbf{r}_a|} \tag{2}$$

Here $\rho_e(\mathbf{r},\mathbf{r}_\alpha)$ is the electronic charge density, Erf is the error function, $\mathbf{r}_a$ is the position of the a^{th} MM atom and the value of σ is the same for all atoms (0.8Å). As the error function (which integrates the Gaussian distribution over a certain radius) asymptotically reaches the value of unity, the above function has the correct asymptotic behavior of the Coulomb interaction at large distance. At short distances, the error function is less than unity and it tends to zero as distance goes to zero, thus removing the discontinuity in the QM/MM interaction potential. We compare the functional behavior of this form of the potential vis-a-vis the pure Coulomb interaction in Fig. 1. It appears that the potential does not saturate near twice the covalent radius of the atom, which is supposed to be a key issue in the modeling of the Coulomb QM/MM interaction. Afterwards, Laio *et al.*[2] introduced another functional form that takes into account the short range effect with the Coulomb potential saturating near the covalent radius of the MM atom. The Coulomb part of their hybrid QM/MM Hamiltonian is given by

$$\mathcal{H}_{\rho_e,q_a} = \int d^3r \rho_e(\mathbf{r},\mathbf{r}_\alpha) q_a \frac{r_{c,a}^n - \tilde{\mathbf{r}}^n}{r_{c,a}^{n+1} - \tilde{\mathbf{r}}^{n+1}} \tag{3}$$

where $\tilde{\mathbf{r}} \equiv |\mathbf{r}-\mathbf{r}_a|$. In the above prescription, the usual Coulomb interaction of $\frac{1}{\tilde{\mathbf{r}}}$ is being replaced by $v(\tilde{\mathbf{r}}) = \frac{r_{c,a}^n - \tilde{\mathbf{r}}^n}{r_{c,a}^(n+1) - \tilde{\mathbf{r}}^(n+1)}$. This functional form also has the correct asymptotic behavior of $\frac{1}{\tilde{\mathbf{r}}}$ and as $\tilde{\mathbf{r}} \to 0$, it smoothly converges to $\frac{1}{r_{c,a}}$. In Fig. 1 we show the behavior of the potential v($\tilde{\mathbf{r}}$) for $r_{c,a} = 0.699 a.u.$ (0.37 Å). This corresponds to the electrostatic potential of a QM electron with a unit positive charge. The functional form, although appears very useful for QM/MM electrostatic interactions, has not been derived theoretically and thus may be considered as empirical. The functional forms of Eichinger *et al.*[27] and Laio *et al.*[2] mentioned above reduce both the attractive and repulsive Coulomb interactions at short distances while having the correct asymptotic behavior. Another crucial aspect of these prescriptions is that they lead to zero forces (finite potential) at very short ranges, thus avoiding the artificial localization of the electronic charge density on a positive MM

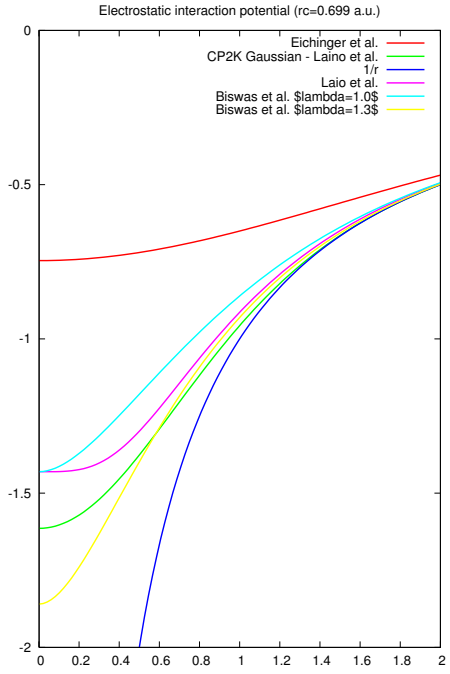


Figure 1. Electrostatic interaction potential between an electron (a.u.) and a unite positive charge. The r_c value is equal to 0.699 a.u..

point charge. Laio *et al.*[2] also remarked that they were not successful in finding a functional form that provides repulsion at short distances and could mimic the Pauli exclusion between electronic charge density and the MM point charge.

The problem has been finally solved by Biswas *et al.*[22] who obtained a regularized and renormalized description for the QM/MM electrostatic interaction by arguing that the point like description of the charge must be valid at interatomic separation but at short distance the Coulomb potential must be given by a localized charge distribution. The derivation below is a summary of the work of Biswas *et al.*[22], to account for the short-range effect.

Let us consider a localized wave function $\phi(\mathbf{r}-\mathbf{r}_a)$ for the charge present at $\mathbf{r}_a$ so that the normalization of the wave function provides the charge q_a,

$$\int |\phi(\mathbf{r}-\mathbf{r}_a)|^2 d^3r = q_a \tag{4}$$

where $\mathbf{r}$ is an arbitrary point in space. For ϕ, a good representation is a partial-wave expansion in terms of an orthonormal basis set $\phi_{lm} = \mathcal{R}_l(u)Y_{lm}(\hat{u})$ of a hydrogen-like wave function and take

$$\phi(\mathbf{u}) = \left(\frac{q_a}{\sum_l |C_l|^2}\right)^{1/2} \cdot \sum_{lm} C_l \mathcal{R}_l(u) Y_{lm}(\hat{u}) \tag{5}$$

where $\mathcal{R}_l(u)$ is similar to the radial part of the hydrogen-like wave function and $Y_{lm}(\hat{u})$,

the spherical harmonics, represent the angular part; C_l is the expansion coefficient. Similar partial-wave expansion to construct a wave function is frequently used in atomic molecular physics[28]. The next approximation is to adopt a first-order approximation (*l*=0) to the expansion scheme which would allow to account for the delocalized effect of the charge in the s-wave approximation

$$\phi(\mathbf{r}-\mathbf{r}_a) = \left(\frac{q_a\xi^3}{\pi}\right)^{1/2} \exp^{-\xi|\mathbf{r}-\mathbf{r}_a|} \tag{6}$$

The Slater function (see Eq. 6) provides a consistent picture with the localized description of a charge and also enables us to arrive at analytical forms for the potential and force as shown below for the s wave. A similar expansion scheme, but using Gaussian orbitals, has been employed earlier by Das *et al.*[21] to study QM/MM systems. Although both the Gaussian and Slater orbitals are known to provide competitive results, the Slater orbitals have the proper behavior (cusp) at the origin while the Gaussian orbitals are generally easier to deal with computationally. However, the analytical form for the Coulomb potential using Slater orbitals provides the same computational advantage as Gaussian orbitals. The parameter ξ of the Slater orbital has the dimension of an inverse length and it is *natural* associate it to the reciprocal of the covalent radius $r_{c,a}$: $\xi \approx \frac{1}{r_{c,a}}$. More generally the parameter ξ can be represented as $\xi = \lambda/r_{c,a}$, where the λ parameter will be used to renormalize the Coulomb energy at $2r_{c,a}$ (the interatomic separation). λ controls the spread of the wave function, and for $\lambda >> 1$ the charge distribution collapses to a point like charge.

With the above wave-function description for the charge, the Coulomb interaction potential (static potential) between the a^{th} MM atom and the QM system can be written as

$$\mathcal{H}_{\rho_e,q_a} = \int d^3r \int d^3r' \frac{\rho_e(\mathbf{r},\mathbf{r}_\alpha)|\phi(\mathbf{r}'-\mathbf{r}_a)|^2}{|\mathbf{r}'-\mathbf{r}|} \tag{7}$$

$$\mathcal{H}_{\rho_N,q_a} = \int d^3r \int d^3r' \frac{\rho_N(\mathbf{r}_\alpha)|\phi(\mathbf{r}'-\mathbf{r}_a)|^2}{|\mathbf{r}'-\mathbf{r}|} \tag{8}$$

where $\mathcal{H}^{Coul}_{QM/MM} = \mathcal{H}_{\rho_e,q_a} + \mathcal{H}_{\rho_N,q_a}$; ρ_N is the charge distribution of the ionic core of the α^{th} QM atom (i.e. sum of the nuclear and inner electron charges). In CP2K the ionic cores are distributed over the grid used also for the electronic charge density. This manner of distributing ionic core charges would not lead to any appreciable modification to the Coulomb energy as the separation between the QM nuclei and the MM atoms becomes of the order of interatomic separation in a molecule and thus would be quite compatible with the point-charge description. Thus, focusing on the effect of the spatial distribution of the MM charges over the QM electron density will not lead to any loss of accuracy. After performing the integrals (as shown in App. A) the following analytical expressions are obtained:

$$\mathcal{H}_{\rho_e,q_a} = \int d^3r q_a \rho_e(\mathbf{r},\mathbf{r}_\alpha) \cdot \left[\frac{1}{|\mathbf{r}-\mathbf{r}_a|} - \frac{\exp^{-2\xi|\mathbf{r}-\mathbf{r}_a|}}{|\mathbf{r}-\mathbf{r}_a|} - \xi * \exp^{-2\xi|\mathbf{r}-\mathbf{r}_a|}\right] \tag{9}$$

$$\mathcal{H}_{\rho_N,q_a} = \int d^3r q_a \rho_N(\mathbf{r}_\alpha) \cdot \left[\frac{1}{|\mathbf{r}-\mathbf{r}_a|} - \frac{\exp^{-2\xi|\mathbf{r}-\mathbf{r}_a|}}{|\mathbf{r}-\mathbf{r}_a|} - \xi * \exp^{-2\xi|\mathbf{r}-\mathbf{r}_a|}\right] \tag{10}$$

From the above, we see that asymptotically (i.e. for $|\mathbf{r}-\mathbf{r}_a| \to \infty$), $\mathcal{H}_{\rho_e,q_a}$ converges to the coulomb potential $\frac{1}{|\mathbf{r}-\mathbf{r}_a|}$. Also for $\xi \to \infty$ (which recovers the point-charge description

of the MM charge), the expression reduces to the usual Coulomb potential, as expected. At short distance the effect of the localized distribution of the MM charge introduces large cancellation to the Coulomb interaction and leads to a finite potential given by ξ (ξ has the dimension of 1/r). Thus this potential leads to zero forces as the distance approaches zero.

It is interesting and worthwhile to mention that the empirical form of the Coulomb potential proposed by Laio *et al.*[2] provides a very similar behavior; the two expressions differ marginally only at low and intermediate ranges, since both potentials converge to the value of $\frac{1}{r_{c,a}}$ at zero distance. As the value of the parameter λ is increased, one gradually approaches towards a point-charge description for the MM atom. At 0.97 Å (typical H-O

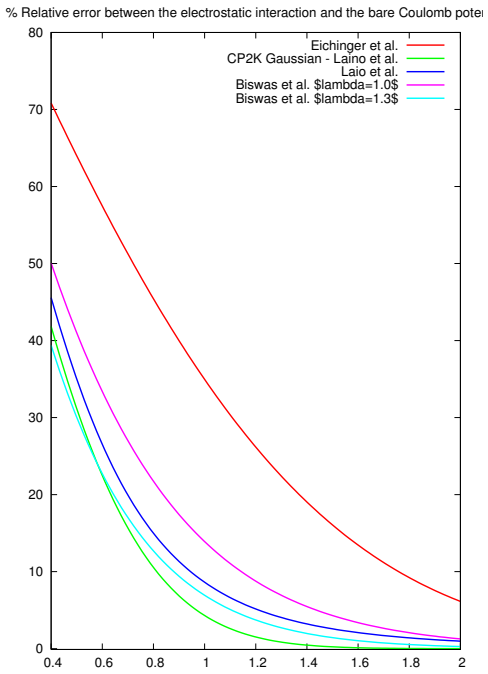


Figure 2. Relative % difference between different electrostatic interaction potential and the bare Coulomb interaction.

separation in water) the value of the electrostatic potential of Eichinger *et al.*[27], Laio *et al.*[2] and the one derived by Biswas *et al.*[22] differ from the Coulomb potential arising from the point-charge description by about 9.5%, 1.3% and 1.9%, respectively (see Fig. 2); all are smaller than the Coulomb potential of a point charge. It is worthwhile to emphasize that for $\lambda = 1.3$, the Coulomb potential approaches the point-charge potential faster and normalizes near 0.74 Å (interatomic separation in hydrogen molecule). The results of[22] show that a value of $\lambda = 1.3$ reduces the above difference of 1.9% (obtained with $\lambda = 1.0$) to 0.5% and the corresponding expansion of the point charge provides the best results[22]. Surprisingly, the empirical form of Laio *et al.*[2] describes the localized distribution of the MM point charge quite effectively and provides an understanding of the importance

of accounting the smearing effect of the MM charge. As compared with the functional form of Laio *et al.*[2] and with the Biswas *et al.* potential[22] obtained with Slater orbital, the potential arising from a Gaussian distribution of the point charge[27] overestimates this effect by about 46%.

Since the modification in the Coulomb potential reflects a delocalization effect of the MM charge, both the QM electron density and the QM ionic cores should experience the same modified Coulomb potential. However, this is not strictly necessary as far as the full Hamiltonian treatment stays consistent in the definition of energy and derivatives. In fact in Laio *et al.*[2] they do not consider the smearing effect of the MM charge when computing the interaction with the ionic cores (see Sec. IV of Ref. 2), thus replacing the modified Coulomb potential with a pure Coulomb interaction $\frac{1}{\mathbf{r}_\alpha - \mathbf{r}_a}$. Though the formalism presented in the next sections (see Sec. 4.1 and Sec. 5) relies on a Gaussian distribution for the point charge, both charge distribution densities (Gaussian and s-wave) are available in the QM/MM driver. Moreover, in the present implementation in CP2K I ensure that both the QM electrons and the ionic cores experience the same external potential.

3 Wave-Function Optimization

The multi-grid implementation is based on the use of an additive[29,24,23] QM/MM scheme. The total energy of the molecular system can be partitioned into three disjointed terms:

$$E_{TOT}(\mathbf{r}_\alpha, \mathbf{r}_a) = E^{QM}(\mathbf{r}_\alpha) + E^{MM}(\mathbf{r}_a) + E^{QM/MM}(\mathbf{r}_\alpha, \mathbf{r}_a) \tag{11}$$

where E^{QM} is the pure quantum energy, E^{MM} is the classical energy and $E^{QM/MM}$ represents the mutual interaction energy of the two subsystems. These energy terms depend parametrically on the coordinates of the quantum nuclei ($\mathbf{r}_\alpha$) and classical atoms ($\mathbf{r}_a$).

The quantum subsystem is described at the density functional theory (DFT) level, exploiting the QUICKSTEP[30] algorithm.

The classical subsystem is described through the use of the MM driver called FIST, also included in the CP2K package. This driver allows the use of the most common force fields employed in molecular mechanics simulations[31,32].

The interaction energy term $E^{QM/MM}$ contains all non-bonded contributions between the QM and the MM subsystem, and in a DFT framework we express it as:

$$E^{QM/MM}(\mathbf{r}_\alpha, \mathbf{r}_a) = \sum_{a \in MM} q_a \int \rho(\mathbf{r}, \mathbf{r}_\alpha) v_a(|\mathbf{r} - \mathbf{r}_a|) d\mathbf{r} + \sum_{a \in MM, \alpha \in QM} v_{\text{NB}}(\mathbf{r}_\alpha, \mathbf{r}_a) \tag{12}$$

where $\mathbf{r}_a$ is the position of the MM atom a with charge q_a, $\rho(\mathbf{r}, \mathbf{r}_\alpha)$ is the total (electronic plus nuclear) charge density of the quantum system, and $v_{\text{NB}}(\mathbf{r}_\alpha, \mathbf{r}_a)$ is the non–bonded interaction between classical atom a and quantum atom α, and finally:

$$v_a(|\mathbf{r} - \mathbf{r}_a|) = \frac{\text{Erf}(|\mathbf{r} - \mathbf{r}_a|/r_{c,a})}{|\mathbf{r} - \mathbf{r}_a|} \tag{13}$$

where $r_{c,a}$ is an atomic parameter, generally close to the covalent radius of the atom a. This function is the exact potential energy function originated by a Gaussian charge distribution $\rho(|\mathbf{r} - \mathbf{r}_a|) = \left(\frac{1}{\sqrt{\pi} * r_{c,a}}\right)^3 \exp(-(|\mathbf{r} - \mathbf{r}_a|/r_{c,a})^2)$. Moreover, the expression in Eq. 13 has

the desired property of tending to 1/r at large distances and going smoothly to a constant for small r (see Sec. 2).

Due to the Coulomb long-range behavior, the computational cost of the integral in Eq. 12 can be very large. When using a localized basis set like GTOs, the most natural way to handle this term is to modify the one-electron Hamiltonian by adding to it the contribution of the MM classical field:

$$H^{\mu\nu}_{QM/MM} = -\int \phi^*_\mu(\mathbf{r},\mathbf{r}_\alpha)\left[\sum_{a\in MM}\frac{q_a}{|\mathbf{r}_a-\mathbf{r}|}\right]\phi_\nu(\mathbf{r},\mathbf{r}_\alpha)d\mathbf{r} \tag{14}$$

ϕ_μ and ϕ_ν being Gaussian basis functions, depending parametrically on the QM nuclei positions $\mathbf{r}_\alpha$, and q_a the atomic charge of classical atom a with coordinates $\mathbf{r}_a$. In this case a suitable pre-screening procedure can be applied for the integral evaluation, in order to effectively compute only the non-zero terms and thus avoiding the quadratically scaling construction of the core Hamiltonian with respect to the number of elements of the basis set. When using a fully delocalized basis set like PWs, on the other hand, the QM/MM interaction term is evaluated by modifying the external potential and collocating on the grid nodes the contribution coming from the MM potential. Unfortunately the number of operations that a direct evaluation of Eq. 12 requires is of the order of $N_u N_{MM}$, where N_u is the number of grid points, usually of the order of 10^6 points, and N_{MM} is the number of classical atoms, usually of the order of 10^4 or more in systems of biochemical interest. It is evident that in a real system a brute force computation of the integral in Eq. 12 is impractical[a].

3.1 GEEP: Gaussian Expansion of the Electrostatic Potential

The key to solve this issue is a decomposition of the electrostatic potential in terms of Gaussian functions with different cutoffs:

$$v_a(\mathbf{r},\mathbf{r}_a) = \frac{\mathrm{Erf}(|\mathbf{r}-\mathbf{r}_a|/r_{c,a})}{|\mathbf{r}-\mathbf{r}_a|} = \sum_{N_g} A_g \exp^{-(|\mathbf{r}-\mathbf{r}_a|/G_g)^2} + R_{low}(|\mathbf{r}-\mathbf{r}_a|) \tag{15}$$

The smoothed Coulomb potential is expressed as a sum of N_g Gaussian functions and of a residual function R_{low}. The A_g are the amplitudes of the Gaussian functions, G_g are their width. If the parameters A_g and G_g are properly chosen, the residual function R_{low} will be smooth, i.e. its Fourier transform will have a compact domain for very small g vectors, and will be approximately zero for $g >> G_{cut}$. The G_{cut} parameter is related to the spacing of the grid on which the R_{low} function will be mapped. We performed the fit of Eq. 15 by a least square approach in Fourier space, using the analytical expression of the g-representation of the modified electrostatic potential[33]:

$$\tilde{v}_a(\mathbf{g}) = \left[\frac{4\pi}{g^2}\right]\exp\left(-\frac{g^2 r^2_{c,a}}{4}\right) \tag{16}$$

In Fig. 3 we show the result of the fitting procedure in g–space with $r_{c,a} = 1.1$ Å, comparing the Fourier components of the modified Coulomb potential with the Fourier components of the residual function R_{low}. In this case the compact support of R_{low} is truncated

[a] In order to overcome this computational bottleneck, most of these methods employ a multipolar expansion for reducing the computational complexity.

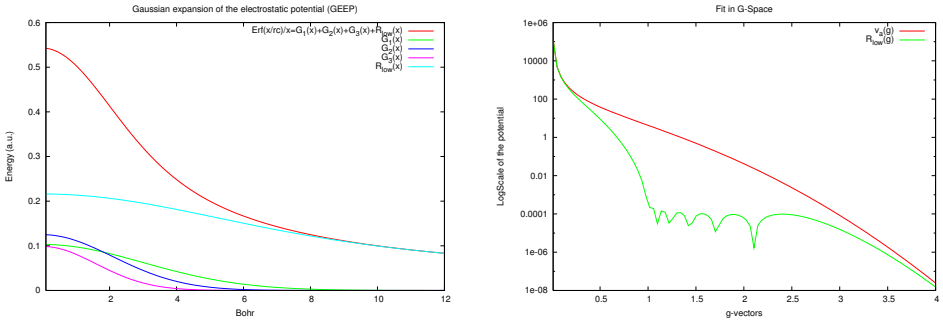


Figure 3. On the left: Gaussian expansion of the electrostatic potential (GEEP). The picture shows the components of the fit for the value $r_{c,a} = 1.1$ Å. On the right: Fourier transform of the potential in Eq. 13 (in red) and Fourier transform of the residual function R_{low} in Eq. 15 (in green). For this particular case ($r_{c,a} = 1.1$) we can define for the residual function a $G_{cut} \approx 1.0$.

at $G_{cut} \approx 1.0$ which should be compared with the value of $G_{cut} \approx 3.0$ needed to achieve the same accuracy when using $v_a(\mathbf{r}, \mathbf{r}_a)$. This implies that the residual function can be mapped on a grid with a spacing one order of magnitude bigger than the one required to map the v_a function.

In Fig. 3 we show the same result of the fit in real space and in Tab. 1 we provide coefficients for selected values of $r_{c,a}$.

	Radius $r_{c,a} = 1.1$ Å		Radius $r_{c,a} = 0.44$ Å	
Number of Gaussians	A_g (a.u.)	G_g (bohr)	A_g (a.u.)	G_g (bohr)
1	0.103103	4.243060	0.230734	1.454390
2	0.125023	2.966300	0.270339	1.094850
3	0.098613	2.254250	0.075855	4.906710
4	-	-	0.190667	0.883485
5	-	-	0.173730	1.965640
6	-	-	0.127689	2.658160
7	-	-	0.095104	3.591640

Table 1. Amplitudes and coefficients of the optimal Gaussians as derived by the fit.

The advantage of this decomposition scheme is that grids of different spacing can be used to represent the different contributions to $v_a(\mathbf{r}, \mathbf{r}_a)$. In fact, the evaluation of a function on a grid relies on the assumption that the grid spacing is optimally chosen on the basis of its Fourier transform spectrum. Writing a function as a sum of terms with compact support and with different cutoffs, the mapping of the function is achieved using different grid levels, in principle as many levels as contribution terms, each optimal to describe the corresponding function. In our case, sharp Gaussians require fine grids while coarser grids are necessary for the smoothest components. In addition the Gaussians can be truncated

beyond a certain threshold value, which makes the collocation of the Gaussians on the grid a very efficient process (see App. B).

The problem of mapping a non-compact function on a fine grid is then converted into the mapping of N_g compact functions on grids with cutoffs lower or at least equal to the fine grid, plus a non-compact very smooth function R_{low} mapped on the coarsest available grid. The sum of the contributions of all the grids, suitably interpolated, will be approximately equal to the function mapped analytically on the fine grid within errors due to the interpolation procedure.

3.2 GEEP library

A library with optimized parameters for the GEEP expansion is available into the CP2K code both for Gaussian and for s-wave (see App. A) charge distribution densities. In particular exploiting the scaling properties of both functional form it is possible to have a proper expansion for whatever value of the $r_{c,a}$ parameter.

3.3 Multi-Grid Framework

Multi-grid methods are well established tools in the solution of partial differential equations[34,35]. In the present implementation multi-grid techniques are employed to combine functions with different cutoffs, i.e. represented on different grid levels.

Let us start by considering two grids, a coarse grid $\mathcal{C}$ with N_c points and a fine grid $\mathcal{F}$ with N_f points, respectively at grid-level k-1 and k. The *interpolation* operator

$$I_{k-1}^{k} : \mathcal{C} \rightarrow \mathcal{F} \tag{17}$$

is by definition a transfer operator of a low cutoff function to a grid with an higher cutoff. The extension of the function to more points requires some regularity assumptions on its behavior. Two limiting cases can be identified: C^1 and C^∞, which can be handled with a simple linear interpolation scheme and with a G–space interpolation, respectively. If the function is C^∞, as in the case of a Gaussian, it is normally better to use an interpolator that assumes a high regularity. This ceases to be true once a collocation threshold is defined for the mapping of the Gaussians. In fact, the function on the grid becomes less regular, and an interpolation of a lower order might perform better. Another reason to avoid G–space interpolation comes from the fact that periodic boundary conditions with respect to the QM grid cannot be applied to the QM/MM potential. This makes the normal G–space interpolation unsuitable for our purpose. Thus we preferred to use an interpolation that works entirely in real space. For simplicity we use a set of commensurate grids, in which all the points of the coarse grid are points of the fine grid. Moreover, the number of points in each direction doubles going from the coarse to the fine grid level immediately above ($N_f = 8N_c$ in 3D). In the case of 1D space, the interpolation operator can be defined as:

$$I_{k-1}^{k}(i,j) = \sum_{n} T(i,n) S^{-1}(n,j) \tag{18}$$

where for the points away from the border $T(i,n) = N_3(n - i/2)$ and $S(i,j) = N_3(j-i)$; N_3 being the characteristic B-spline function of order 3[36] (see App. C). The border was treated as a non-uniform B-Spline. Higher dimensional spaces can be treated using the

direct product of the transformation along the single dimensions. The opposite operation, the *restriction* J_k^{k-1} is defined through the condition that the integral of the product of a function defined on the coarse grid with a function defined on a fine grid should give the same result both on the fine and on the coarse grid. Thus the restriction is simply the transpose of the interpolation

$$J_k^{k-1}(i,j) = \left[I_{k-1}^k(i,j)\right]^T = \sum_n S^{-1}(i,n)T(n,j) \tag{19}$$

Using N_{grid} grid levels and choosing the finer and coarser grid levels in order to treat correctly the sharpest and smoothest Gaussian components respectively, we can achieve good accuracy and performance.

4 QM/MM Coupling for Isolated Systems

4.1 QM/MM Energy

The QM/MM electrostatic energy within DFT can be expressed with the following equation:

$$E^{QM/MM}(\mathbf{r}_\alpha, \mathbf{r}_a) = \int d\mathbf{r}\rho(\mathbf{r}, \mathbf{r}_\alpha)V^{QM/MM}(\mathbf{r}, \mathbf{r}_a) \tag{20}$$

where $V^{QM/MM}$ is the electrostatic QM/MM potential evaluated on the finest grid, the same on which the final QM total density is evaluated. The overall description of the algorithm used to evaluate the QM/MM electrostatic potential on the finest grid can be outlined as follows:

- Each MM atom is represented as a continuous Gaussian charge distribution. The electrostatic potential generating from this charge is fitted through a Gaussian expansion using functions with different cutoffs, Sec. 3.1.

- Every Gaussian function is mapped on one of the available grid levels, chosen to be the first grid whose cutoff is equal to or bigger than the cutoff of that particular Gaussian function. Using this collocation criterion, every Gaussian will be represented on the same number of grid points irrespective of its width. In practice a submesh of size $\approx$ 25x25x25 is sufficient for an optimal Gaussian representation. Moreover, once a collocation threshold is defined, the Gaussian can be considered a compact domain function, i.e. it is zero beyond a certain distance, usually called Gaussian radius. Thus only MM atoms embedded into the QM box, or close to it, will contribute to the finest grid levels, as shown in Fig. 4.

  The result of this collocation procedure is a multi-grid representation of the QM/MM electrostatic potential $V_i^{QM/MM}(\mathbf{r}, \mathbf{r}_a)$, where i labels the grid level, represented by a sum of single atomic contributions $V_i^{QM/MM}(\mathbf{r}, \mathbf{r}_a) = \sum_{a \in MM} v_a^i(\mathbf{r}, \mathbf{r}_a)$, on that particular grid level. In a realistic system the collocation represents most of the computational time spent in the evaluation of the QM/MM electrostatic potential, that is around $60 - 80\%$.

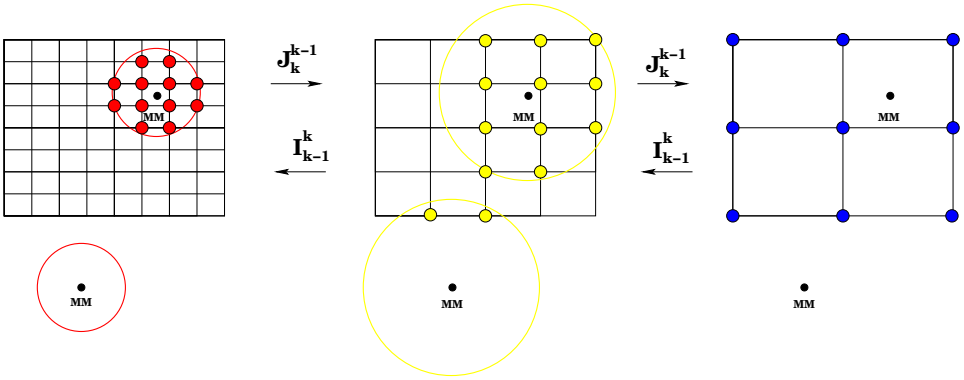


Figure 4. Schematic representation of the collocation procedure. Two MM atoms and three grid levels have been depicted. The circles (in the first and second grid levels) are the collocation region of the Gaussian centered on the two MM atoms. Atoms whose distance from the QM box is greater than the Gaussian collocation radius do not contribute to the potential on that grid level. However, all atoms contribute to the coarsest grid level through the long-range R_{low} part.

- Afterwards, the multi-grid expansion $V_i^{QM/MM}(\mathbf{r}, \mathbf{r}_a)$ is sequentially interpolated starting from the coarsest grid level up to the finest level. The QM/MM electrostatic potential on the finest grid level can then be expressed as:

$$V^{QM/MM}(\mathbf{r}, \mathbf{r}_a) = \sum_{i=coarse}^{fine} \prod_{k=i}^{fine-1} I_{k-1}^k V_i^{QM/MM}(\mathbf{r}, \mathbf{r}_a) \qquad (21)$$

  where $V_i^{QM/MM}(\mathbf{r}, \mathbf{r}_a)$ is the electrostatic potential mapped on grid level i and I_{k-1}^k is the interpolation operator in real space. This operation does not depend on the number of MM atoms but only on the number of grid points, i.e. on the cutoff used in the calculation and on the dimensions of the QM box. For realistic systems the computational cost is around $20 - 40\%$ of the overall cost of the evaluation of the QM/MM electrostatic potential.

Using the real space multi-grid technique together with the GEEP expansion, the prefactor in the evaluation of the QM/MM electrostatic potential has been lowered from $N_f * N_f * N_f$ to $N_c * N_c * N_c$, where N_f is the number of grid points on the finest grid and N_c is the number of grid points on the coarsest grid. The computational cost of the other operations for evaluating the electrostatic potential, such as the mapping of the Gaussians and the interpolations, becomes negligible in the limit of a large MM system, usually more than 600-800 MM atoms.

Using the fact that grids are commensurate ($N_f/N_c = 2^{3(N_{grid}-1)}$), and employing for every calculation 4 grid levels, the speed-up factor is around 512 (2^9); this means that the present implementation is 2 orders of magnitude faster than the direct analytical evaluation of the potential on the grid. The number of grid levels that can be used is limited by two technical factors. The first is that the coarsest grid needs to have at least as

many points per dimension as the ones corresponding to the cutoff of the residual function R_{low} in order to perform the interpolation/restriction in an efficient manner. The second limitation is due to the constraint of using commensurate grid levels. The more grid levels are required in the calculation, the more the finest grid level cutoff will increase. This leads to an increment in memory requirements and to an unnecessary precision when handling the higher cutoff grids. Usually it is a combination of cutoff and grid levels that provides maximum efficiency. The two parameters can be chosen by checking that the coarsest grid level has no more than 5-10 grid points per dimension within the specified cutoff for the finest grid. Following the previous rule, the number of operations required for the direct evaluation of Eq. 12 is of the order of N*100*N_{MM}, where N is an integer between 1 and 10 and N_{MM} is the number of classical atoms.

4.2 QM/MM Forces

The forces on classical atoms due to the interaction Hamiltonian Eq. 20 are obtained by taking the derivative of Eq. 20 with respect to the classical atomic positions $\mathbf{r}_a$:

$$-\frac{\partial E^{QM/MM}}{\partial \mathbf{r}_a} = -\int \rho(\mathbf{r},\mathbf{r}_\alpha)\frac{\partial V^{QM/MM}(\mathbf{r},\mathbf{r}_a)}{\partial \mathbf{r}_a}d\mathbf{r} \tag{22}$$

The integral evaluation can be divided into terms deriving from the different grid levels:

$$-\frac{\partial E^{QM/MM}}{\partial \mathbf{r}_a} = -\sum_{i=coarse}^{fine}\int \rho(\mathbf{r},\mathbf{r}_\alpha)\frac{\partial V_{fine}^{i,QM/MM}(\mathbf{r},\mathbf{r}_a)}{\partial \mathbf{r}_a}d\mathbf{r} \tag{23}$$

where the $V_{fine}^{i,QM/MM}$ labels the potential term on the finest grid level coming from the corresponding grid level i. Using the multi-grid expression for terms $V_{fine}^{i,QM/MM} = \prod_{k=i}^{fine-1} I_{k-1}^{k} V_i^{QM/MM}$, the derivatives can be written as:

$$-\frac{\partial E^{QM/MM}}{\partial \mathbf{r}_a} = -\sum_{i=coarse}^{fine}\int \rho(\mathbf{r},\mathbf{r}_\alpha)\prod_{k=i}^{fine-1} I_{k-1}^{k}\frac{\partial V_i^{QM/MM}(\mathbf{r},\mathbf{r}_a)}{\partial \mathbf{r}_a}d\mathbf{r} \tag{24}$$

$$= -\sum_{i=coarse}^{fine}\int \left[\prod_{k=i+1}^{fine} J_k^{k-1}\right]\rho(\mathbf{r},\mathbf{r}_\alpha)\frac{\partial V_i^{QM/MM}(\mathbf{r},\mathbf{r}_a)}{\partial \mathbf{r}_a}d\mathbf{r} \tag{25}$$

In the previous equation the property that the interpolation operator is equal to the transpose of the restriction operator (and vice-versa) was used. The MM derivatives are then evaluated applying the restriction operator to the converged QM $\rho(\mathbf{r},\mathbf{r}_\alpha)$. This leads to a multi-grid expansion of the density and each integral is evaluated on the appropriate grid level. The overall derivative is the sum of the contributions of the different grid levels.

We now consider the forces on the QM atoms. If $n_c^\alpha(\mathbf{r})$ is the Gaussian density used to represent the core charge distribution of the α^{th} quantum ions and labeling with $P^{\mu\nu}$ the $\mu\nu$ element of the density matrix in the Gaussian basis set $\{\ \phi_\mu\ \}$, the forces on quantum

ions[30] due to the QM/MM interaction potential are

$$-\frac{\partial E^{QM/MM}}{\partial \mathbf{r}_\alpha} = -\sum_{\mu\nu}(\frac{\partial P^{\mu\nu}}{\partial \mathbf{r}_\alpha})V^{QM/MM}_{\mu\nu} - \tag{26}$$

$$2\sum_{\mu\nu} P^{\mu\nu}\int(\frac{\partial \phi_\mu(\mathbf{r},\mathbf{r}_\alpha)}{\partial \mathbf{r}_\alpha})V^{QM/MM}(\mathbf{r},\mathbf{r}_a)\phi_\nu(\mathbf{r},\mathbf{r}_\alpha)d\mathbf{r} - \tag{27}$$

$$\int(\frac{\partial n^\alpha_c(\mathbf{r},\mathbf{r}_\alpha)}{\partial \mathbf{r}_\alpha})V^{QM/MM}(\mathbf{r},\mathbf{r}_a)d\mathbf{r} \tag{28}$$

where $V^{QM/MM}_{\mu\nu} = \int \phi_\mu(\mathbf{r},\mathbf{r}_\alpha)V^{QM/MM}(\mathbf{r},\mathbf{r}_\alpha)\phi_\nu(\mathbf{r},\mathbf{r}_\alpha)d\mathbf{r}$ is the QM/MM Hamiltonian interaction term in the Gaussian basis set $\{\ \phi_\mu\ \}$. The first term is the so-called Pulay term[37] and is present because the basis set depends explicitly on the atomic position[30]. It vanishes if Gaussians form a complete basis set. The evaluation of the gradients on QM atoms is relatively inexpensive compared to a full quantum calculation. All considerations raised in Sec. 4.1, concerning the scaling of the present scheme in the evaluation of the QM/MM potential, remain valid in the evaluation of the forces on classical atoms.

The calculation of the forces within the present implementation has been compared with the calculation of the forces using the method described in Ref. 2, which is an implementation of QMMM in the CPMD code[5]. Comparison with the CPMD-QMMM code is complicated by the fact that in Ref. 2 a multipolar expansion is used for the long-range part of the QM/MM electrostatic coupling. For this reason we compare only forces on atoms of the first MM solvation shell, which are treated exactly also in CPMD-QMMM code. We consider a system of 215 classical SPC[38] water molecules and 1 QM water molecule. Although the system size is relatively small, the number of molecules present is comparable to the number of molecules normally treated exactly in CPMD-QMMM. In Fig. 6 we show the relative error between the previous and the present implementations. The highest relative errors (less than 1.0 %) correspond to forces which have small modules ($\leq 10^{-3} a.u.$). The average relative error is $\approx 0.01\%$ with a speed-up in the energy and derivative evaluation of a factor of 40 with respect to CPMD.

An important figure of merit for QM/MM codes that are aimed at molecular dynamics (MD) simulation is their ability to conserve the energy. In order to address this issue we studied a system composed of 3 water molecules, 2 MM and 1 QM, equilibrated at 400K. We simulate this system for 1 picosecond. The results are shown in Fig. 5. For comparison the energy of the pure classical and the quantum run are shown in the same picture. No drift is observed during 1 picosecond of simulation. We also show the potential energy during the simulation, whose oscillation is ≈ 3 orders of magnitude bigger than the total energy oscillation.

5 Extension to Periodic Boundary Conditions

Assuming the overall charge neutrality condition, the electrostatic interaction energy of a QM/MM simulation within PBC can be easily evaluated:

$$E^{TOT} = \frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} dr dr' \frac{\rho(r)\rho(r')}{|r-r'|} \tag{29}$$

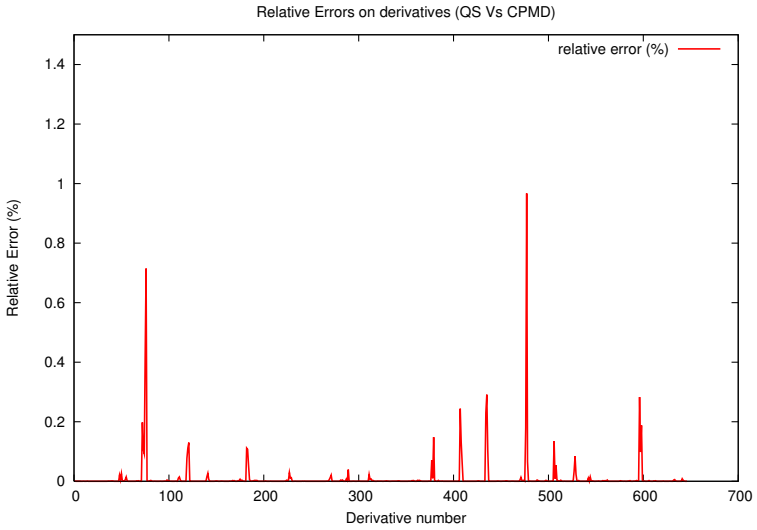


Figure 5. Relative errors on derivatives evaluated with the different functional form of Eq. 13 implemented in CPMD code and the new scheme implemented in CP2K. The average relative error is 0.01 %.

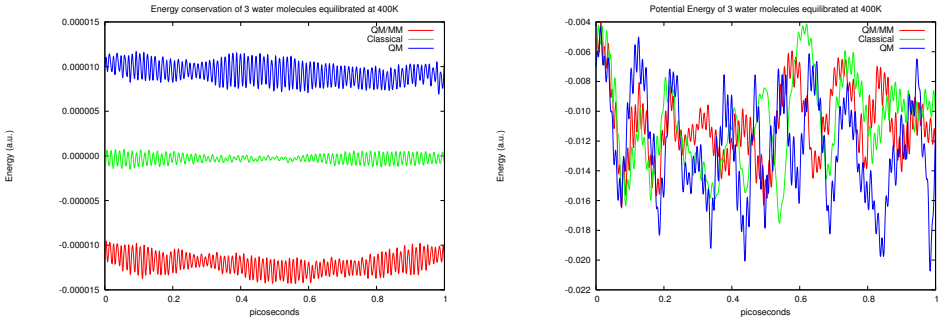


Figure 6. On the left: energy conservation of a system composed of 3 water molecules equilibrated at 400K during 1 ps of simulation. The red line shows the total energy for the QM/MM run, the green line represents a pure classical run and the blue line shows a pure quantum run. The total energies have been shifted for better visualization. No drift is observed and all energy conservation is consistent. On the right: potential energy during the same run. Its variation is 3 orders of magnitude larger than the total energy variation.

$\rho = \rho^{QM} + \rho^{MM}$ being the total charge density of the system (see Fig. 7-a). Volumes integrals cover the full space $\mathbb{R}^3$ and will be omitted in the following to make notation lighter.

Once the total density is split into a QM and a MM part both sub-systems could in principle possess an overall net charge different from zero. Therefore the use of a neutralizing

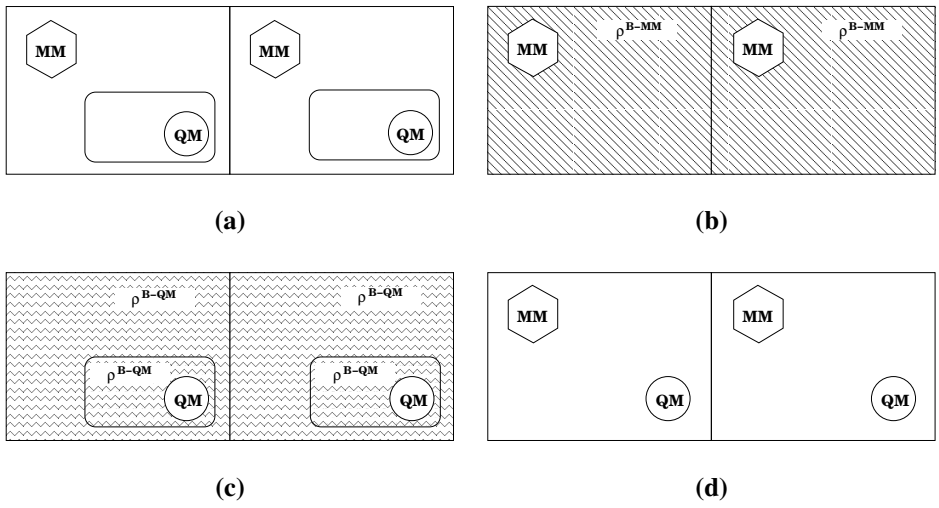


Figure 7. These frames show the decomposition of the total QM/MM energy. In each frame two of the many periodic replica have been shown. Frame Fig. 7-a shows the total system. Frame Fig. 7-b shows the energy of the MM sub-system embedded in the neutralizing background charge (deriving from the division of the QM and MM sub-systems). Frame Fig. 7-c shows the energy of the QM sub-system with the neutralizing background charge of the QM cell and that relating to the MM cell. The last frame Fig. 7-d depicts the QM/MM pure electrostatic mutual interaction term.

background charge (ρ^B) is necessary to avoid divergence in treating electrostatic within PBC. The total energy (see Fig. 7-a) term can be split into three separate terms:

$$E^{MM} = \frac{1}{2}\int\int drdr' \frac{(\rho^{MM}(r)+\rho^{B,MM})(\rho^{MM}(r')+\rho^{B,MM})}{|r-r'|} \tag{30}$$

$$E^{QM} = \frac{1}{2}\int\int drdr' \frac{(\rho^{QM}(r)+\rho^{B,QM})(\rho^{QM}(r')+\rho^{B,QM})}{|r-r'|} \tag{31}$$

$$E^{QM/MM} = \int\int drdr' \frac{(\rho^{QM}(r)+\rho^{B,QM})(\rho^{MM}(r')+\rho^{B,MM})}{|r-r'|} \tag{32}$$

The physical nature of these terms is illustrated pictorially in Fig. 7. Assuming the total charge of the system is zero (although this assumption can be relaxed with no modifications to the formalism) the mixed terms involving the neutralizing background charge of the $E^{QM/MM}$ cancel the interaction terms of the QM and MM density with their own

background charges. The expression for the three terms is:

$$E^{MM} = \frac{1}{2}\int\int dr dr' \frac{(\rho^{MM}(r))(\rho^{MM}(r'))}{|r-r'|} \tag{33}$$

$$E^{QM} = \frac{1}{2}\int\int dr dr' \frac{(\rho^{QM}(r))(\rho^{QM}(r'))}{|r-r'|} \tag{34}$$

$$E^{QM/MM} = \int\int dr dr' \frac{(\rho^{QM}(r))(\rho^{MM}(r'))}{|r-r'|} \tag{35}$$

The first term (Eq. 33 and Fig. 7-b) is evaluated using standard techniques such as particle-particle or particle-mesh schemes[39,40]. The second term (Eq. 34 and Fig. 7-c) is the Hartree energy of the QM sub-system. Since the total energy of the QM sub-system is usually evaluated exploiting a smaller cell, care needs to be taken to include the correct electrostatic interactions of the periodic QM replicas, i.e. restore the correct periodicity (MM cell). The last term (Eq. 35 and Fig. 7-d) is the evaluation of the periodic MM electrostatic potential, partitioned into a real space contribution and a periodic correction. The real space term contains the interaction due to the short-range part of the electrostatic potential of the MM charges with the total quantum charge distribution (electrons plus nuclei). Only MM atoms close to the QM region will contribute to this term. The periodic term contains instead the long-range effects of the MM sub-system.

In the next subsection, the standard Ewald method is briefly reviewed for a N-point charge particle system interacting in an orthorhombic box of edge L_x, L_y, L_z. We then introduce the Ewald lattice summation with the GEEP scheme[41]. Finally we discuss the algorithm to decouple/recouple multiple QM images.

5.1 Ewald Lattice Summation for Electrostatic Interactions

Given an N-point charge particle system, the electrostatic potential $\Phi_{tot}(\mathbf{r})$ at position $\mathbf{r}$ is evaluated using the Ewald lattice sum technique[43]. In this approach, $\Phi_{tot}(\mathbf{r})$ is split into the sum of two potentials, using a Gaussian screening charge of width κ:

$$\Phi_{tot}(\mathbf{r}) = \Phi_{rec}(\mathbf{r}) + \Phi_{real}(\mathbf{r}) \tag{36}$$

The reciprocal space potential term $\Phi_{rec}(\mathbf{r})$ is determined using the Fourier series:

$$\Phi_{rec}(\mathbf{r}) = \frac{4\pi}{V}\sum_{\mathbf{k}\neq 0}\frac{e^{-\frac{k^2}{4\kappa}}}{k^2}\sum_{a}^{MM} q_a e^{-\imath\mathbf{k}\cdot|\mathbf{r}-\mathbf{r}_a|} \tag{37}$$

where $\mathbf{k} = [2\pi n_x/L_x^2, 2\pi n_y/L_y^2, 2\pi n_z/L_z^2]$ and V is the volume ($\mathrm{V}=\mathrm{L}_x\cdot\mathrm{L}_y\cdot\mathrm{L}_z$) of the primary unit cell. The real space part of the Ewald potential is given by:

$$\Phi_{real}(\mathbf{r}) = \sum_{a}^{MM}\sum_{|\mathbf{L}|\leq L_{cut}} q_a \frac{\mathrm{Erfc}(\kappa|\mathbf{r}-\mathbf{r}_a+\mathbf{L}|)}{|\mathbf{r}-\mathbf{r}_a+\mathbf{L}|} \tag{38}$$

where $\mathbf{L} = [n_x L_x, n_y L_y, n_z L_z]$ counts the periodic images and n_x, n_y and n_z are integers. As the Erfc has a real space short-range property, only the $|\mathbf{L}| \leq L_{cut}$ periodic images will contribute to the real space term of the electrostatic potential.

5.2 QM/MM Periodic Potential

The QM/MM periodic potential (see Eq. 35 and Fig. 7-d) on a generic point i of the finest grid level can be computed using the real space lattice sum:

$$V^{fine}(\mathbf{r}_a)_i = \sum_{a}^{MM} \sum_{\mathbf{L}}^{\infty\ \prime} q_a v_a(\mathbf{r}_i, \mathbf{r}_a + \mathbf{L}) \tag{39}$$

where $\mathbf{r}_i$ is the coordinate of the point i of the finest grid level and $\mathbf{r}_a$ indexes the functional dependence from the set of MM atomic coordinates and v is given by Eq. 13. The summation over $\mathbf{L}$ involves all integer translations of the real space lattice vectors $\mathbf{L} = [n_x L_x, n_y L_y, n_z L_z]$ for integers n_k and the prime symbol indicates that when $\mathbf{L} = 0$ the term $|\mathbf{r}_i - \mathbf{r}_a| = 0$ is neglected. The summation in Eq. 39 has the same convergence properties as the standard Ewald summation schemes[43].

The total QM/MM electrostatic energy can be split into two rapidly convergent terms[44,43], one over real space and the other over reciprocal space lattice vectors:

$$E^{QM/MM}(\mathbf{r}_\alpha, \mathbf{r}_a) = E^{QM/MM}_{real}(\mathbf{r}_\alpha, \mathbf{r}_a) + E^{QM/MM}_{recip}(\mathbf{r}_\alpha, \mathbf{r}_a) \tag{40}$$

where:

$$E^{QM/MM}_{real}(\mathbf{r}_\alpha, \mathbf{r}_a) = \int d\mathbf{r} \rho(\mathbf{r}, \mathbf{r}_\alpha) V^{QM/MM}_{real}(\mathbf{r}, \mathbf{r}_a) \tag{41}$$

and

$$E^{QM/MM}_{recip}(\mathbf{r}_\alpha, \mathbf{r}_a) = \int d\mathbf{r} \rho(\mathbf{r}, \mathbf{r}_\alpha) V^{QM/MM}_{recip}(\mathbf{r}, \mathbf{r}_a) \tag{42}$$

The definition of the two terms is strictly connected to the type of functional form used to describe the Coulomb interactions. For Gaussian charge distributions (but similar expressions are available as well for the s–wave charge expansion, i.e. App. A), the electrostatic potential function has the analytical form:

$$v_a(\mathbf{r}, \mathbf{r}_a) = \frac{\mathrm{Erf}(|\mathbf{r} - \mathbf{r}_a|/r_{c,a})}{|\mathbf{r} - \mathbf{r}_a|} \tag{43}$$

easily represented as a sum of two terms[41]:

$$v_a(\mathbf{r}, \mathbf{r}_a) = \frac{\mathrm{Erf}(|\mathbf{r} - \mathbf{r}_a|/r_{c,a})}{|\mathbf{r} - \mathbf{r}_a|} = \sum_{N_g} A_g \exp(-\frac{|\mathbf{r} - \mathbf{r}_a|^2}{G_g^2}) + R_{low}(|\mathbf{r} - \mathbf{r}_a|) \tag{44}$$

The best choice is to use the mathematical properties of the two functional forms (short-range term and long-range term) to define the division into real and reciprocal space contributions:

$$\begin{aligned} v_a(\mathbf{r}, \mathbf{r}_a) = \frac{\mathrm{Erf}(|\mathbf{r} - \mathbf{r}_a|/r_{c,a})}{|\mathbf{r} - \mathbf{r}_a|} &= \sum_{N_g} A_g \exp(-\frac{|\mathbf{r} - \mathbf{r}_a|^2}{G_g^2}) + R_{low}(|\mathbf{r} - \mathbf{r}_a|) \\ &= v_a^{rs}(\mathbf{r}, \mathbf{r}_a) + v_a^{recip}(\mathbf{r}, \mathbf{r}_a) \end{aligned} \tag{45}$$

All short-range interactions will be evaluated in the real space while all long-range interactions will be taken into account in the reciprocal space formalism. The real space term $V_{real}^{QM/MM}(\mathbf{r}, \mathbf{r}_a)$ is defined as:

$$V_{real}^{QM/MM}(\mathbf{r}, \mathbf{r}_a) = \sum_{|\mathbf{L}| \leq L_{cut}} \sum_a q_a v_a^{rs}(\mathbf{r}, \mathbf{r}_a + \mathbf{L})$$
$$= \sum_{|\mathbf{L}| \leq L_{cut}} \sum_a q_a \left[\sum_{N_g} A_g \exp(-\frac{|\mathbf{r} - \mathbf{r}_a + \mathbf{L}|^2}{G_g^2}) \right] \quad (46)$$

where a labels the MM atoms. The radii of the Gaussians are such that only a few periodic images ($|\mathbf{L}| \leq L_{cut}$, ideally only one) are needed to achieve convergence of the real space term, while others give zero contribution. As in[41], each Gaussian of Eq. 46 is mapped on the appropriate grid level. The same approach outlined here for Gaussian charge distribution holds for the s-wave charge expansion.

The effect of the periodic replicas of the MM sub-system is only in the long-range term, and it comes entirely from the residual function $R_{low}(\mathbf{r}, \mathbf{r}_a)$ of Eq. 45:

$$V_{recip}^{QM/MM}(\mathbf{r}, \mathbf{r}_a) = \sum_{\mathbf{L}}^{\infty}{}' \sum_a q_a v_a^{recip} = \sum_{\mathbf{L}}^{\infty}{}' \sum_a q_a R_{low}(|\mathbf{r} - \mathbf{r}_a + \mathbf{L}|) \quad (47)$$

Performing the same manipulation used in Ewald summation[43] (see App. B) the previous equation can be computed more efficiently in the reciprocal space:

$$V_{recip}^{QM/MM}(\mathbf{r}_i, \mathbf{r}_a) = L^{-3} \sum_{\mathbf{k}}^{k_{cut}}{}' \sum_a \tilde{R}_{low}(\mathbf{k}) q_a \cos\left[2\pi \mathbf{k} \cdot (\mathbf{r}_i - \mathbf{r}_a)\right] \quad (48)$$

The term $\tilde{R}_{low}(\mathbf{k})$, representing the Fourier transform of the smooth electrostatic potential, can be evaluated analytically:

$$\tilde{R}_{low}(\mathbf{k}) = \left[\frac{4\pi}{|\mathbf{k}|^2}\right] \exp\left(-\frac{|\mathbf{k}|^2 r_{c,a}^2}{4}\right) - \sum_{N_g} A_g (\pi)^{\frac{3}{2}} G_g^3 \exp(-\frac{G_g^2 |\mathbf{k}|^2}{4}) \quad (49)$$

The potential in Eq. 48 can be mapped on the coarsest grid. In fact, the long-range contribution is physically very smooth and a good representation can be achieved with large grid spacings. Furthermore, since the R_{low} function is a low cutoff function, $\tilde{R}_{low}(\mathbf{k})$ is zero for all k-vectors larger than a well defined k_{cut}. The k_{cut} parameter depends strongly on the number of Gaussian functions used in the GEEP scheme (as described in Sec. 3.1).

Once the electrostatic potential of a single MM charge within periodic boundary conditions is derived, the evaluation of the electrostatic potential due to the MM sub-system is easily computed employing the same multi-grid operators (interpolation and restriction) described in Sec. 3.3 and in App. C.

5.3 Periodic Coupling with QM Images

In the present section we complete the description of the electrostatic coupling, discussing the interaction between the periodic images of the QM replicas (see Fig. 7-c). The Quickstep[30,45] algorithm uses a mixed plane wave / Gaussian basis set to solve the DFT equations

for the quantum sub-system. The plane waves are used to compute efficiently the Hartree potential. Therefore, unless the quantum box and the MM box have the same dimensions, the QM images, interacting by PBC implicit in the evaluation of the Hartree potential, have the wrong periodicity.

In order to avoid this error, the QM problem is usually solved using standard decoupling techniques[46,47]. This approximation is legitimate when the evaluation of the QM/MM potential is performed using spherical truncation schemes for Coulomb interactions.

Since we want to describe the long-range QM/MM interaction with periodic boundary conditions, we may not neglect the QM/QM periodic interactions, which play a significant role if the QM sub-system has a net charge different from zero or a significant dipole moment. Therefore we exploit a technique proposed few years ago by Blöchl[47], for decoupling the periodic images and restoring the correct periodicity also for the QM part. A full and comprehensive description of the methods to evaluate energy corrections and derivatives is given in Ref. 47. Here we summarize Blöchl's decoupling scheme. Given a QM total density charge $\rho(\mathbf{r}, \mathbf{r}_\alpha)$, the electrostatic energy of this isolated density is:

$$E = \frac{1}{2} \int_V d\mathbf{r} \int d\mathbf{r}' \frac{\rho(\mathbf{r}, \mathbf{r}_\alpha)\rho(\mathbf{r}', \mathbf{r}_\alpha)}{|\mathbf{r} - \mathbf{r}'|} \tag{50}$$

Let us introduce a new model charge density $\hat{\rho}(\mathbf{r}, \mathbf{r}_\alpha)$, which is localized within the same volume V as $\rho(\mathbf{r}, \mathbf{r}_\alpha)$ and which reproduces the multipole moments of the correct charge distribution. The representation adopted in Ref. 47 is given by the sum:

$$\hat{\rho}(\mathbf{r}, \mathbf{r}_\alpha) = \sum_\alpha q_\alpha g_\alpha(\mathbf{r}, \mathbf{r}_\alpha) \tag{51}$$

of atom-centered spherical Gaussians, which are normalized such that they posses a charge of one:

$$g_i(\mathbf{r}, \mathbf{r}_\alpha) = \frac{1}{(\sqrt{\pi} r_{c,\alpha})^3} \exp(-\frac{|\mathbf{r} - \mathbf{r}_\alpha|^2}{r_{c,\alpha}^2}) \tag{52}$$

where $\mathbf{r}_\alpha$ denotes a particular atomic site. Every atomic site may be the center of various Gaussians with different decay lengths $r_{c,\alpha}$. By construction, the multipole moments of the model charge density agree with those of the original charge distribution. Since the electrostatic interaction of separated charge distribution (the array of periodic QM charge densities) depends only on its multipole moments, the model charge density is used to modify the Hartree potential and to cancel the electrostatic interactions between the periodic images. In App. D, we briefly summarize with a matrix formalism the charge fit scheme as derived in Ref. 47. In the same way as the Blöchl scheme cancels the electrostatic interactions between periodic images, it is possible to use it to include the electrostatic interactions between periodic images with the periodicity of the MM box.

5.4 QM/MM Forces

The derivatives on MM atoms can be easily evaluated taking the derivative of both terms in real space and in reciprocal space, and summing the contribution of the different grid levels. The derivatives of the real space term are the same as the one presented in Sec. 4.2. The derivatives of the reciprocal space term need to be evaluated by deriving the MM

nuclei potential energy contribution and integrating this derivative with the quantum charge distribution:

$$\frac{\partial E_{recip}^{QM/MM}(\mathbf{r}_\alpha, \mathbf{r}_a)}{\partial \mathbf{r}_a} = \int d\mathbf{r}\rho(\mathbf{r}, \mathbf{r}_\alpha)\frac{\partial V_{recip}^{QM/MM}(\mathbf{r}, \mathbf{r}_a, \mathbf{r}_\alpha)}{\partial \mathbf{r}_a} = \tag{53}$$

$$\Delta\omega \sum_{\mathbf{r}_i} \rho(\mathbf{r}_i, \mathbf{r}_\alpha) L^{-3} {\sum_{\mathbf{k}}}^{k_{cut}}{}' {\sum_{a}}^{MM} \tilde{R}_{low}(\mathbf{k}) q_a \frac{\partial \cos\left[2\pi\mathbf{k}\cdot(\mathbf{r}_i - \mathbf{r}_a)\right]}{\partial \mathbf{r}_a} \tag{54}$$

where $\Delta\omega$ is the volume element of the coarsest grid level. This contribution is summed with the terms in real space to obtain the total derivatives on MM atoms. The derivatives on QM atoms are computed in the same way as we described in Sec. 4.2, the only difference being that the QM derivatives are modified by the coupling/decoupling terms. These corrections have been derived and extensively discussed in Ref. 47.

6 Tests and Applications

Three systems were selected to test the new method. The first one, an infinite array of Gaussian alternating opposite charges, can be solved analytically and therefore provides a clear and unambiguous test of the accuracy of our new approach.

The second system is a periodic model of α-quartz (α-SiO_2) where a bulk fragment, described at the DFT level, is embedded in the environment of classical atoms described with MM force fields. The third system analyzes a charged oxygen vacancy defect in α-quartz, in the same periodic model. These two systems do not possess an analytical solution but both have been extensively studied experimentally[48–55] and theoretically[56–64].

6.1 Analytical Test

In order to validate this new algorithm, we consider the electrostatic interaction of an array of Gaussian charge distributions:

$$\rho(\mathbf{r}_\alpha) = (\kappa/\pi)^{3/2} \exp(-\kappa^2 |\mathbf{r}_\alpha|^2) \tag{55}$$

κ being the width of the Gaussian charge density.

The charges (32 positively charged (+1) and 32 negatively charged (-1)) are arranged on a cubic array of points forming a NaCl lattice. Neighboring charges have opposite sign. The potential generated by such a set of charges can be calculated exactly by noting that the electrostatic potential of a single charge density (Eq. 55) at an arbitrary distance $\mathbf{r}$ can be determined analytically, $V_{ext}(\mathbf{r}) = \mathrm{Erf}(\kappa\mathbf{r})/\mathbf{r}$. We now construct a test QM/MM model, selecting two neighboring charges (see Fig. 8) and calculating the Hartree potential in a smaller orthorhombic cell centered around the two chosen charges. This calculation would have been a necessary step had we treated the two selected centers quantum mechanically instead of with a fixed nuclear charge distribution. The calculation was performed using a plane wave cutoff of 25 Ry and 3 Gaussians were used for each selected atom to build the model density used to decouple/recouple the periodic images.

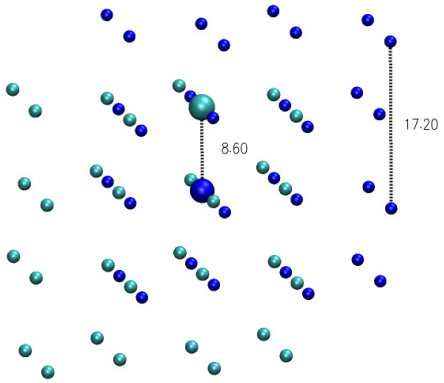


Figure 8. Orthorhombic cell of face centered cubic lattice of Gaussian charges. The two big spheres represent the QM atoms. Lattice parameter 17.2 Å. The Gaussian charges have a width of $0.5\sqrt{2}$ Å.

QM Cell (x,y,z) (Å)	Num. Gauss.	k_{cut} (bohr^{-1})	E_{tot} (Hartree)	ΔE (mHartree)
34.4 34.4 34.4	Analytical Calculation		3.441010	
34.4 34.4 34.4	6	0.5	3.440520	0.49
34.4 34.4 34.4	6	0.7	3.441176	-0.17
34.4 34.4 34.4	6	1.0	3.441119	-0.11
34.4 34.4 34.4	6	2.0	3.441070	-0.06
34.4 34.4 34.4	6	0.5	3.440520	0.49
34.4 34.4 34.4	9	0.5	3.440687	0.33
34.4 34.4 34.4	12	0.5	3.440885	0.12
34.4 34.4 34.4	15	0.5	3.440895	0.11
34.4 34.4 34.4	15	0.5	3.440895	0.11
27.0 27.0 27.0	15	0.5	3.440978	0.03
34.4 27.0 27.0	15	0.5	3.440951	0.06
22.0 22.0 12.0	15	0.5	3.440865	0.14
12.0 12.0 12.0	15	0.5	3.441356	-0.35
34.4 34.4 34.4	QM/MM non-periodic*		3.443106	2.10

Table 2. The interaction of a Gaussian charge distribution in a 3-dimensional lattice as shown in Fig. 8 as a function of the number of Gaussians used in GEEP and as a function of the QM cell. * The QM/MM non-periodic calculation was performed with 64000 MM atoms arranged in a cube cell of 344.0 Å.

In Tab. 2 we show how this pseudo QM/MM calculation depends on parameters like the QM cell dimension (affecting the coupling/decoupling between QM periodic images), the k_{cut} parameter of Eq. 48 and the number of Gaussians used in the GEEP scheme. In particular we note that the number of Gaussians is strictly correlated to the k_{cut} value. In fact, the more Gaussians that are used in the GEEP scheme, the more the R_{low} will be a

low cutoff function. This permits a smaller k_{cut} parameter to be used in order to reach the same accuracy (see Tab. 2).

The choice of the dimension of the QM box is almost irrelevant for the accuracy of the results (see Tab. 2). In fact even using a box of 12.0 Å, which is the smallest possible box size usable with this QM sub-system, we find accurate results. We remark that other decoupling techniques[46,18] require boxes twice the size of the minimum box, leading to a substantial computational overhead.

Moreover we computed the pseudo QM/MM interaction energy for the non-periodic pseudo QM/MM calculation, using an MM environment of 64000 atoms (MM cell side of 344.0 Å). The result shows that for ordered structures surface effects are very important and the only way to include correctly the electrostatic interactions is by using PBC. Overall this test indicates that the new proposed scheme is both valid and efficient. In terms of computational time no additional overhead was noted when performing pseudo QM/MM calculation with or without PBC.

6.2 SiO_2

Let us now consider a realistic problem, a crystal of α-SiO_2 (α-quartz) in an orthorhombic cell, subject to periodic boundary conditions. Several QM/MM schemes have been proposed in the literature for silica-based systems[65–72], differing in the description of the quantum-classical interface and of the classical region. All of them treat the QM/MM long-range interaction with a truncation scheme, properly optimizing the charges of the H-atoms terminating the MM cluster or its shape in order to recover the correct long-range effects.

The MM crystal we used for this test is made up of 15552 atoms (5184 SiO_2 units) in an orthorhombic cell of 49.94, 57.66 and 63.49 Å. The system was optimized using the empirical pair potential of van Beest[26] which is known to provide a reliable description of bulk α-SiO_2[73]. A fragment of 160 atoms was treated at the QM level Fig. 9, describing the oxygen boundary atoms with a core charge increased by 0.4 in order to maintain the neutrality of the overall system. This boundary scheme will be described in details in Sec. 7.1. DFT calculations with Gödecker-Tetter-Hutter (GTH) pseudo-potentials[74] using local density approximation to describe the exchange-correlation functional were performed on the QM site using a cutoff of 200 Ry. We optimized the wave-function with and without the use of periodic boundary conditions. The results show that the use of periodicity is essential to treat highly ordered crystal structures. Without periodic boundary conditions we find the Kohn-Sham gap to be 0.12 eV which is much lower than the experimental band gap of about 9 eV[75,76] and than the computed Kohn-Sham gap of 5.8 eV[57]. Also the population analysis gives an indication that the lack of PBC leads to an incorrect description of the system. In fact by population analysis[47] we find that many oxygen atoms have a positive charge while some silicon atoms have a negative charge. If we use periodic boundary conditions, on the other hand, we find results that agree with those previously published. In particular, using PBC, we find for the Kohn-Sham band gap a value of 6.23 eV using the same computational parameters as in the case of non-PBC. The population analysis shows the proper charge distribution with charges close to +2.0 and -1.0 for silicon and oxygen respectively.

After removing the atom depicted in Fig. 9 from the same crystal structure, we studied

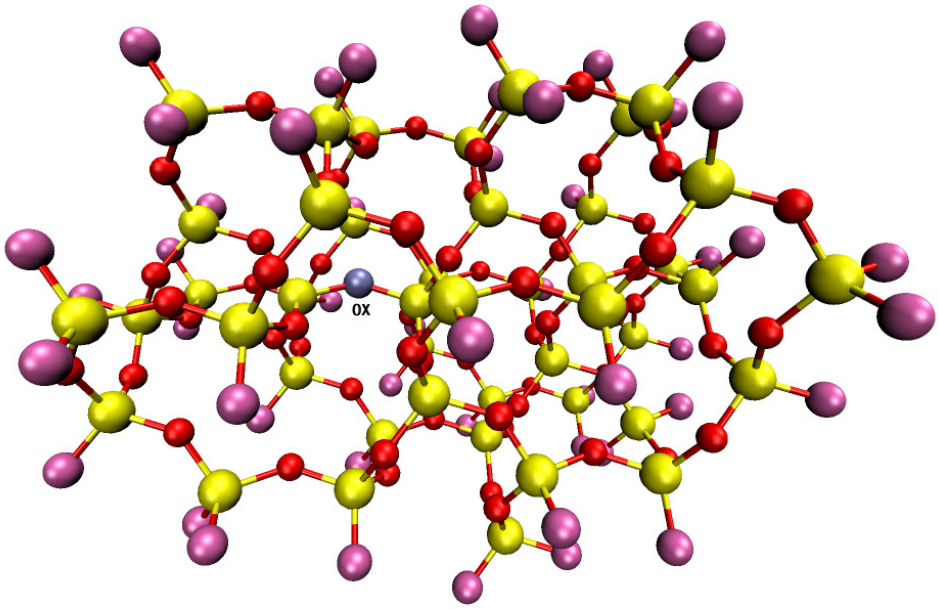


Figure 9. The picture shows the QM cluster. Silicon atoms in yellow, oxygen in red, boundary oxygen atoms (treated increasing the core charge by 0.4) in purple and in blue the oxygen atom (OX) removed to create the oxygen vacancy defect.

the charged oxygen vacancy defect in SiO_2 with the same computational setup used for stoichiometric SiO_2.

As for quartz the lack of PBC leads to an incorrect description for both the electronic structure and the population analysis. The use of the present scheme gives a Kohn-Sham band gap of 3.18 eV, as against the theoretical result[57] of 3.30 eV. The value obtained without PBC is 0.0089 eV. Unlike the other QM/MM schemes used for silica we do not use any additional charge to terminate the MM cluster and no particular attention was paid to the choice of its shape. The computational cost for the evaluation of the QM/MM-PBC electrostatic potential on this system accounts for 5% of the total CPU time of a single MD step.

7 QM/MM Study on Silica: Motivation

Silica is pervasive in present technologies, its applications ranging from optical fibers to metal-oxide-semiconductor devices and even to car tires. *Ab initio* studies have provided important insight on the properties of bulk phases, defects and surfaces of silica[77–80]. The usual approach in the *ab initio* modeling of condensed matter systems makes use of supercells with periodic boundary conditions containing at most few hundreds of atoms. However, if one is interested in the study of point defects as an impurity atom or a vacancy, the use of a full quantum periodic model implies an extremely high concentration of defects with a consequently strong defect-defect interaction due to the limited supercell

size. On the other hand, the use of a full quantum cluster model, popular in the chemistry community, suffers from other limitations, since also in this case the size of the system can not exceed, typically, one hundred atoms. Long-range electrostatic interactions are not kept into account and local relaxation associated, e.g. with defect formation, is partially hindered by the boundary atoms that have to be held fixed in order to prevent a global rearrangement of the cluster. However the properties to be addressed are often local in nature, such as the structure and spectroscopic properties of point defects or the chemical reactivity of specific sites. In these cases, a quantum mechanical description is necessary only for a small number of atoms around the site of interest, the rest of the system affects the local properties only via long-range electrostatic interactions and geometrical constraints. For this class of problems the QM/MM approach offers a satisfactory compromise between accuracy and computational efficiency[29]. By embedding a quantum mechanics calculation in a classical molecular mechanics model of the environment, the hybrid QM/MM schemes attempt to incorporate environmental effects at an atomistic level, including such influences as mechanical constraints, electrostatic perturbations and dielectric screening.

Several QM/MM schemes have been proposed in literature for silica-based systems[65–69,81,82,70,71], differing in the description of the quantum-classical interface and of the classical region.

Our specialization of the general QM/MM scheme to silica has been validated by computing the structural and dynamical properties of an oxygen vacancy in α-quartz, a prototypical defect in silica. For this benchmark case, we consider the effect on the accuracy of the description of several factors: i) the total size of the system (MM+QM); ii) the size of the QM subsystem; iii) the manner in which the valence at the boundary of the QM system is saturated; iv) the basis set. In this manner we provide an optimized setup for performing molecular dynamics QM/MM simulations in silicon dioxide. The quality of the description is demonstrated by performing a long molecular dynamics at finite temperature on the oxygen vacancy described with a minimal QM/MM model. We have found that convergence in the properties of the defect is already achieved with a very small quantum subsystem composed of eight atoms only. The combination of the QM/MM approach with the use of a localized basis set for the quantum cluster calculations makes long molecular dynamics simulations affordable at a low computational cost.

7.1 Modeling and Computational Details

The validity of a QM/MM scheme relies on few ingredients: the way in which bonded interactions between atoms in the classical and quantum region are described; the way in which electrostatic interaction between the two subsystem is treated; the quality of the classical force field. Finally, if the QM/MM scheme is aimed at performing molecular dynamics, a variational formulation of the total energy with respect to the atomic positions is also required.

For what concerns the classical force field, the most sophisticated QM/MM scheme presently available in literature is probably that proposed by Sulimov *et al.* in Ref. 66 where they use a classical region which includes up to several hundred polarizable atoms within a shell model , surrounded by a first region with non-polarizable point charges ions and by an outer region treated as a polarizable continuum. Here, we use the van Beest, Kramer, van Santen (BKS) potential[26]. Even this simplified description of the classical

subsystem does not affect significantly the accuracy of the description as we will show in Sec. 7.2 for the test case we considered. The condition of neutrality of the system imposes that $q_O = -\frac{1}{2}q_{Si}$. The charge of silicon ion is +2.4 e. The parameters for the short range interaction A,b,C are given in Ref. 26. This potential has been successfully applied to the study of the phase diagram of crystalline silica[83] and also provides a useful model of the amorphous phase[73].

Performing a QM/MM calculation on silica requires the description of a "pseudobond" between an MM Si and a QM O, or, vice versa, between an MM O and a QM Si. If, for example, the QM system is a six member ring, each Si atom will have two dangling bonds, whose valence has to be saturated in some way. Several strategies have been explored in the literature, for silica or similar systems, involving e.g. the introduction of extra "dummy" hydrogen atoms[84–86], or the ad hoc parameterization of a pseudo potential for the boundary atoms[20,87–90]. This latter approach has been adopted, for instance,in the QM/MM scheme for silica of Ref. 66. We here propose a manner for capping the QM region that does not require the introduction of extra atoms, and makes use of the ordinary pseudo potential also for the boundary atoms. We choose the QM region in such a manner that the boundary QM atoms are always oxygen atoms. The pseudo potential of these atoms (that from now we indicate with O*) is the ordinary one. To saturate the valence of the boundary oxygen atoms we add one electron for each O*, which would ideally come from the neighboring MM Si. In order to enforce global charge neutrality, we change the ionic charge of the boundary oxygen pseudo potential from 6 to 6.4. Hence, the total charge of the QM system is $(0.4-1)\, n_{O^*} = -0.6 n_{O^*}$ which is equal to the total charge of the classical atoms that have been replaced by the QM atoms. In fact, in a system in which all the silicon atoms are four-fold coordinated while all the O-atoms are two-fold coordinated the number of boundary oxygens is given by $n_{O^*} = 4n_{Si} - 2n_O$, where n_{Si}, n_O are the number of QM Si and QM O respectively. Therefore, since the charge of classical Si and O are 2.4 and -1.2 respectively, the total classical charge of the QM subsystem is indeed $2.4 n_{Si} - 1.2 n_O = -0.6\,(+4n_{Si} - 2n_O) = -0.6 n_{O^*}$, equal to the QM charge.

We perform DFT calculation in the local density approximation (LDA), by using Gaussian based pseudo potentials[74] with a DZVP atomic basis set and expanding the electron density in plane-waves with an energy cutoff of 240 Ry.

The interaction energy term, $E_{QM/MM}$ is expressed as:

$$E_{QM/MM}(\mathbf{r}^{QM}, \mathbf{r}^{MM}) = \sum_{i \in MM} q_i \int d\mathbf{r} \frac{\mathrm{erf}(\frac{|\mathbf{r}-\mathbf{r}_i^{MM}|}{r_{c,i}})\rho_{QM}(\mathbf{r})}{|\mathbf{r}-\mathbf{r}_i^{MM}|} + \sum_{\substack{i \in MM \\ j \in QM}} V_{\mathrm{NB}}(\mathbf{r}_i^{MM}, \mathbf{r}_j^{QM}) \tag{56}$$

where $\rho_{QM}(\mathbf{r})$ is the total (electronic plus nuclear) charge density of the quantum system. $V_{\mathrm{NB}}(\mathbf{r}_i^{MM}, \mathbf{r}_j^{QM})$ is the non–bonded interaction.

All the classical steric and electrostatic interactions between QM atoms are set to zero. Instead, a non–bonded[26] term is introduced between O^* and the first classical silicon atoms. The parameters of the interaction are obtained by performing a series of full QM calculations on a H_3Si-O-Si-O-Si-H_3 cluster by varying the distance between the central Si and one of the two oxygen atoms while keeping other angles and distances fixed (see Fig. 10). This distance dependence has then been fitted with the functional form of the BKS potential[26]. The parameters obtained with this procedure are A=603935406. K,

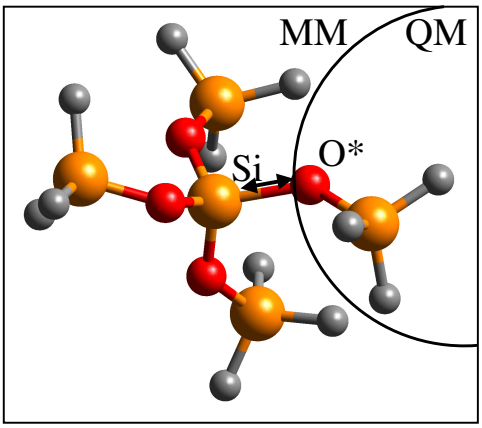


Figure 10. Structure of the small cluster used for the fitting of the short range potential between the boundary O^* and the first classical silicon. The arrows indicate the Si-O bond elongated. Si, O and H atoms are depicted in yellow, red, and grey, respectively.

b=5.6077 $Å^{-1}$, C=2244282. K/$Å^6$. These values are only slightly different from those of the van Beest, Kramer, van Santen potential.

In order to test the quality of our QM/MM model, we consider a big cluster of α-quartz made by an integer number of SiO_2 units saturated with hydrogens. The system is divided in two regions, one treated within the *ab initio* framework (QM region), the second treated by a classical force field (MM region). In real applications, the QM subregion should ideally be as small as possible for reasons of computational efficiency. With this goal in mind, we benchmark our QM/MM model by considering the structural properties of a 6-member ring embedded in MM SiO_2 and the formation energy of the neutral oxygen vacancy with QM subsystems of various size.

7.2 Validation of the QM/MM Approach

7.2.1 Geometry of the 6-Member Ring in α-Quartz.

We first consider a rather small SiO_2 cluster composed of 164 atoms saturated with H atoms. This system is chosen because it can be optimized at the full QM level. The QM region is a ring made by six member ring approximately at the center of this cluster (see Fig. 11). In order to reduce possible long-range electrostatic effects we optimized the charges of the classical H-atoms terminating the cluster in order to reproduce the full QM dipole moment. The position of the H-atoms and of the Si and O atoms connected to them are held fixed in the geometry optimization using both the QM/MM or MM Hamiltonian.

The difference between the QM/MM and the full QM geometry is used as a measure of the quality of the capping and of the QM/MM Hamiltonian.

By using the capping scheme described in Sec. 7.1, we perform a geometry optimization of the system using the QM/MM Hamiltonian (Eq. 56). The results are shown in Fig. 12, in which the full QM, MM and QM/MM structures are superimposed. The differences between these structures are small, especially for what concerns the QM subsystem.

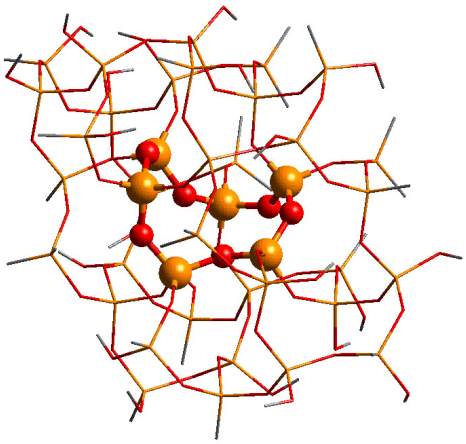


Figure 11. Structure of the cluster with a QM six-membered ring (depicted with spheres). The color code is the same as in Fig. 10.

The value of the root mean square deviation (RMSD) between the QM/MM and full QM geometries is computed for the QM subsystem, the full cluster and the boundary atoms. The results are reported in Tab. 3.

For a comparison, we also considered capping schemes in which the valence of the QM system is saturated by dummy hydrogen atoms. If the last QM atom is an oxygen, the H is placed in the direction of the first MM Si. Hence, the QM subsystem will be

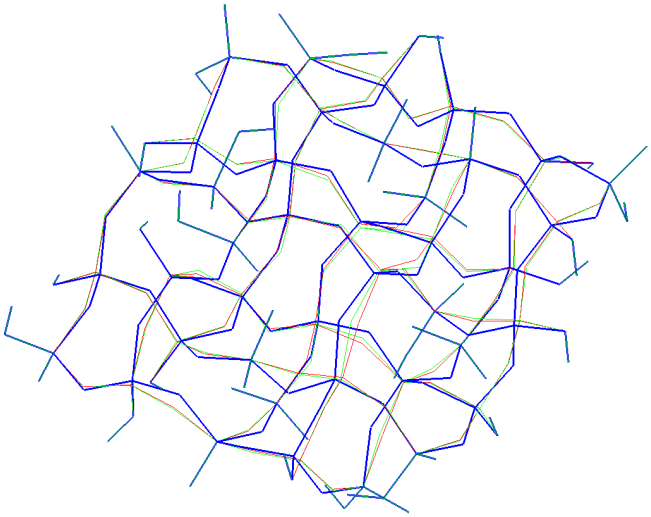


Figure 12. Superposition of the geometries obtained from a full quantum (blue line), a full classical (red line) and a QM/MM optimization (green line).

RMSD	RING [Å]	INTERFACE [Å]	FULL [Å]
MM	0.248	0.254	0.199
QM/MM O*	0.183	0.221	0.196
QM/MM O-H	0.275	0.403	0.254
QM/MM Si-H	0.191	0.276	0.206

Table 3. RMSD of the MM structure and the QM/MM O*, O-H, Si-H terminated with respect to the QM structure in three different case: in case RMSD_{RING} we compare only the 6SiO atoms of the ring; in case $\text{RMSD}_{INTERFACE}$ we compare the positions of all the QM atoms and the MM boundary atoms (for MM case we chose the same atoms of O* and O-H cases); in case RMSD_{FULL} we compare the positions of all the atoms.

terminated by -OH moieties (Model "OH"). If the last QM atom is a Si, the H is placed in the direction of the first MM oxygen (Model "H"). The RMSD for these two models are also reported in Tab. 3. The ratio of the SiO/SiH bond lengths for "H" capping or the ratio OH/SiO for "OH" capping have been fixed at the values determined in a preliminary full quantum optimization. The model "OH" shows a large RMSD for both the ring and interface regions. This is due to a large difference in the SiOH angle (115^o) of the silanol with respect to the SiOSi angle (145^o) in α-quartz, as already pointed out by by Sauer *et al* in Ref. 65. However the "OH" capping might perform better for small silica clusters with smaller SiOSi angles[81,82]. The model "SiH" performs better than the "OH" model, but still show large deviation in the interface region. However, since we have not attempted to reparameterize the short range potential at the interface as we did for the O* capping, we must say that there is still room for improvement for the "H" capping.

7.2.2 Formation Energy of an Oxygen Vacancy in α-Quartz.

We now consider the formation energy and the structure of the neutral oxygen vacancy defect in α-quartz. We use the experimental structural parameters: a=4.913 Å, c/a=1.100[91,48]. Removal of an oxygen atom produces a relaxation of the lattice with a formation of a Si-Si covalent bond, whose length is much shorter than the equilibrium Si-Si distance in a perfect lattice (3.08 Å). Theoretical studies report that the equilibrium distance of the Si-Si bond is in the range 2.3-2.6 Å[66,92,59,93,94,64]. The predicted value is strongly affected by the size of quantum system in cluster and periodic calculations. Also for the formation energy, the values reported in literature depend significantly on the model (full QM cluster, full QM periodic or QM/MM) and on the basis set. Boureau and Carniato[95] found that the formation energy of the neutral oxygen vacancy must be larger than 7.3 eV from purely thermodynamic arguments. Density-functional-theory calculations in periodic models give 6.97 eV[92], 7.85 eV[59], 9.6 eV[93] at the LDA level and 8.64 eV at the GGA level[93]. Hartree-Fock calculations on an isolated cluster give 6.7 eV and 5.5 eV with and without the d functions in the basis set and 8.5 eV including correlation energy at the MP2 level[94,64]. Sulimov *et al.*[66], using a QM/MM approach with the QM region treated at the unrestricted Hartree-Fock level (UHF), have obtained a formation energy of 6.08 eV with the 6-31G* basis set used, which corresponds to ours.

They have also found that the formation of a Si-Si bond induces a strong anisotropic relaxation of the lattice that extends up to 13 Å from the defect. They also find a Si-Si distance of 2.32-2.40 Å depending from the basis set used (2.37 Å with a basis set

equivalent to ours).

We compute the formation energy using the QM/MM Hamiltonian described in Sec. 7.1, considering the effect of several factors that could influence the accuracy of the calculation.

We first consider the effect of the size of QM subsystem, computing the formation energy for the three QM subsystems shown in Fig. 13. The smaller system (9 atoms) is the

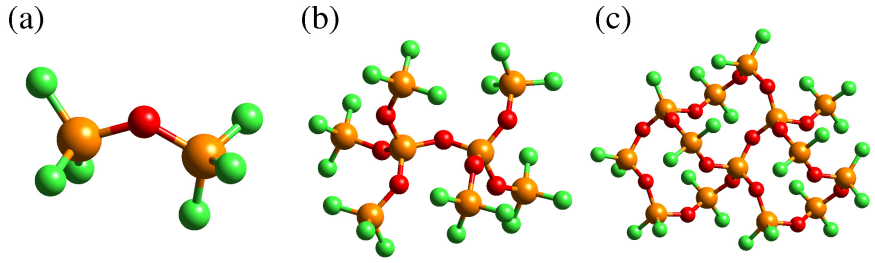


Figure 13. Structure of three different QM clusters used for the study of the oxygen vacancy. (a) $Si_2OO_6^*$, (b) $Si_8O_7O_{18}^*$, (c) $Si_{14}O_{16}O_{24}^*$. Si, O and O* are depicted in yellow, red, and green, respectively.

$O\left(SiO_3^*\right)_2$ moiety. The oxygen atoms in the SiO_3^* groups are boundary atoms, while the central O is removed to generate the vacancy. The average and large QM subsystems are composed of all the SiO_2 units within three and five bond separation from the oxygen that is removed, i.e. 33 or 54 atoms, respectively. For all the three cases, the QM subsystem is embedded in a classical cluster composed of 508 SiO_2 units. The Si and the O atoms at the boundary of the classical cluster are saturated by hydrogen ions of charges -0.6 and +0.6, respectively.

The vacancy formation energy is given by

$$\Delta E^{form} = E(O) + E(vacancy) - E(quartz). \tag{57}$$

The energy of the isolated oxygen $E\left(O\right)$ is obtained as $E\left(O\right) = \frac{1}{2}\left(E\left(O_2\right) + E^{diss}\left(O_2\right)\right)$ where $E\left(O_2\right)$ is the *ab initio* total energy of the O_2 molecule in the triplet state and $E^{diss}\left(O_2\right) = 5.16$ eV is the experimental dissociation energy of O_2[96]. The correction due to the basis set superposition error (BSSE) is about 0.1 eV. The results are shown in Tab. 4. We estimated the BSSE with the counterpoise correction[97] separately for the three energy terms in Eq. 57 as follows: i) the correction to perfect quartz is the difference in total energy due to the addition to the basis set of a ghost oxygen atom which forms a O_2 molecule with the oxygen removed in the vacancy formation; ii) E(O) is calculated with a full basis set of i); iii) the correction to the E(vacancy) is obtained by using the full basis set of i) in the unrelaxed vacancy configuration.

We have also checked the dependence of the geometry and formation energy on the basis set by performing additional calculations on the smaller cluster ($Si_2OO_6^*$) with the TZVP and TZV2P basis sets. The results are reported in Tab. 5 and show that the DZVP basis set is accurate enough for structural properties, but formation energies change sizably with the basis set as already found in Hartree-Fock calculation with smaller basis sets in Ref. 66.

	dist Si-Si [Å]	ΔE [eV]	ΔE^{CP} [eV]
QM+MM regions:	1764 atoms		
(A) $Si_2OO_6^*$	2.35	7.34	-0.11
(B) $Si_8O_7O_{18}^*$	2.36	7.31	-0.12
(C) $Si_{14}O_{16}O_{24}^*$	2.40	7.36	-0.13
QM+MM regions:	773 atoms		
(B) $Si_8O_7O_{18}^*$	2.35	7.36	-0.12

Table 4. Si-Si bond length, vacancy formation energy (ΔE) and the counterpoise correction ΔE^{CP} (included in ΔE) for different size of the QM and MM regions. The DZVP basis set has been used.

	dist. Si-Si [Å]	ΔE [eV]	ΔE^{CP} [eV]
DZVP	2.35	7.34	-0.11
TZVP	2.34	7.44	-0.08
TZV2P	2.34	7.91	-0.07

Table 5. Si-Si bond length, vacancy formation energy (ΔE) and the counterpoise correction ΔE^{CP} (included in ΔE) for different basis sets.

As a final remark, we note that for a crystalline system the Madelung field in the quantum region would strongly depend on the value of the classical charges in the MM region. Different MM models with different charges might provide similar bulk properties, e.g. bulk structure of the glass, along with different local Madelung fields. Therefore, particular care must be paid in using QM/MM when the properties of charged defects are addressed, e.g. the heterolytic breaking of a siloxane bond. The QM/MM scheme we propose is expected to correctly describe the elastic response of the system surrounding the quantum region. Its applicability to study of any other local properties of the quantum region which would depend on the details of the Madelung field must be carefully checked. In this respect, the BKS potentials we have used is probably better than others available in literature also in describing the local Madelung field since the classical charges are fitted on *ab initio* data. In our benchmark application the vacancy is a neutral defect and the problems outlined above are probably less severe. To check further this point, we have computed the formation energy of the unrelaxed oxygen vacancy in model B (Cfr. Tab. 4) by changing the charge of classical silicon from 2.4 (BKS) to several values in the range 1.6-3.6. The charges of the hydrogen atoms capping the MM cluster and of the boundary quantum oxygen atoms have been scaled accordingly. It turns out that the change in the formation energy of the unrelaxed oxygen vacancy is always smaller than 20 meV.

7.2.3 Molecular Dynamics

In order to check the validity of our setup we have performed molecular dynamics simulations of the QM/MM system of size 8/1764 starting from the structure of the defect ($Si_2O_6^*$) optimized with the DZVP basis set. We have first equilibrated the system at high temperature (1000 K) by velocity rescaling for 0.3 ps. Observables are measured by averaging over a run 10 ps long. The time step used in the velocity Verlet algorithm is 0.5 fs.

In Fig. 14a, we report the fluctuation in the potential energy and the total energy, constant of motion, of our microcanonical simulation. The fluctuation in the constant of motion is two order of magnitude smaller than the thermal fluctuations in the potential energy which prove the robustness of our scheme. The Si-Si bond is stable and undergoes stretching deformation with a characteristic frequency that we have identified by Fourier transforming the autocorrelation function $< \dot{\mathbf{R}}(t)\dot{\mathbf{R}}(0) >$ where $\dot{\mathbf{R}}(t)$ is the instantaneous Si-Si bond vector. The correlation function is computed up to 2.5 ps by averaging over three independent sections of a run 10 ps long. The results are well converged up to 0.3 ps. For longer times, a longer simulation run would be needed. The autocorrelation function is therefore windowed with a Fermi-Dirac function which smoothly brings $< \dot{\mathbf{R}}(t)\dot{\mathbf{R}}(0) >$ to zero above 0.25 ps. The resulting power spectrum is shown in Fig. 14b. The peak at

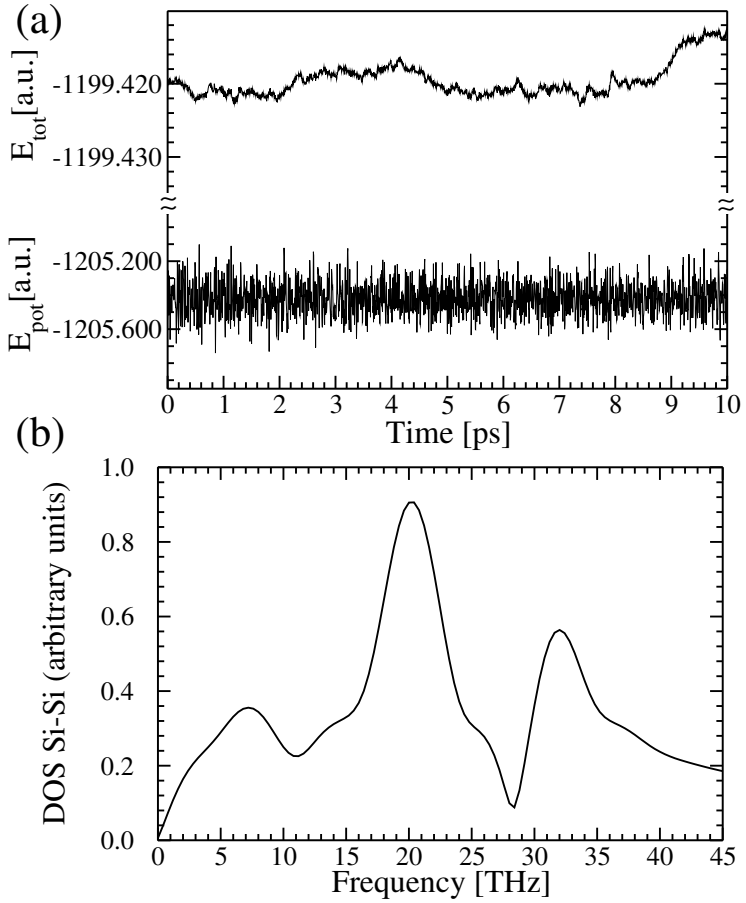


Figure 14. Potential energy (E_{pot}) and total energy (E_{tot}) as a function of time in the molecular dynamics simulation. b) The power spectrum of the velocity-velocity autocorrelation function for the Si-Si bond length only (see text).

$\sim$20.5 THz corresponds to the main stretching mode of The Si-Si bond. Its position in frequency (20.5 THz) compares well with that of the Si-Si stretching mode of the disilane molecule H_5C_2OSi-$SiOC_2H_5$ we have identified at 21.3 THz from a molecular dynamics simulation 2.2 ps long at 300 K or at 22.4 THz from the diagonalization of the dynamical matrix computed within linear response theory[98] with the code CPMD[5,99]. The peak at 20.5 THz is also in good agreement with a prominent structure in the vibrational spectra of the Si-Si bond which emerges from the difference in the vibrational density of states, computed fully *ab initio* (LDA) in Ref. 100, for two periodic models (36 atoms large) of α-quartz with and without the oxygen vacancy (see Fig.1 of Ref. 100). The computational load for the molecular dynamics simulation on a single Opteron processor (2.2 GHz) is 28 hours/ps for the small quantum cluster ($Si_2O_6^*$) and 72 hours/ps for the larger cluster ($Si_8O_6O_{24}^*$) both with a classical cluster of 1764 atoms.

8 Conclusion

In these pages, I reviewed an algorithm for evaluating the QM/MM coupling term with a fast linear scaling implementation both for periodic and non–periodic systems. The main result is the dropping of the prefactor in the linear scaling, with a gain in the number of floating point operations proportional to $2^{3(N_{grid}-1)}$, where N_{grid} is the number of grid levels used in the multi-grid framework. The evaluation of the electrostatic potential on a grid is proportional to the number of MM atoms times the number of grid points. In real systems the linear scaling evaluation of the potential is therefore characterized by a prefactor $\approx 10^6$. In this scheme the prefactor is instead $\approx 10^3$. The number of floating point operations is reduced several orders of magnitude and the computational time is 10-100 times smaller.

The algorithm is presently implemented in the package CP2K, released under GPL license and freely available on the internet[1].

The performance analysis confirms the present algorithm as the state of the art for the evaluation of QM/MM interaction coupling within a GPW scheme. Moreover, at variance with the majority of present-day QM/MM methods, our scheme does not rely on electrostatic cutoffs and so avoids all related problems. Consequently, the present method offers a fast, easy-to-use code for QM/MM calculations of large biological and inorganic systems.

Finally, I have shown how to employ the QM/MM scheme to model Silica. In this framework, the capping of the QM region consist of boundary oxygen atoms with a modified charge to enforce total charge neutrality. This scheme makes long molecular dynamics simulations, needed for instance to simulate local chemical reactivity, easily affordable. The method has been tested calculating structural and dynamical properties of an oxygen vacancy in α-quartz. We have found that good convergence in the Si-Si bond length and formation energy is achieved by using a quantum cluster as small as eight atoms in size.

Acknowledgments

The work presented in this chapter is the outcome of my PhD thesis. The support of Fawzi Mohamed, Alessandro Laio, Michele Parrinello was vital to this project. Federico Zipoli deserves all my gratitude for being one of the very first beta–tester of the QM/MM implementation, being the principal investigator of the Silica project.

Appendix A: Derivation of Coulomb Potential for Delocalized Point-Like Charges

To perform the integration in Eq. 7, we consider the integral:

$$I(\mathbf{r}) = \int d^3r' \frac{|\phi(\mathbf{r}' - \mathbf{r}_a)|^2}{|\mathbf{r}' - \mathbf{r}|} \tag{58}$$

Using Eq. 6 for ϕ and taking the Fourier transformations for $\exp^{-2\xi|\mathbf{r}'-\mathbf{r}_a|}$ and $\frac{1}{|\mathbf{r}'-\mathbf{r}_a|}$, we obtain

$$I = \frac{q_a\xi^3}{\pi} \int d^3r' \left(\frac{2\xi}{\pi^2} \int d^3p \frac{\exp^{\imath\mathbf{p}\cdot(\mathbf{r}'-\mathbf{r}_a)}}{(p^2+4\xi^2)^2} \right) \cdot \left(\frac{1}{2\pi^2} \int d^3q \frac{\exp^{\imath\mathbf{q}\cdot(\mathbf{r}'-\mathbf{r})}}{q^2} \right) \tag{59}$$

Rearranging and performing the integration over d^3r', we get

$$I = \frac{q_a\xi^3}{\pi} \frac{2\xi}{\pi^2} \frac{1}{2\pi^2} (2\pi)^3 \int d^3p \int d^3q \delta(\mathbf{p}-\mathbf{q}) * \frac{\exp^{-\imath\mathbf{p}\cdot\mathbf{r}_a}}{(p^2+4\xi^2)^2} \frac{\exp^{\imath\mathbf{q}\cdot\mathbf{r}}}{q^2} \tag{60}$$

Performing the integration over d^3q using δ-function integration one obtains

$$I = \frac{q_a\xi^3}{\pi} \frac{8\xi}{\pi} \int d^3p \frac{\exp^{\imath\mathbf{p}\cdot(\mathbf{r}-\mathbf{r}_a)}}{p^2 * (p^2+4\xi^2)^2} \tag{61}$$

Decomposing $\frac{1}{p^2*(p^2+4\xi^2)^2}$ we rewrite the above integral as

$$I = \frac{q_a\xi^3}{\pi} \frac{8\xi}{\pi} \int d^3p \left[\frac{1}{\zeta^4 p^2} - \frac{1}{\zeta^4 * (p^2+\zeta^2)} - \frac{1}{\zeta^2 * (p^2+\zeta^2)^2} \right] \cdot \exp^{\imath\mathbf{p}\cdot(\mathbf{r}-\mathbf{r}_a)} \tag{62}$$

where $\zeta = 2\xi$. Taking the inverse Fourier transforms for all the three integrals and simplifying for the constants, we finally obtain

$$I = q_a \left[\frac{1}{|\mathbf{r}-\mathbf{r}_a|} - \frac{\exp^{-2\xi|\mathbf{r}-\mathbf{r}_a|}}{|\mathbf{r}-\mathbf{r}_a|} - \xi \exp^{-2\xi|\mathbf{r}-\mathbf{r}_a|} \right] \tag{63}$$

Appendix B: Derivation of the Long-Range QM/MM Potential

The effect of the periodic copies of the MM sub-system is only in the long-range term, and it comes entirely from the residual function $R_{low}(\mathbf{r}, \mathbf{r}_a)$ of Eq. 45:

$$V_{recip}^{QM/MM}(\mathbf{r}, \mathbf{r}_a) = \sum_{\mathbf{L}}^{\infty} {}' \sum_a v_a^{recip} = \sum_{\mathbf{L}}^{\infty} {}' \sum_a R_{low}(|\mathbf{r} - \mathbf{r}_a + \mathbf{L}|) \tag{64}$$

This summation has the same convergence properties as the Ewald series, and can be efficiently computed in the reciprocal space. To derive the expression of this modified Ewald sum, let us assume we know the analytical expression of the density $\sigma(\mathbf{r}, \mathbf{r}_a)$ originating from the atomic potential R_{low}. The potential at point $\mathbf{r}_i$ due to the charge distribution $\sigma(\mathbf{r}, \mathbf{r}_a)$ is:

$$\begin{aligned} V_{recip}^{QM/MM}(\mathbf{r}_i, \mathbf{r}_a) &= \int d\mathbf{r} \frac{\sigma(\mathbf{r}+\mathbf{r}_i, \mathbf{r}_a)}{\mathbf{r}} \\ &= L^{-3} \int d\mathbf{r} \sum_{\mathbf{k}}^{k_{cut}} {}' \frac{\tilde{\sigma}(\mathbf{k}) \exp[-\imath 2\pi \mathbf{k}(\mathbf{r}+\mathbf{r}_i-\mathbf{r}_a)]}{\mathbf{r}} \end{aligned} \tag{65}$$

The use of the identity[101]

$$\begin{aligned}
&\int d\mathbf{r}\frac{\exp[-\imath 2\pi\mathbf{k}(\mathbf{r}+\mathbf{r}_i-\mathbf{r}_a)]}{\mathbf{r}} \\
&= \int_0^\infty r dr \int_0^{2\pi} d\phi \int_0^\pi \sin\theta d\theta \exp[-\imath 2\pi|\mathbf{k}||\mathbf{r}+\mathbf{r}_i-\mathbf{r}_a|\cos\theta] \\
&= \frac{4\pi}{k^2}\cos\left[2\pi\mathbf{k}\cdot(\mathbf{r}_i-\mathbf{r}_a)\right]
\end{aligned} \tag{66}$$

in Eq. 65 leads to

$$V_{recip}^{QM/MM}(\mathbf{r}_i,\mathbf{r}_a) = 4\pi L^{-3}\sum_{\mathbf{k}}^{k_{cut}\ \prime}\frac{\tilde{\sigma}(\mathbf{k})}{k^2}\cos\left[2\pi\mathbf{k}\cdot(\mathbf{r}_i-\mathbf{r}_a)\right] \tag{67}$$

Using the Maxwell equation $\nabla^2 V = 4\pi\rho$ and its representation in Fourier space, the term in Eq. 67

$$4\pi\frac{\tilde{\sigma}(\mathbf{k})}{k^2} = \tilde{R}_{low}(\mathbf{k}) \tag{68}$$

is the Fourier transform of the potential originated by the density of charge $\sigma(\mathbf{r},\mathbf{r}_a)$. Then the previous equation can be written

$$V_{recip}^{QM/MM}(\mathbf{r}_i,\mathbf{r}_a) = L^{-3}\sum_{\mathbf{k}}^{k_{cut}\ \prime}\sum_a \tilde{R}_{low}(\mathbf{k})q_a\cos\left[2\pi\mathbf{k}\cdot(\mathbf{r}_i-\mathbf{r}_a)\right] \tag{69}$$

Appendix C: Splines

C.1 Multi Grid

Multi grid methods instead of just a fine grid $\mathcal{G}_f$ use other coarser grids. These grid levels are ordered from the most coarse $\mathcal{G}_c = \mathcal{G}_1$ to the finest $\mathcal{G}_f$ ($1 = c \leq f$). In 3D all the coarser grids do not cost much in term of memory (typically 0.14-0.4 times the memory of the fine grid). Adding these extra grids is useful because each one can represent a given wavelength in an optimal way (i.e. with a minimal number of points), and perform operations on this wavelength efficiently. Typically operations on each grid level are local and work on patches of neighboring points, and after a series of them one collects the result on the fine grid.

Multi grids methods can be used to solve linear equations on a grid, for example partial differential equations, but they aren't yet used for this purpose in CP2K . We use multi grids to transfer the density from the Gaussian basis set to the grid trying to use a constant number of points per Gaussian, as described in Ref. 45, and in QM/MM to transfer the MM potential on the grid.

Multi grid is interesting only if there is an efficient way to transfer the operations done on one grid level to the others. For $i < j$ the transfer functions

$$P_j^i: \quad \mathcal{G}_i \rightarrow \mathcal{G}_j \tag{70}$$

$$R_i^j: \quad \mathcal{G}_j \rightarrow \mathcal{G}_i \tag{71}$$

are called prolongation and restriction respectively.

If one wants that the integration of a function h defined on a finer grid $\mathcal{G}_j$ with a function g defined on a coarser grid $\mathcal{G}_i$ give the same result both transferring g to the fine grid and summing there or (more efficiently) transferring h to the coarse grid and then summing there one has

$$\langle P_j^i g, h\rangle = \langle g, R_i^j h\rangle \tag{72}$$

i.e. the projection is the dual of the restriction.

The prolongation can be seen as an interpolation: given the values on a coarse grid try to find the values on a finer grid. In general one can also imagine a continuous function that underlies the prolongation operation. A very good method for grids with periodic boundary condition is the G-space interpolation. With a fast Fourier transform (FFT) one can find the G-space representation $\hat{n}_{ijk}$ of the points on the grid. Then a continuous representation of them would be

$$n(\boldsymbol{r}) = \sum_{ijk} \hat{n}_{ijk} exp(\boldsymbol{G}(i,j,k)\cdot\boldsymbol{r}), \tag{73}$$

where $\boldsymbol{G}(i,j,k) = 2\pi h^{-1}[i,j,k]$, h^{-1} is the inverse of the cell vectors matrix, and i,j,k are evenly distributed between the positive and negative values. The G-space interpolation can be performed directly in the G-space, without going in the direct space. Indeed the $\boldsymbol{G}(i,j,k)$ of the coarser grid are a subset of the ones of the finer grid, and the mapping is trivial, taking care that for an even number of grid points you assign half the value to $N/2$ and half to $-N/2$.

The continuous function underlying the G-space interpolation is C^∞ (i.e. smooth, infinitely often differentiable), and is the best interpolation scheme (with respect to L^2 norm) for points that come from a periodic C^∞ function. Unfortunately if the points come from a function which is not smooth or for a non periodic function this is no longer true.

In CP2K non smoothness is present because at the core there is a jump in the derivative (cusp condition), and the exchange-correlation functionals, especially the gradient corrected ones, exacerbate the problem. This is due also to the pseudo potential we use in CP2K[30]. Also introducing a cutoff for the Gaussian loses their smoothness. It was in this setting that we initially introduced the spline approach. This turned out to be more useful that we thought and an extension of it was used to cope with the non periodicity (with respect to the QM cell) of the potential in a QM/MM setting.

C.2 Periodic Uniform Splines

A uniform cardinal B-Spline of order 3 in 3d is a function $\boldsymbol{R}^3 \to \boldsymbol{R}$

$$f(x,y,z) = \sum_{ijk} c_{ijk} N^3(x-i) N^3(y-j) N^3(z-k), \tag{74}$$

that is controlled by the coefficients c_{ijk}.

N^3 is a piecewise polynomial function in C^2 with compact support that can be seen as the convolution of the characteristic function of $[-1/2, 1/2]$ ($\chi_{[-1/2,1/2]}$) with itself three

times.

$$N^3(t) = \begin{cases} \frac{1}{6}(t+2)^3 & -2 \leq t < -1 \\ -\frac{1}{2}t^3 - t^2 + \frac{2}{3} & -1 \leq t < 0 \\ \frac{1}{2}t^3 - t^2 + \frac{2}{3} & 0 \leq t < 1 \\ -\frac{1}{6}(t-2)^3 & 1 \leq t < 2 \\ 0 & \text{otherwise} \end{cases} \tag{75}$$

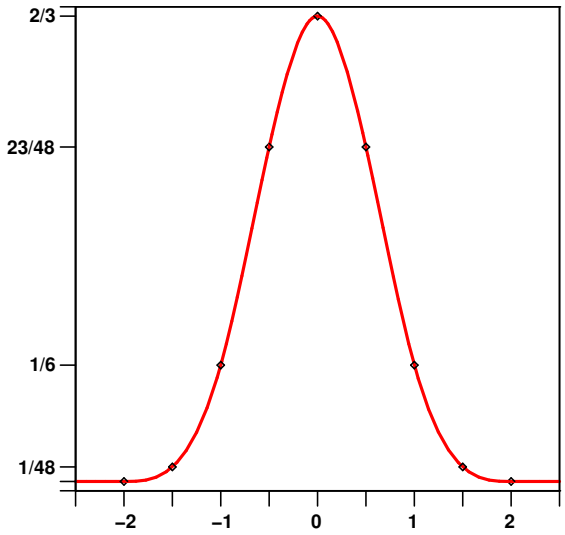


Figure 15. The N^3 function.

C.3 Periodic Prolongation/Restriction

With this the prolongation operation can be defined as follow:

1. find the coefficients c_{ijk} that interpolate the values v_{ijk} on the coarse grid
2. evaluate the spline Eq. 74 on the fine grid to obtain the final values w_{ijk}

We define the function

$$S^i : \mathcal{G}_i \to \mathcal{G}_i \qquad (S^i)_{klm,nop} = N^3(||k-n||)N^3(||l-o||)N^3(||m-p||) \tag{76}$$

where $||x||$ is introduced because of periodic boundary conditions, and means the smallest distance, for example for the dimension x

$$||x|| = ((x + N_x/2) \mod N_x) - N_x/2, \tag{77}$$

where N_x is the number of grid points in the dimension of x. S^i maps the coefficients c_{ijk} to the values v_{ijk}. This matrix is very sparse because the N^3 is different from 0 only for the nearest neighbor, i.e. for an integer value i

$$N^3(i) = \begin{cases} \frac{1}{6} \text{ if } i = -1 \\ \frac{2}{3} \text{ if } i = 0 \\ \frac{1}{6} \text{ if } i = 1 \\ 0 \text{ otherwise} \end{cases} \tag{78}$$

The application $S^i c$ can be seen as the convolution of the grid with the 3x3x3 stencil (indexed from -1 to 1) with values

$$S_4 = \frac{2}{3}^{3-|i|-|j|-|k|} \frac{1}{6}^{|i|+|j|+|k|} \tag{79}$$

which has values

$$[[\frac{8}{27}, \frac{2}{27}, \frac{1}{54}, \frac{1}{216}]] \tag{80}$$

for center, face centers,edges, and vertices of the 3x3x3 cube.

Then the first step of the prolongation is

$$c = (S^i)^{-1} v \tag{81}$$

which we calculate iteratively with a conjugated gradient solver, using

$$[[2 - \frac{8}{27}, -\frac{2}{27}, -\frac{1}{54}, -\frac{1}{216}]] \tag{82}$$

as approximate inverse for the first guess, and

$$[[4.096, -1.28, 0.4, -0.125]] \tag{83}$$

as pre-conditioner. The pre-conditioner is generated by the 1d-values $[-1.6/4, 1.6, -1.6/4]$ in each direction. It was found by minimizing the condition number of S^i multiplied by operators generated from 1d-values, and then (slightly) further optimized in the program. With this in 10-15 iterations, independently of the size of the grid, a convergence to less than 10^{-10} for both argument and residual can be achieved.

To evaluate the spline on the fine grid we use commensurate grids for efficiency reasons, which means that each grid has exactly the double of the number of points in every direction than the previous grid level. In this case it is useful to introduce the (rectangular) matrix.

$$(T^i_{i+1})_{klm,nop} = N^3(\frac{k}{2} - n) N^3(\frac{l}{2} - o) N^3(\frac{m}{2} - p) \tag{84}$$

which is very sparse as for half integer the only nonzero values are

$$N^3(\frac{i}{2})_{i=-4..4} = [0, \frac{1}{48}, \frac{1}{6}, \frac{23}{48}, \frac{2}{3}, \frac{23}{48}, \frac{1}{6}, \frac{1}{48}, 0] \tag{85}$$

Thus we have

$$P^i_{i+1} = T^i_{i+1} (S^i)^{-1}, \tag{86}$$

and

$$R^{i+1}_i = (P^i_{i+1})^T = (S^i)^{-1} (T^i_{i+1})^T. \tag{87}$$

The interpolation between the other grid levels can be defined as the product of the cascade prolongation/restrictions from grid i to grid j

$$P_j^i = \prod_{k=j-1}^{i} P_{k+1}^k = P_j^{j-1}...P_{i+2}^{i+1}P_{i+1}^i, \tag{88}$$

and

$$R_i^j = \prod_{k=i}^{j-1} R_k^{k+1} = R_i^{i+1}R_{i+1}^{i+2}...R_{j-1}^j. \tag{89}$$

This approach works very well with periodic boundary conditions. The coefficients of the spline can be seen as the G-space coefficients of a Fourier transform. Like them they depend in a unique and global way from the values on the grid (direct space): any coefficient depends on the values of all the grid, but with the splines the weight of far away points decreases faster than with G-space interpolation, splines are more localized.

The coefficients define a continuous function that on the grid has exactly the values of the direct space, but that is defined everywhere, not just on the grid, and thus they can be used to interpolate the values, or transfer the function between grid levels. The continuous function defined by the cubic splines is C^2 (twice continuously differentiable). This is not optimal to interpolate smooth functions, but if the function to interpolate is not so regular (due for example to cutoff effects, or numerical instabilities) then the spline interpolation becomes better.

C.4 Non-Periodic Uniform Splines

If one wants to go beyond the periodic boundary conditions the function N^3 cannot be used for the coefficients close to the border. Indeed using the N^3 function would force the function to go at 0 and with derivative 0 two units after the border, and what is worse (one can argue that what happens beyond the border is not relevant and is an artifact) a simple linear function cannot be interpolated exactly. This gives rise to border effects that cannot be neglected. This problem is important for QM/MM where the potential generated by the MM atoms is not periodic with respect to the QM cell. As already stated the solution is to modify the form of the N^3 functions for the coefficients close to the border.

To find out how to modify the functions we will look at a generalization of the uniform cardinal splines. To simplify the discussion we will first look at a non-uniform B-Spline of order 3 in just 1 dimension. This is a parametric 1d line in a 2d dimensional space, i.e. a $\boldsymbol{R} \to \boldsymbol{R}^2$ function

$$g(u) = \sum_i P_i N^3(u - i), \tag{90}$$

where P_i is an array (indexed by the integer i) of 2-dimensional vectors.

This looks complicated, but if one sets

$$P_i = [i, v_i] \tag{91}$$

then if we call the first component of g, x and the second h

$$[x(u), h(u)] := g(u), \tag{92}$$

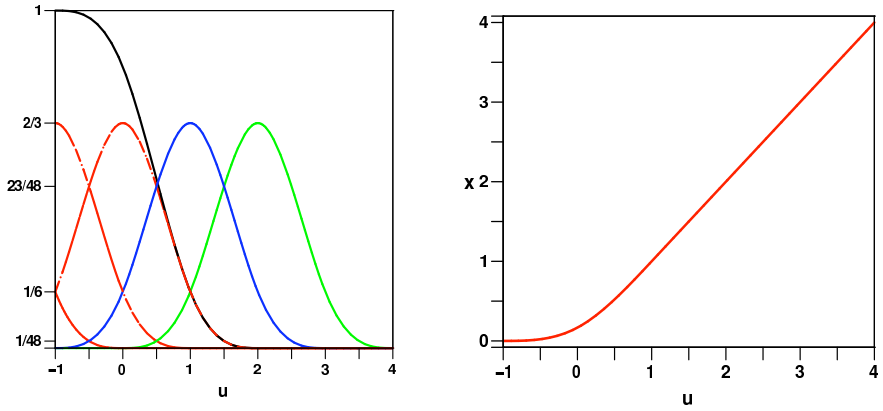


Figure 16. The left panel shows the weights of the splines for $i = -2..2$. The dotted red splines (with $i = -2..0$) have the same coefficient, so they have been summed up into the continuous black line. The right panel shows the value of x as function of u.

we see that the mapping $x(u)$ is the identity and

$$h(u) = h(x) = \sum_i v_i N(x - i) \tag{93}$$

and so h is just a uniform cardinal spline.

Assuming that the lower boundary is at 0, we want to look at

$$P_i = [\max(0, i), v_{\max(0,i)}]. \tag{94}$$

As we can see for $u \geq 1$ $u = x$, but for smaller values the correspondence breaks and the function gets really parametric. x begins to change more and more slowly, and finally freezes at 0 when u reaches -1. Now the correct way to redefine the $N^3(x-i)$ to functions $M_i(x)$ for i close to the border (i.e. to 0) is

$$\begin{aligned} M_0(x(u)) &= N^3(u+2) + N^3(u+1) + N^3(u) \\ M_1(x(u)) &= N^3(u-1) \\ M_2(x(u)) &= N^3(u-2) \end{aligned} \tag{95}$$

and for $i > 2$ $M_i(x) = N^3(x - i)$.

To be able to directly represent $M_{0..2}(x)$ one has to invert $x(u)$

$$x^{-1}(t) = \begin{cases} \text{undefined} & t < 0 \\ \sqrt[3]{6u} - 1 & t < \frac{1}{6} \\ 2\sqrt{2}\cos\left(\frac{1}{3}(\pi + \arccos(\frac{3\sqrt{2}(t-1)}{4})\right) + 1 & t < 1 \\ t & t \geq 1 \end{cases} \tag{96}$$

With an explicit inverse one obtains a direct representation of the functions $M_{0..2}$ shown in Fig. 17 We see that for the evaluation on a grid with spacing 1 only the weight

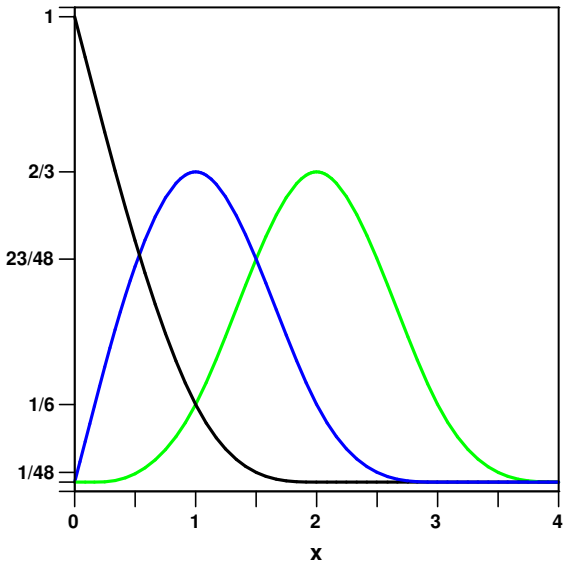


Figure 17. The border functions $M_{0..2}$ as function of x.

exactly at the border has to be changed (to 1), whereas for an uniform refinement, i.e. to prolongate to a grid with spacing $1/2$, the border and the points just before it have to be changed. Approximately $M_{0...2}$ have the following values at the important points:

$$\begin{aligned} M_0(0) &= 1 & M_0(\tfrac{1}{2}) &= 0.51797770339331435652 9532 \\ M_1(0) &= 0 & M_1(\tfrac{1}{2}) &= 0.46404459321337128694 0937 \\ M_2(0) &= 0 & M_2(\tfrac{1}{2}) &= 0.01797770339331435652 9531. \end{aligned} \tag{97}$$

Thus using the weights given by M_0, M_1, M_2 instead of the ones given by N^3 at the border the simplicity of the uniform spline schema can be kept and linear functions can be correctly interpolated. The upper border is just symmetric.

In 3d we have to look at non-uniform B-Spline of order 3 in 3 dimensions, which are parametric 3d surfaces in a 4d dimensional space, i.e. a $\boldsymbol{R}^3 \to \boldsymbol{R}^4$ function

$$g(u,v,t) = \sum_{ijk} P_{ijk} N^3(u-i) N^3(v-j) N^3(t-k), \tag{98}$$

where P_{ijk} is a 3d grid (indexed by the integer i, j, k) of 4-dimensional vectors.

Looking at it one can see that the fact that the weight functions are just a direct product of the 1d weighting functions is preserved with boundaries along the border of a box. Assuming that the lower left corner of the box is for $(i,j,k) = (0,0,0)$

$$f(x,y,z) = \sum_{ijk} v_{ijk} M_i(x) M_j(y) M_k(z), \tag{99}$$

with M_i as defined in the 1d case.

C.5 Non-Periodic Prolongation/Restriction

The prolongation and restriction operation can be calculated just as before

$$P_j^i = \tilde{T}_j^i (\tilde{S}^i)^{-1} \tag{100}$$

$$R_i^j = (P_j^i)^T = ((\tilde{S}^i)^T)^{-1} (\tilde{T}_j^i)^T \tag{101}$$

where $\tilde{S}^i$ and $\tilde{T}_j^i$ are different from S^i and T_j^i because they use $M_i(x)$ instead of the $N^3(x-i)$. This means that $\tilde{S}^i$ differs from S^i only at the border where in each dimension 1 is used instead of $2/3$ as weight and the $1/6$ contribution from the neighboring point is ignored. This breaks the symmetry of S^i and makes the sum of the contributions of the weights c close to the border differ from 1. Likewise $\tilde{T}_j^i$ differs from T_j^ionly at the border using the values of Eq. 97.

The inversion of $\tilde{S}^i$ is performed using the same approximate inverse as in the non-periodic case, but setting the weight to 1 instead of $2/3$ at the border, and removing the $1/6$ contribution from the point next to the border, as with $\tilde{S}$. For the pre-conditioner the contribution from the weight c at the border are scaled in such a way that at the value at the border is 1 (i.e. not just setting the border to one, but also changing the contribution to the close-by v. With this method the same performance as in the periodic case can be achieved on big grids: ≈ 12 iterations for 10^{-10} accuracy, ≈ 20 for machine accuracy (10^{-14}). For small grids other approximate inverse and pre-conditioners (not based on the the periodic solution) would be better, but $\approx 1/3$ more iterations on the small grids is not costly, and not worth extra optimization.

Such a function can describe exactly hyper planes, is efficient to evaluate and has worked very well for the QM/MM implementation in CP2K .

Appendix D: Construction of the Model Charge Density

The model density $\hat{\rho}(\mathbf{r}, \mathbf{r}_\alpha)$, introduce in Sec. 5.3, can be derived by minimizing the multipole moments and the net charge of the system:

$$\Delta Q_L = \left| \int d\mathbf{r} \mathbf{r}^l Y_L(\mathbf{r}) (\rho(\mathbf{r}, \mathbf{r}_\alpha) - \hat{\rho}(\mathbf{r}, \mathbf{r}_\alpha)) \right| \tag{102}$$

$$\Delta W = \left| \int d\mathbf{r} \mathbf{r}^2 (\rho(\mathbf{r}, \mathbf{r}_\alpha) - \hat{\rho}(\mathbf{r}, \mathbf{r}_\alpha)) \right| \tag{103}$$

The parameters of the model density are obtained from a fit to the original charge density, which is biased by a weight function. In the reciprocal space, both requirements Eq. 102 and Eq. 103 can be translated into expressions that are sensitive only to the intermediate neighborhood of the origin. Thus the fit uses a weighting function of the form:

$$w(\mathbf{k}) = 4\pi \frac{(|\mathbf{k}|^2 - |\mathbf{k}_{cut}|^2)^2}{|\mathbf{k}|^2 |\mathbf{k}_{cut}|^2} \tag{104}$$

for $|\mathbf{k}| < |\mathbf{k}_{cut}|$ and zero elsewhere. The weight function enhances the importance of the low k-vectors while ignoring the high k-vectors of the density.

Using the method of Lagrange multipliers, the parameters of the model density q_α are obtained from the extremal condition of

$$\mathcal{L}(q_\alpha, \lambda) = \frac{V}{2} \sum_{\mathbf{k}\neq 0} w(\mathbf{k}) \left| \rho(\mathbf{k}) - \sum_\alpha q_\alpha g_\alpha(\mathbf{k}) \right|^2 - \lambda V \left[\rho(\mathbf{k}=0) - \sum_\alpha q_\alpha g_\alpha(\mathbf{k}=0) \right] \tag{105}$$

In matrix form the equation can be written in

$$\mathbf{Aq} + \lambda \mathbf{C} = \mathbf{BCq} = N \tag{106}$$

where the matrix element of $\mathbf{A}$, $\mathbf{C}$ and $\mathbf{B}$ are given by:

$$A_{i,j} = V \sum_{\mathbf{k}\neq 0} w(\mathbf{k}) [g_i^\dagger(\mathbf{k}) g_j(\mathbf{k})] \tag{107}$$

$$C_i = V g_i(\mathbf{k}=0) = 1 \tag{108}$$

$$B_i = V \sum_{\mathbf{k}\neq 0} w(\mathbf{k}) Re[\rho^\dagger(\mathbf{k}) g_i(\mathbf{k})] \tag{109}$$

and $\mathbf{q}$ is the array of parameters of the model charge density. The solution to this linear equation system is given by:

$$\mathbf{q} = \mathbf{A}^{-1} \left[\mathbf{B} - \mathbf{C} \frac{\mathbf{CA}^{-1}\mathbf{B} - N}{\mathbf{CA}^{-1}\mathbf{C}} \right] \tag{110}$$

References

1. The CP2K developers group, "Cp2k", freely available at the URL: `http://www.cp2k.org`, released under GPL license., 2012.
2. A. Laio, J. VandeVondele, and U. Röthlisberger, *A Hamiltonian electrostatic coupling scheme for hybrid Car-Parrinello molecular dynamics simulations*, J. Chem. Phys. B, **106**, no. 16, 7300–7307, 2002.
3. G. Karlström, R. Lindh, P.-A. Malmqvist, B. O. Roos, U. Ryde, V. Veryazov, P.-O. Widmark, M. Cossi, B. Schimmelpfennig, P. Neogrady, and L. Seijo, *MOLCAS: a program package for computational chemistry*, Computational Material Science, **28**, no. 2, 222, 2003.
4. M. W. Schmidt, K. K. Baldridge, J. A. Boatz, S. T. Elbert, M. S. Gordon, J. H. Jensen, S. Koseki, N. Matsunaga, K. A. Nguyen, S. Su, T. L. Windus, M. Dupuis, and J. A. Montgomery, *General atomic and molecular electronic-structure system*, J. Comput. Chem., **14**, no. 11, 1347–1363, 1993.
5. CPMD, Version 3.9.1, copyright IBM Corp. 1990-2005, copyright MPI für Festkörperforschung Stuttgart 1997-2005; `http://www.cpmd.org/`.
6. L. Fusti-Molnar and P. Pulay, *Accurate molecular integrals and energies using combined plane wave and Gaussian basis sets in molecular electronic structure theory*, J. Chem. Phys., **116**, no. 18, 7795–7805, 2002.
7. M. Eichinger, P. Tavan, J. Hutter, and M. Parrinello, *A hybrid method for solutes in complex solvents: Density functional theory combined with empirical force fields*, J. Chem. Phys., **110**, no. 21, 10452–10467, 1999.

8. A. Crespo, D. A. Scherlis, M. A. Martí, P. Ordejon, A. E. Roitberg, and D. A. Estrin, *A DFT-based QM-MM approach designed for the treatment of large molecular systems: Application to chorismate mutase*, J. Phys. Chem. B, **107**, no. 49, 13728–13736, 2003.
9. A. Tongraar, K. R. Liedl, and B. M. Rode, *Born-Oppenheimer ab initio QM/MM dynamics simulations of Na+ and K+ in water: From structure making to structure breaking effects*, J. Phys. Chem. A, **102**, no. 50, 10340–10347, 1998.
10. A. Tongraar and B. M. Rode, *A Born-Oppenheimer ab initio quantum mechanical/-molecular mechanical molecular dynamics simulation on preferential solvation of Na+ in aqueous ammonia solution*, J. Phys. Chem. A, **105**, no. 2, 506–510, 2001.
11. C. F. Schwenk, H. H. Loeffler, and B. M. Rode, *Structure and dynamics of metal ions in solution: QM/MM molecular dynamics simulations of Mn2+ and V2+*, J. Am. Chem. Soc., **125**, no. 6, 1618–1624, 2003.
12. S. Chalmet and M. F. Ruiz-Lopez, *The reaction field of a water molecule in liquid water: Comparison of different quantum/classical models*, J. Chem. Phys., **115**, no. 11, 5220–5227, 2001.
13. S. Chalmet, D. Rinaldi, and M. F. Ruiz-Lopez, *A QM/MM/continuum model for computations in solution: Comparison with QM/MM molecular dynamics simulations*, Int. J. Quant. Chem., **84**, no. 5, 559–564, 2001.
14. P. Bandyopadhyay and M. S. Gordon, *A combined discrete/continuum solvation model: Application to glycine*, J. Chem. Phys., **113**, no. 3, 1104–1109, 2000.
15. J. Gao and C. Alhambra, *A hybrid semiempirical quantum mechanical and lattice-sum method for electrostatic interactions in fluid simulations*, J. Chem. Phys., **107**, no. 4, 1212–1217, 1997.
16. K. Nam, J. Gao, and D. M. York, *An efficient linear-scaling Ewald method for long-range electrostatic interactions in combined QM/MM calculations*, J. Chem. Theory Comp., **1**, no. 1, 2–13, 2005.
17. F. Dehez, M. T. C. Martins-Costa, D. Rinaldi, and C. Millot, *Long-range electrostatic interactions in hybrid quantum and molecular mechanical dynamics using a lattice summation approach*, J. Chem. Phys., **122**, no. 23, No. 234503, 2005.
18. D. A. Yarne, M. E. Tuckerman, and G. J. Martyna, *A dual length scale method for plane-wave-based, simulation studies of chemical systems modeled using mixed ab initio/empirical force field descriptions*, J. Chem. Phys., **115**, no. 8, 3531–3539, 2001.
19. Y.Q. Tu and A. Laaksonen, *On the effect of Lennard-Jones parameters on the quantum mechanical and molecular mechanical coupling in a hybrid molecular dynamics simulation of liquid water*, J. Comput. Chem., **111**, no. 16, 7519–7525, 1999.
20. J. Gao, P. Amara, C. Alhambra, and M.J. Field, *A Generalized Hybrid Orbital (GHO) Method for treatment of boundary atoms in QM/MM calculations*, J. Phys. Chem., **102**, no. 24, 4714–4721, 1998.
21. D. Das, K. P. Eurenius, E. M. Billings, P. Sherwood, D. C. Chatfield, M. Hodoscek, and B. R. Brooks, *Optimization of quantum mechanical molecular mechanical partitioning schemes: Gaussian delocalization of molecular mechanical charges and the double link atom method*, J. Chem. Phys., **117**, no. 23, 10534–10547, 2002.
22. P.K. Biswas and V. Gogonea, *A regularized and renormalized electrostatic coupling Hamiltonain for hybrid quantum mechanical - molecular mechanical calculations.*, J. Chem. Phys., **123**, no. 16, No. 164114, 2005.

23. U.C. Singh and P.A. Kollman, *A combined* ab initio *quantum mechanical and molecular mechanical method for carrying out simulations on complex molecular systems: Applications to the $CH_3Cl + Cl^-$ exchange reaction and gas-phase protonation of polyethers*, J. Comput. Chem., **7**, no. 6, 718–730, 1986.
24. M. J. Field, P. A. Bash, and M. Karplus, *A combined quantum-mechanical and molecular mechanical potential for molecular-dynamics simulations*, J. Comput. Chem., **11**, no. 6, 700–733, 1990.
25. D. A. McQuarrie, *Statistical Mechanics*, University Science Books, Sausalito, CA, 2000, p. 234.
26. B. W. H. van Beest, G. J. Kramer, and R. A. van Santen, *Force fields for silicas and aluminophosphates based on ab initio calculations*, Phys. Rev. Lett., **64**, no. 16, 1955–1958, 1990.
27. M. Eichinger, P. Tavan, J. Hutter, and M. Parrinello, *A hybrid method for solutes in complex solvents: Density functional theory combined with empirical forcefields*, J. Chem. Phys., **110**, no. 21, 10452–10467, 1999.
28. P. K. Biswas, *A new ab initio method of calculating Zeff and hence the positron annihilation rates using T-matrix scattering amplitudes*, Eur. Phys. J. D, **29**, no. 1, 321–327, 2004.
29. P. Sherwood, *Modern Methods and Algorithms of Quantum Chemistry*, vol. 1 of *NIC Series*, chapter Hybrid quantum mechanics/molecular mechanics approaches, pp. 257–277, John von Neumann Institute for Computing, 2000.
30. J. VandeVondele, M. Krack, F. Mohamed, M. Parrinello, T. Chassaing, and J. Hutter, *QUICKSTEP: Fast and accurate density functional calculations using a mixed Gaussian and plane waves approach*, Comp. Phys. Comm., **167**, no. 2, 103–128, 2005.
31. B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J. States, and S. Swaminathan, *CHARMM - A program for macromolecular energy, minimization and dynamics calculations*, J. Comput. Chem., **4**, no. 2, 187–217, 1983.
32. D. Case, D. Pearlman, J. Caldwell, T.E. Cheatham III, W. Ross, C. Simmerling, T. Darden, K. Merz, R. Stanton, A. Cheng, J. Vincent, M. Crowley, V. Tsui, R. Radmer, Y. Duan, J. Pitera, I. Massova, G. Seibel, U. Singh, P. Weiner, and P.A. Kollman, *AMBER v.9.0*, Tech. Rep., University of California, San Francisco, 2002.
33. T. Laino, The Mathematica Notebook used to develop the GEEP technology is part of the CP2K distribution and can be freely downloaded. Released under GPL license.
34. W. Hackbusch, *Multi-Grid Methods and Applications*, vol. 4 of *Series in Computational Mathematics*, Springer Verlag, Berlin, 1985.
35. W.L. Briggs, *A Multigrid Tutorial*, SIAM Books, Philadelphia, 1987.
36. G. Feng, *Data smoothing by cubic spline filters*, IEEE Trans.on Signal Process., **46**, no. 10, 2790–2796, 1998.
37. P. Pulay, *Ab initio calculation of force constants and equilibrium geometries in polyatomic molecules .I. Theory*, Mol. Phys., **17**, no. 2, 197–204, 1969.
38. H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, and J. Hermans, *Intermolecular Forces*, chapter Interaction models for water in relation to protein hydration, pp. 331–342, Reidel, Dordrecht, The Netherlands, 1981.
39. M. Deserno and C. Holm, *How to mesh up Ewald sums. I. A theoretical and numerical comparison of various particle mesh routines*, J. Chem. Phys., **109**, no. 18, 7678–7693, 1998.

40. T. Darden, D. York, and L. Pedersen, *Particle mesh ewald - an N log(N) method for ewald sums in large systems*, J. Chem. Phys., **98**, no. 12, 10089–10092, 1993.
41. T. Laino, F. Mohamed, A. Laio, and M. Parrinello, *An efficient real space multigrid QM/MM electrostatic coupling*, J. Chem. Theory Comp., **1**, no. 6, 1176–1184, 2005.
42. T. Laino, F. Mohamed, A. Laio, and M. Parrinello, *An Efficient Linear-Scaling Electrostatic Coupling for treating periodic boundary conditions in QM/MM Simulations*, J. Chem. Theory Comp., **2**, no. 5, 1370–1378, 2005.
43. P. P. Ewald, *The calculation of optical and electrostatic grid potential*, Ann. Phys., **64**, no. 3, 253–287, 1921.
44. M. Allen and D. Tildesley, *Computer Simulation of Liquids*, Oxford University Press, Oxford, 1987.
45. G. Lippert, J.E. Hutter, and M. Parrinello, *The Gaussian and augmented-plane-wave density functional method for ab initio molecular dynamics simulations*, Theor. Chem. Acc., **103**, no. 2, 124–140, 1999.
46. G. J. Martyna and M. E. Tuckerman, *A reciprocal space based method for treating long range interactions in ab initio and force-field-based calculations in clusters*, J. Chem. Phys, **110**, no. 6, 2810–2821, 1999.
47. P. E. Blöchl, *Density derived atomic point charges*, J. Chem. Phys., **103**, no. 17, 7422–7428, 1995.
48. L. Levien, C. T. Prewitt, and D. J. Weidner, *Structure and elastic properties of quartz at pressure*, Am. Mineral., **65**, no. 9-10, 920–930, 1980.
49. C. M. Nelson and R. A. Weeks, *Trapped electrons in irradiated quartz and silica. 1. Optical Absorption*, J. Am. Ceram. Soc., **43**, no. 8, 396–399, 1960.
50. R. A. Weeks, *Paramagnetic resonance of lattice defects in irradiated quartz*, J. Appl. Phys., **27**, no. 11, 1376–1381, 1956.
51. R. A. Weeks and C. M. Nelson, *Trapped electrons in irradiated quartz and silica .2. Electron spin resonance*, J. Am. Ceram. Soc., **43**, no. 8, 399–404, 1960.
52. R. H. Silsbee, *Electron spin resonance in neutron-irradiated quartz*, J. Appl. Phys., **32**, no. 8, 1459–1461, 1961.
53. M. G. Jani, R. B. Bossoli, and L. E. Halliburton, *Further characterization of the E_1' center in crystalline SiO2*, Phys. Rev. B, **27**, no. 4, 2285–2293, 1983.
54. W. L. Warren, E. H. Poindexter, M. Offenberg, and W. Müller-Warmuth, *Paramagnetic point-defects in amorphous-silicon dioxide and amorphous-silicon nitride thin-films .1. a-SiO2*, J. Electrochem. Soc., **139**, no. 3, 872–880, 1992.
55. E. H. Poindexter and W. L. Warren, *Paramagnetic point-defects in amorphous thin-films of SiO2 and Si3N4 - Updates and additions*, J. Electrochem. Soc., **142**, no. 7, 2508–2516, 1995.
56. D. R. Hamann, *Generalized Gradient Theory for Silica Phase Transitions*, Phys. Rev. Lett., **76**, no. 4, 660–663, 1996.
57. P. E. Blöchl, *First-Principles calculations of defects in oxygen-deficient silica exposed to hydrogen*, Phys. Rev. B, **62**, no. 10, 6158–6179, 2000.
58. K. C. Snyder and W. B. Fowler, *Oxygen vacancy in alpha-quartz - A possible bistable and metastable defect*, Phys. Rev. B, **48**, no. 18, 13238–13243, 1993.
59. D. C. Allan and M. P. Teter, *Local density approximation total energy calculations for silica and titania structure and defects*, J. Am. Ceram. Soc., **73**, no. 11, 3247–3250, 1990.

60. M. Boero, A. Pasquarello, J. Sarnthein, and R. Car, *Structure and hyperfine parameters of E'(1) centers in a-quartz and in vitreous SiO2*, Phys. Rev. Lett., **78**, no. 5, 887–890, 1997.
61. G. Pacchioni and G. Ierano, *Optical absorption and nonradiative decay mechanism of E'_1 center in silica*, Phys. Rev. Lett., **81**, no. 2, 377–380, 1998.
62. A. H. Edwards and W. B. Fowler, *Semi-empirical molecular-orbital techniques applied to silicon dioxide - MINDO/3*, J. Phys. Chem. Solids, **46**, no. 7, 841–857, 1985.
63. J. K. Rudra and W. B. Fowler, *Oxygen vacancy and the E'_1 center in crystalline SiO2*, Phys. Rev. B, **35**, no. 15, 8223–8230, 1987.
64. G. Pacchioni, A. M. Ferrari, and G. Ierano, *Cluster model calculations of oxygen vacancies in SiO2 and MgO - Formation energies, optical transitions and EPR spectra*, Faraday Discuss., **106**, no. 1, 155–172, 1997.
65. J. Sauer and M. Sierka, *Combining quantum mechanics and interatomic potential functions in ab initio studies of extended systems*, J. Comput. Chem., **21**, no. 16, 1470–1493, 2000.
66. V.B. Sulimov, P.V. Sushko, A.L. Shluger A.H. Edwards, and A.M. Stoneham, *Asymmetry and long-range character of lattice deformation by neutral oxygen vacancy in alpha-quartz*, Phys. Rev. B, **66**, no. 2, 24108–24114, 2002.
67. A.S. Mysovsky, P.V. Sushko, A.H. Edwards S. Mukhopadhyay, and A.L. Shluger, *Calibration of embedded-cluster method for defect studies in amorphous silica*, Phys. Rev. B, **69**, no. 8, No.085202, 2004.
68. V.B. Sulimov, S. Casassa, C. Pisani, J. Garapon, and B. Poumellac, *Embedded cluster ab initio study of the neutral oxygen vacancy in quartz and cristobalite*, Modell. Simul. Mater. Sci. Eng., **8**, no. 5, 763–773, 2000.
69. C. Pisani, M. Busso, F. Lopez-Gejo, S. Casassa, and L. Maschio, *Quasi-periodic ab initio models in material science: the case of oxygen-deficient centers in optical fibers*, Theor. Chem. Acc., **111**, no. 2-6, 246–254, 2004.
70. D. Erbetta, D. Ricci, and G. Pacchioni, *Simplified embedding schemes for the quantum-chemical description of neutral and charged point defects in SiO2 and related dielectrics*, J. Chem. Phys., **113**, no. 23, 10744–10752, 2000.
71. V.A. Nasluzov, E.A. Ivanova, A.M. Shor, G.N. Vayssilov, U. Birkenheuer, and N. Rosch, *Elastic polarizable environment cluster embedding approach for covalent oxides and zeolites based on a density functional method*, J. Phys. Chem. B, **107**, no. 10, 2228–2241, 2003.
72. F. Zipoli, T. Laino, A. Laio, M. Bernasconi, and M. Parrinello, *A* QUICKSTEP *-based quantum mechanics/molecular mechanics approach for silica*, J. Chem. Phys., **124**, no. 15, No. 154707, 2006.
73. J.S. Tse, D.D. Klug, and Y.L. Page, *High-pressure densification of amorphous silica*, Phys. Rev. B, **46**, no. 10, 5933–5938, 1992.
74. S. Goedecker, M. Teter, and J. Hutter, *Separable dual-space Gaussian pseudopotentials*, Phys. Rev. B, **54**, no. 3, 1703–1710, 1996.
75. H. R. Philipp, *Optical transitions in crystalline and fused quartz*, Solid State Comm., **4**, no. 1, 73–75, 1966.
76. S. Miyazaki, H. Nishimura, M. Fukuda, L. Ley, and J. Ristein, *Structure and electronic states of ultrathin SiO2 thermally grown on Si(100) and Si(111) surfaces*, Appl. Surf. Sci., **114**, 585–589, 1997.

77. G. Pacchioni, L. Skuja, and D. L. Griscom, (Eds.), *Defects in SiO_2 and Related Dielectrics: Science and Technology*, Kluwer Academic Publisher, 2000.
78. J. Sarnthein, A. Pasquarello, and R. Car, *Origin of the high-frequency doublet in the vibrational spectrum of vitreous SiO2*, Science, **275**, no. 5308, 1925–1927, 1997.
79. N. Binggeli and J.R. Chelikowsky, *Structural transformation of quartz at high-pressures*, Nature, **353**, no. 6342, 344–346, 1991.
80. L.S. Dubrovinsky, S.K. Saxena, P. Lazor, R. Ahuja, O. Eriksson, J.M. Wills, and B. Johansson, *Experimental and theoretical identification of a new high-pressure phase of silica*, Nature, **388**, no. 6640, 362–365, 1997.
81. M.-H. Du and H.-P. Cheng, *Transparent interface between classical molecular dynamics and first-principles molecular dynamics*, Int. J. Quant. Chem., **93**, no. 1, 1–8, 2003.
82. M.-H. Du, A. Kolchin, and H.-P. Cheng, *Water-silica surface interactions: A combined quantum-classical molecular dynamic study of energetics and reaction pathways*, J. Chem. Phys., **119**, no. 13, 6418–6422, 2003.
83. I. Laika-Voivod, F. Sciortino, T. Grande, and P.H. Poole, *Phase diagram of silica from computer simulation*, Phys. Rev. E, **70**, no. 6, No. 61507, 2004.
84. C.S. Carmer, B. Weiner, and M. Frenklach, *Molecular-dynamics with combined quantum and empirical potentials - C2H2 adsorption on Si(100)*, J. Chem. Phys., **99**, no. 2, 1356–1372, 1993.
85. J.R. Shoemaker, L.W. Burggraf, and M.S. Gordon, *SIMOMM: An integrated molecular orbital/molecular mechanics optimization scheme for surfaces*, J. Phys. Chem. A, **103**, no. 17, 3245–3251, 1999.
86. M. Frenklach, M. Skokov, and B. Weiner, *An atomistic model for stepped diamond growth*, Nature, **372**, no. 6506, 535–537, 1994.
87. G. Monard, M. Loos, V. Thery, K. Baka, and J.-L. Rivail, *Hybrid classical quantum force field for modeling very large molecules*, Int. J. Quant. Chem., **58**, no. 2, 153–159, 1996.
88. Y.K. Zhang, T.S. Lee, and W.T. Yang, *A pseudobond approach to combining quantum mechanical and molecular mechanical methods*, J. Chem. Phys., **110**, no. 1, 46–54, 1999.
89. V. Kairys and J.H. Jensen, *QM/MM Boundaries Across Covalent Bonds: A Frozen Localized Molecular Orbital-Based Approach for the Effective Fragment Potential Method*, J. Phys. Chem. A, **104**, no. 28, 6656–6665, 2000.
90. R. Murphy, D. Philipp, and R. Friesner, *A mixed quantum mechanics/molecular mechanics (QM/MM) method for large-scale modeling of chemistry in protein environments*, J. Comput. Chem., **21**, no. 16, 1442–1447, 2000.
91. T. Narasimhamurty, *Photoelastic constants of alpha-quartz*, J. Opt. Soc. Am., **59**, no. 6, 682–686, 1969.
92. C.M. Carbonaro, V. Fiorentini, and S. Massidda, *Ab initio study of oxygen vacancies in alpha-quartz*, J. Non-Cryst. Solids, **221**, no. 1, 89–96, 1997.
93. N. Capron, S. Carniato, A. Lagraa, and G. Boureau, *Local density approximation and generalized gradient approximation calculations for oxygen and silicon vacancies in silica*, J. Chem. Phys., **112**, no. 21, 9543–9548, 2000.
94. B.B. Stefanov and K. Raghavachari, *Photoabsorption of the neutral oxygen vacancy in silicate and germanosilicate glasses: First-principles calculations*, Phys. Rev. B, **56**, no. 9, 5035–5038, 1997.

95. G. Boureau and S. Carniato, *Apparent discrepancies between thermodynamic data and theoretical calculations of the formation energy of an oxygen vacancy in silica*, Solid State Commun., **98**, no. 6, 485–487, 1996, Correction in **Solid State Commun.** 99(1), R1-R4 (1996).
96. D.R. Lide, (Ed.), *CRC Handbook of Chemistry and Physics*, CRC Press, 78th edition edition, 1997-1998.
97. Frank Jensen, *Introduction to Computational Chemistry,*, John Wiley & Sons, 1999.
98. S. Baroni, S. de Gironcoli, A. Dal Corso, and P. Giannozzi, *Phonons and related crystal properties from density-functional perturbation theory*, Rev. Mod. Phys., **73**, no. 2, 515, 2001.
99. T. Laino, "Cpmd comparison", For the CPMD calculations we have used norm-conserving pseudopotentials and plane wave expansion of the Kohn-Sham orbitals up to a kinetic cutoff of 90 Ry. The disilane molecule is at the center of a cubic box of edge 15 Å.
100. G. Roma, Y. Limoge, and S. Baroni, *Oxygen self-diffusion in alpha-quartz*, Phys. Rev. Lett., **86**, no. 20, 4564–4567, 2001.
101. R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman Lectures on Physics*, vol. 1, Addison-Wesley, xs, 1963, 30-11.

Accelerated Molecular Dynamics Methods

Danny Perez[1], Blas. P. Uberuaga[2], and Arthur F. Voter[1]

[1] Theoretical Division
Los Alamos National Laboratory, Los Alamos, NM, 87545, USA
E-mail: {danny_perez, afv}@lanl.gov

[2] Materials Science and Technology Division
Los Alamos National Laboratory, Los Alamos, NM, 87545, USA
E-mail: blas@lanl.gov

A long-standing limitation in the use of molecular dynamics (MD) simulation is that it can only be applied directly to processes that take place on very short timescales: nanoseconds if empirical potentials are employed, or picoseconds if we rely on electronic structure methods. Many processes of interest in chemistry, biochemistry, and materials science require study over microseconds and beyond, due either to the natural timescale for the evolution or to the duration of the experiment of interest. Ignoring the case of liquids, the dynamics on these time scales is typically characterized by infrequent-event transitions, from state to state, usually involving an energy barrier. There is a long and venerable tradition of using transition state theory (TST)[1–3] to directly compute rate constants for these kinds of activated processes. If needed, dynamical corrections to the TST rate, and even quantum corrections, can be computed to achieve an accuracy suitable for the problem at hand. These rate constants then allow us to understand the system behavior on longer time scales than we can directly reach with MD. For complex systems with many reaction paths, the TST rates can be fed into a stochastic simulation procedure such as kinetic Monte Carlo, and a direct simulation of the advance of the system through its possible states can be obtained in a probabilistically exact way.

A problem that has become more evident in recent years, however, is that for many systems of interest, there is a complexity that makes it difficult, if not impossible, to determine all the relevant reaction paths to which TST should be applied. This is a serious issue, as omitted transition pathways can have uncontrollable consequences on the simulated long-time kinetics.

Over the last 15 years or so, we have been developing a new class of methods for treating the long-time dynamics in these complex, infrequent-event systems. Rather than trying to guess in advance what reaction pathways may be important, we return instead to a molecular dynamics treatment, in which the trajectory itself finds an appropriate way to escape from each state of the system. Since a direct integration of the trajectory would be limited to nanoseconds, while we are seeking to follow the system for much longer times, we modify the dynamics in some way to cause the first escape to happen much more quickly, thereby accelerating the dynamics. The key is to design the modified dynamics in a way that does as little damage as possible to the probability for escaping along a given pathway – i.e., we try to preserve the relative rate constants for the different possible escape paths out of the state. We can then use this modified dynamics to follow the system from state to state, reaching much longer times than we could reach with direct MD. The dy-

namics within any one state may no longer be meaningful, but the state-to-state dynamics, in the best case (as we discuss below), can be exact. We have developed three methods in this "accelerated molecular dynamics" (AMD) class, in each case appealing to TST, either implicitly or explicitly, to design the modified dynamics. Each of the methods has its own advantages, and we and others have applied these methods to a wide range of problems. The purpose of this article is to give the reader a brief introduction to how these methods work. Note that this brief review does not claim to be exhaustive: various other methods aiming at similar goals have been proposed in the literature. For the sake of brevity, our focus will exclusively be on the methods developed by our group.

These lecture notes are organized as follow: Section 1 introduces the basic concepts (infrequent event systems, TST, etc.) that are later called upon. Sections 2,3, and 4 introduce the AMD methods – Parallel-Replica Dynamics, Hyperdynamics and Temperature Accelerated Dynamics – and provide examples of their use. Finally, Section 5 provides guidelines to assist in the choice of the right AMD method for a given problem.

1 Background

1.1 Infrequent Event Systems

We begin by defining an "infrequent-event" system, as this is the type of system for which the accelerated dynamics methods are ideal. The dynamical evolution of such a system is characterized by the occasional activated event that takes the system from basin to basin, events that are separated by possibly millions of thermal vibrations within one basin. A simple example of an infrequent-event system is an adatom on a metal surface at a temperature that is low relative to the diffusive jump barrier. We will exclusively consider thermal systems, characterized by a temperature T, a fixed number of atoms N, and a fixed volume V; i.e., the canonical ensemble. Typically, there is a large number of possible paths for escape from any given basin. As a trajectory in the $3N$-dimensional coordinate space in which the system resides passes from one basin to another, it crosses a ($3N$-1)-dimensional "dividing surface" at the ridgetop separating the two basins. While on average these crossings are infrequent, successive crossings can sometimes occur within just a few vibrational periods; these are termed "correlated dynamical events" (e.g., see Ref. 4–6). An example would be a double jump of the adatom on the surface. For this discussion it is sufficient, but important, to realize that such events can occur. In most of the methods presented below, we will assume that these correlated events do not occur – this is the primary assumption of transition state theory – which is actually a very good approximation for many solid-state diffusive processes. We define the "correlation time" (τ_{corr}) of the system as the duration of the system memory. A trajectory that has resided in a particular basin for longer than τ_{corr} is assumed to have no memory of its history and, consequently, how it got to that basin, in the sense that when it later escapes from the basin, the probability for escape along a given path is independent of how it entered the state. The relative probability for escape to a given adjacent state is proportional to the rate constant for that escape path, which we will define below.

An infrequent event system, then, is one in which the residence time in a state (τ_{rxn}) is much longer than the correlation time (τ_{corr}). We will focus here on systems with energetic barriers to escape, but the infrequent-event concept applies equally well to entropic

bottlenecks.[a] The key to the accelerated dynamics methods described here is recognizing that to obtain the right sequence of state-to-state transitions, we need not evolve the vibrational dynamics perfectly, as long as the relative probability of finding each of the possible escape paths is preserved.

1.2 Transition State Theory

Transition state theory (TST)[8,9,1–3] is the formalism underpinning all of the accelerated dynamics methods, directly or indirectly. In the TST approximation, the classical rate constant for escape from state A to some adjacent state B is taken to be the equilibrium flux through the dividing surface between A and B (Fig. 1). If there are no correlated dynamical events, the TST rate is the exact rate constant for the system to move from state A to state B.

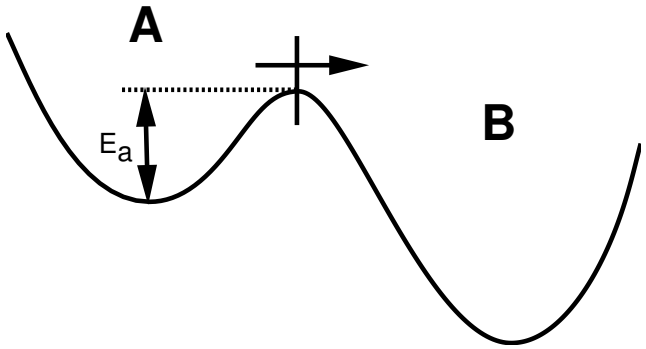


Figure 1. A two-state system illustrating the definition of the transition state theory rate constant as the outgoing flux through the dividing surface bounding state A.

The power of TST comes from the fact that this flux is an *equilibrium* property of the system. Thus, we can compute the TST rate without ever propagating a trajectory. The appropriate ensemble average for the rate constant for escape from A, $k_{A\rightarrow}^{\mathrm{TST}}$, is

$$k_{A\rightarrow}^{\mathrm{TST}} = \langle |dx/dt|\, \delta(\mathrm{x}-\mathrm{q}) \rangle_A \,, \tag{1}$$

where $x \in \mathbf{r}$ is the reaction coordinate and $x = q$ the dividing surface bounding state A. The angular brackets indicate the ratio of Boltzmann-weighted integrals over $6N$-dimensional phase space (configuration space $\mathbf{r}$ and momentum space $\mathbf{p}$). That is, for some property $P(\mathbf{r},\mathbf{p})$,

$$\langle P \rangle = \frac{\int\int P(\mathbf{r},\mathbf{p}) \exp[-H(\mathbf{r},\mathbf{p})/k_B T] d\mathbf{r} d\mathbf{p}}{\int\int \exp[-H(\mathbf{r},\mathbf{p})/k_B T] d\mathbf{r} d\mathbf{p}}, \tag{2}$$

where k_B is the Boltzmann constant and $H(\mathbf{r},\mathbf{p})$ is the total energy of the system, kinetic plus potential. The subscript A in Eq. 1 indicates the configuration space integrals are

[a] For systems with entropic bottlenecks, the parallel-replica dynamics method can be applied very effectively[7].

restricted to the space belonging to state A. If the effective mass (m) of the reaction coordinate is constant over the dividing surface, Eq. 1 reduces to a simpler ensemble average over configuration space only[10],

$$k_{A\rightarrow}^{\mathrm{TST}} = \sqrt{2k_BT/\pi m}\ \langle \delta(x-q) \rangle_A \,. \tag{3}$$

The essence of this expression, and of TST, is that the Dirac delta function picks out the probability of the system being at the dividing surface, relative to everywhere else it can be in state A. Note that there is no dependence on the nature of the final state B.

In a system with correlated events, not every dividing surface crossing corresponds to a reactive event, so that, in general, the TST rate is an upper bound on the exact rate. For diffusive events in materials at moderate temperatures, these correlated dynamical events typically do not cause a large change in the rate constants, so TST is often an excellent approximation. This is a key point; this behavior is markedly different than in some chemical systems, such as molecular reactions in solution or the gas phase, where TST is just a starting point and dynamical corrections can lower the rate significantly (e.g., Ref. 11).

While in the traditional use of TST, rate constants are computed after the dividing surface is specified, in the accelerated dynamics methods we exploit the TST formalism to design approaches that do not require knowing in advance where the dividing surfaces will be, or even what product states might exist.

1.3 Harmonic Transition State Theory

If we have identified a saddle point on the potential energy surface for the reaction pathway between A and B, we can use a further approximation to TST. We assume that the potential energy near the basin minimum is well described, out to displacements sampled thermally, with a second-order energy expansion – i.e., that the vibrational modes are harmonic – and that the same is true for the modes perpendicular to the reaction coordinate at the saddle point. Under these conditions, the TST rate constant becomes simply

$$k_{A\rightarrow B}^{HTST} = \nu_0 e^{-E_a/k_BT}, \tag{4}$$

where

$$\nu_0 = \frac{\prod_i^{3N} \nu_i^{min}}{\prod_i^{3N-1} \nu_i^{sad}}. \tag{5}$$

Here E_a is the static barrier height, or activation energy [the difference in energy between the saddle point and the minimum of state A (c.f., Fig. 1)], $\{\nu_i^{min}\}$ are the strictly positive normal mode frequencies at the minimum of A, and $\{\nu_i^{sad}\}$ are the strictly positive, non-imaginary, normal mode frequencies at the saddle separating A from B. This is often referred to as the Vineyard[12] equation. The analytic integration of Eq. 1 over the whole phase space thus leaves a very simple Arrhenius temperature dependence.[b] To the extent that there are no recrossings and the modes are truly harmonic, this is an exact expression for the rate. This harmonic TST expression is employed in the temperature accelerated dynamics method (without requiring calculation of the prefactor ν_0).

[b] Note that although the exponent in Eq. 4 depends only on the static barrier height E_a, in this HTST approximation there is no assumption that the trajectory passes exactly through the saddle point.

1.4 Complex Infrequent Event Systems

The motivation for developing accelerated molecular dynamics methods becomes particularly clear when we try to understand the dynamical evolution of what we will term complex infrequent event systems. In these systems, we simply cannot guess where the state-to-state evolution might lead. The underlying mechanisms may be too numerous, too complicated, and/or have an interplay whose consequences cannot be predicted by considering them individually. In very simple systems we can raise the temperature to make diffusive transitions occur on an MD-accessible time scale. However, as systems become more complex, changing the temperature causes corresponding changes in the relative probability of competing mechanisms. Thus, this strategy will cause the system to select a different sequence of state-to-state dynamics, ultimately leading to a completely different evolution of the system, and making it impossible to address the questions that the simulation was attempting to answer.

Many, if not most, materials problems are characterized by such complex infrequent events. We may want to know what happens on the time scale of milliseconds, seconds or longer, while with MD we can barely reach one microsecond. Running at higher T or trying to guess what the underlying atomic processes are can mislead us about how the system really behaves. Often for these systems, if we could get a glimpse of what happens at these longer times, even if we could only afford to run a single trajectory for that long, our understanding of the system would improve substantially. This, in essence, is the originally motivation for the development of the methods described here. Coupled with the constant increase in computing power, AMD methods have demonstrated the capability to go beyond this initial mission and can now be used to parameterize higher-scales model or to compute long time averages of relevant quantities.

1.5 Dividing Surfaces and Transition Detection

We have implied that the ridge tops between basins are the appropriate dividing surfaces in these systems. For a system that obeys TST, these ridgetops are the optimal dividing surfaces; recrossings will occur for any other choice of dividing surface. A ridgetop can be defined in terms of steepest-descent paths – it is the $3N$-1-dimensional boundary surface that separates those points connected by steepest descent paths to the minimum of one basin from those that are connected to the minimum of an adjacent basin. This definition also leads to a simple way to detect transitions as a simulation proceeds, a requirement of parallel replica dynamics and temperature accelerated dynamics. Intermittently, the trajectory is interrupted and minimized via steepest descent. If this minimization leads to a basin minimum that is distinguishable from the minimum of the previous basin, a transition has occurred. An appealing feature of this approach is that it requires virtually no knowledge of the type of transition that might occur. Often only a few steepest descent steps are required to determine that no transition has occurred. While this is a fairly robust detection algorithm, more efficient approaches can be tailored to the system being studied, for example, defining transitions as changes in atomic coordination.

In what follows, we describe the accelerated dynamics methods. There are currently three accelerated dynamics that have been developed: parallel replica dynamics, hyperdynamics, and temperature accelerated dynamics.

2 Parallel-Replica Dynamics

The parallel replica method[13] is the simplest and most accurate of the accelerated dynamics techniques, with the only assumption being that the infrequent events obey first-order kinetics (exponential decay); i.e., for any time $t > \tau_{corr}$ after entering a state, the probability distribution function for the time of the next escape is given by

$$p(t) = k_{tot} e^{-k_{tot} t} \tag{6}$$

where k_{tot} is the rate constant for escape from the state. For example, Eq. 6 arises naturally for ergodic, chaotic exploration of an energy basin. Parallel replica allows for the parallelization of the state-to-state dynamics of such a system on M processors. We sketch the derivation here for equal-speed processors. For a state in which the rate to escape is k_{tot}, on M processors the effective escape rate will be Mk_{tot}, as the state is being explored M times faster. Also, if the time accumulated on one processor is t_1, on the M processors a total time of $t_{sum} = Mt_1$ will be accumulated. Thus, we find that

$$p(t_1)dt_1 = Mk_{tot} e^{-Mk_{tot} t_1} dt_1 \tag{7}$$

$$= k_{tot} e^{-k_{tot} t_{sum}} dt_{sum} \tag{8}$$

$$= p(t_{sum})dt_{sum} \tag{9}$$

and the probability to leave the state per unit time, expressed in t_{sum} units, is the same whether it is run on one or M processors. A variation on this derivation shows that the M processors need not run at the same speed, allowing the method to be used on a heterogeneous or distributed computer; see Ref. 13.

The algorithm is schematically shown in Fig. 2. Starting with an N-atom system in a particular state (basin), the entire system is replicated on each of M available parallel or distributed processors. After a short dephasing stage during which each replica is evolved forward with independent noise for a time $\Delta t_{deph} \geq \tau_{corr}$ to eliminate correlations between replicas, each processor carries out an independent constant-temperature MD trajectory for the entire N-atom system, thus exploring phase space within the particular basin M times faster than a single trajectory would. Whenever a transition is detected

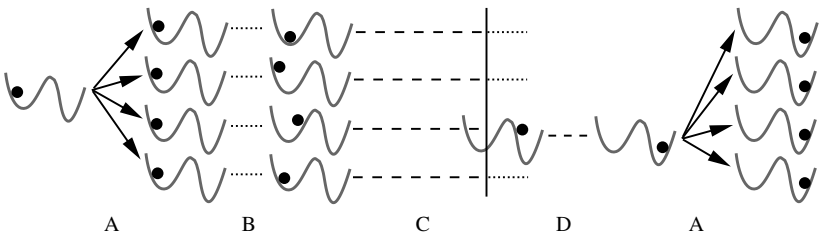


Figure 2. Schematic illustration of the parallel replica method (after Ref. 7). The four steps, described in the text, are (A) replication of the system into M copies, (B) dephasing of the replicas, (C) independent trajectories until a transition is detected in any of the replicas, and (D) brief continuation of the transitioning trajectory to allow for correlated events such as recrossings or follow-on transitions to other states. The resulting configuration is then replicated, beginning the process again.

on any processor, all processors are alerted to stop. The simulation clock is advanced by the accumulated trajectory time summed over all replicas, i.e., the total time τ_{rxn} spent exploring phase space within the basin until the transition occurred.

The parallel replica method also correctly accounts for correlated dynamical events (i.e., there is no requirement that the system obeys TST), unlike the other accelerated dynamics methods. This is accomplished by allowing the trajectory that made the transition to continue on its processor for a further amount of time $\Delta t_{corr} \geq \tau_{corr}$, during which recrossings or follow-on events may occur. The simulation clock is then advanced by Δt_{corr}, the final state is replicated on all processors, and the whole process is repeated. Parallel replica dynamics then gives an arbitrarily accurate state-to-state dynamical evolution, because the escape times obey the correct probability distribution, nothing about the procedure corrupts the relative probabilities of the possible escape paths, and the correlated dynamical events are properly accounted for.

The efficiency of the method is limited by both the dephasing stage, which does not advance the system clock, and the correlated event stage, during which only one processor accumulates time. (This is illustrated schematically in Fig. 2, where dashed line trajectories advance the simulation clock but dotted line trajectories do not.) Thus, the overall efficiency will be high when

$$\tau_{rxn}/M \gg \Delta t_{deph} + \Delta t_{corr}. \tag{10}$$

Some tricks can further reduce this requirement. For example, whenever the system revisits a state, on all but one processor the interrupted trajectory from the previous visit can be immediately restarted, eliminating the dephasing stage. Also, the correlation stage (which only involves one processor) can be overlapped with the subsequent dephasing stage for the new state on the other processors, in the hope that there are no correlated crossings that lead to a different state.

While the derivation of the parallel-replica presented above does not impose a specific definition of a "state" of the system, the operational definition used in practice often corresponds to a single basin of the potential energy surface (c.f. Section 1.5). An exponential distribution of escape times is then obtained if the typical timescale for a transition out of the state is long compared to the characteristic vibrational period of the system around that fixed point, i.e., if there is a separation of timescale between vibrations and transitions between basins. While this definition has the virtue of being conceptually and computationally simple, it limits the range of possible applications to systems where the basins are deep enough (relative to $k_B T$) and well separated from each other and leaves many other, more complex, systems out of reach. There is thus a clear need to develop strategies to capitalize on more general definitions of states and hence higher-level gaps in the characteristic timescales spectrum. For example, in the case of pyrolysis of hexadecane, it was shown that a state could be defined as the ensemble of all configuration space points that share the same network of covalent bonds[14]. In that case, these "superstates" contain a large number of simple energy basins of the potential energy surface, each corresponding to different global conformations of the molecular backbone. There, the method exploited the separation of timescale between the rapid changes of dihedral angles of the backbone (intrasuperstate transitions) and the slow covalent bond breaking process (intersuperstates transitions) rather than between the vibrational timescale and that of sampling of the different dihedral angles. This enables one to ignore the "irrelevant" fast transitions that would

demand incessant dephasing and decorrelation and concentrate directly on the real kinetic bottlenecks. This has allowed for simulations over timescales of microseconds and to the observation of various non-trivial reactions, such as the isomerization process shown in Fig. 3.

Note that parallel replica dynamics can also be extended to more general classes of problems, such as systems with some externally applied strain rate. The requirement here is that the drive rate is slow enough that at any given time the rates for the processes in the system depend only on the instantaneous configuration of the system[15]. Note that in this case, different processors must run at the same speed (or synchronization must be enforce by some other mean).

Parallel replica dynamics has the advantage of being fairly simple to program, with very few "knobs" to adjust – Δt_{deph} and Δt_{corr}, which can be conservatively set at a few ps for most systems. As multiprocessing environments are now ubiquitous, parallel replica dynamics provides a very powerful simulation tool.

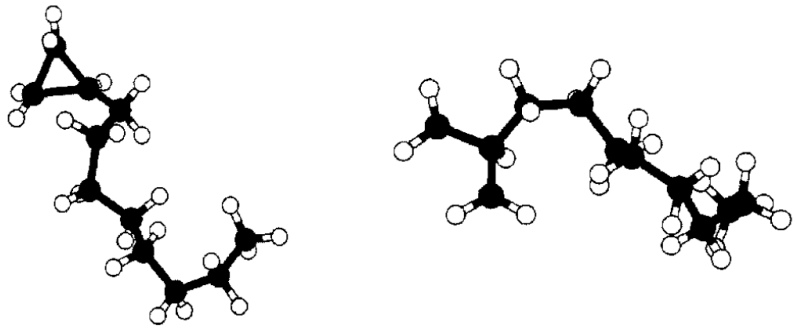


Figure 3. Isomerization of $C_{11}H_{22}$ from a cyclopropyl structure (left) to a branched diradical (right) as obtained through parallel replica dynamics simulations at 2500K. Taken from Ref. 14.

3 Hyperdynamics

Hyperdynamics[7,16] builds on the basic concept of importance sampling[17,18], extending it into the time domain. In the hyperdynamics approach[16], the potential surface $V(\mathbf{r})$ of the system is modified by adding to it a nonnegative *bias* potential $\Delta V_b(\mathbf{r})$. The dynamics of the system is then evolved on this biased potential surface, $V(\mathbf{r}) + \Delta V_b(\mathbf{r})$. A schematic illustration is shown in Fig. 4. The derivation of the method requires that the system obeys TST – that there are no correlated events. There are also important requirements on the form of the bias potential. It must be zero at all the dividing surfaces, and the system must still obey TST for dynamics on the modified potential surface. If such a bias potential can be constructed, a challenging task in itself, we can substitute the modified potential $V(\mathbf{r}) + \Delta V_b(\mathbf{r})$ into Eq. 1 to find

$$k_{A\rightarrow}^{\mathrm{TST}} = \frac{\langle |v_A| \, \delta(\mathrm{r}) \rangle_{A_b}}{\left\langle e^{\beta \Delta V_b(\mathrm{r})} \right\rangle_{A_b}}, \tag{11}$$

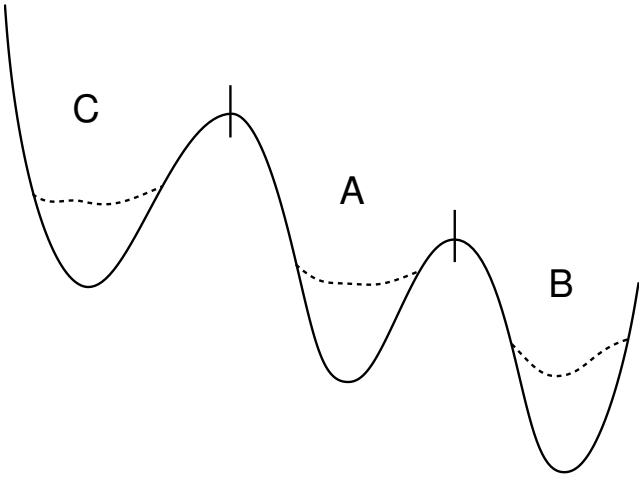


Figure 4. Schematic illustration of the hyperdynamics method. A bias potential ($\Delta V(\mathbf{r})$), is added to the original potential ($V(\mathbf{r})$, solid line). Provided that $\Delta V(\mathbf{r})$ meets certain conditions, primarily that it be zero at the dividing surfaces between states, a trajectory on the biased potential surface ($V(\mathbf{r}) + \Delta V(\mathbf{r})$, dashed line) escapes more rapidly from each state without corrupting the relative escape probabilities. The accelerated time is estimated as the simulation proceeds.

where $\beta = 1/k_B T$ and the state A_b is the same as state A but with the bias potential ΔV_b applied. This leads to a very appealing result: a trajectory on this modified surface, while relatively meaningless on vibrational time scales, evolves *correctly* from state to state at an accelerated pace. That is, the relative rates of events leaving A are preserved:

$$\frac{k^{\mathrm{TST}}_{A_b \to B}}{k^{\mathrm{TST}}_{A_b \to C}} = \frac{k^{\mathrm{TST}}_{A \to B}}{k^{\mathrm{TST}}_{A \to C}}. \tag{12}$$

This is because these relative probabilities depend only on the numerator of Eq. 11 which is unchanged by the introduction of ΔV_b since, by construction, $\Delta V_b = 0$ at the dividing surface.

Moreover, the accelerated time is easily estimated as the simulation proceeds. For a regular MD trajectory, the time advances at each integration step by Δt_{MD}, the MD time step (often on the order of 1 fs). In hyperdynamics, the time advance at each step is Δt_{MD} multiplied by an instantaneous boost factor, the inverse Boltzmann factor for the bias potential at that point, so that the total time after n integration steps is

$$t_{hyper} = \sum_{j=1}^{n} \Delta t_{MD}\, e^{\Delta V(\mathbf{r}(t_j))/k_B T}. \tag{13}$$

Time thus takes on a statistical nature, advancing monotonically but nonlinearly. In the long-time limit, it converges on the correct value for the accelerated time with vanishing relative error. The overall computational speedup is then given by the average boost factor,

$$\mathrm{boost(hyperdynamics)} = t_{hyper}/t_{MD} = \langle e^{\Delta V(\mathbf{r})/k_B T} \rangle_{A_b}, \tag{14}$$

divided by the extra computational cost of calculating the bias potential and its forces. If all the visited states are equivalent (e.g., this is common in calculations to test or demonstrate a particular bias potential), Eq. 14 takes on the meaning of a true ensemble average.

The rate at which the trajectory escapes from a state is enhanced because the positive bias potential within the well lowers the effective barrier. Note, however, that the shape of the bottom of the well after biasing is irrelevant; no assumption of harmonicity is made.

The ideal bias potential should give a large boost factor, have low computational overhead (though more overhead is acceptable if the boost factor is very high), and, to a good approximation, meet the requirements stated above. This is very challenging, since we want, as much as possible, to avoid utilizing any prior knowledge of the dividing surfaces or the available escape paths. Most bias potentials typically are either computationally intensive, tailored to very specific systems, assume localized transitions, or are limited to low-dimensional systems. An important step in that the design of generic and efficient bias potentials has however been recently taken by Miron and Fichthorn, with the introduction of their "bond-boost" bias potential[19]. As the name suggests, the bond-boost potential is composed of pairwise terms that tend to soften the bonds between atoms. The key assumption here is that transitions between states will involve the formation or breaking of some bond so that the proximity to a transition state will be signaled by an unusually large distortion of a bond. If the overall bias potential is then designed to vanish when any bond in the system distorts by more than some critical amount (say by more than 20% of its equilibrium length), then it should be possible to safely turn off the bias before a dividing surface is reached. This approach is not without difficulty (mostly because of the problem of choosing a suitable critical distortion amount), but opens the door to a new generation of bias potentials.

The reader interested in experimenting with hyperdynamics can find relevant examples, both of model and realistic systems, in Ref. 16. [c]

4 Temperature Accelerated Dynamics

In the temperature accelerated dynamics (TAD) method[20], the idea is to speed up the transitions by increasing the temperature, while filtering out the transitions that should not have occurred at the original temperature. This filtering is critical, since without it the state-to-state dynamics will be inappropriately guided by entropically favored higher-barrier transitions. The TAD method is more approximate than the previous two methods, as it relies on harmonic TST, but for many applications this additional approximation is acceptable, and the TAD method often gives substantial boost, with no need for designing bias potential or harnessing parallel computers. Consistent with the accelerated dynamics concept, the trajectory in TAD is allowed to wander on its own to find each escape path, so that no prior information is required about the nature of the reaction mechanisms.

In each basin, the system is evolved at a high temperature T_{high} (while the temperature of interest is some lower temperature T_{low}). Whenever a transition out of the basin is detected, the saddle point for the transition is found. The trajectory is then reflected back into the basin and continued. This "basin constrained molecular dynamics" (BCMD) procedure generates a list of escape paths and attempted escape times for the high-temperature

[c]Note that Eq. 20 in that paper has an error in the d_2 term, which should have a $(2\pi)^2$ factor rather than a (2π) factor.

system. Assuming that TST holds and that the system is chaotic and ergodic, the probability distribution for the first-escape time for each mechanism is an exponential (Eq. 6). Because harmonic TST gives an Arrhenius dependence of the rate on temperature (Eq. 4), depending only on the static barrier height, we can then extrapolate each escape time observed at T_{high} to obtain a corresponding escape time at T_{low} that is drawn correctly from the exponential distribution at T_{low}. This extrapolation, which requires knowledge of the saddle point energy, but not the preexponential factor, can be illustrated graphically in an Arrhenius-style plot ($ln(1/t)$ vs. $1/T$), as shown in Fig. 5. The time for each event seen at T_{high} extrapolated to T_{low} is then

$$t_{low} = t_{high} e^{E_a(\beta_{low} - \beta_{high})}, \tag{15}$$

where, again, $\beta = 1/k_B T$ and E_a is the energy of the saddle point. The event with the shortest time at low temperature is the correct transition for escape from this basin.

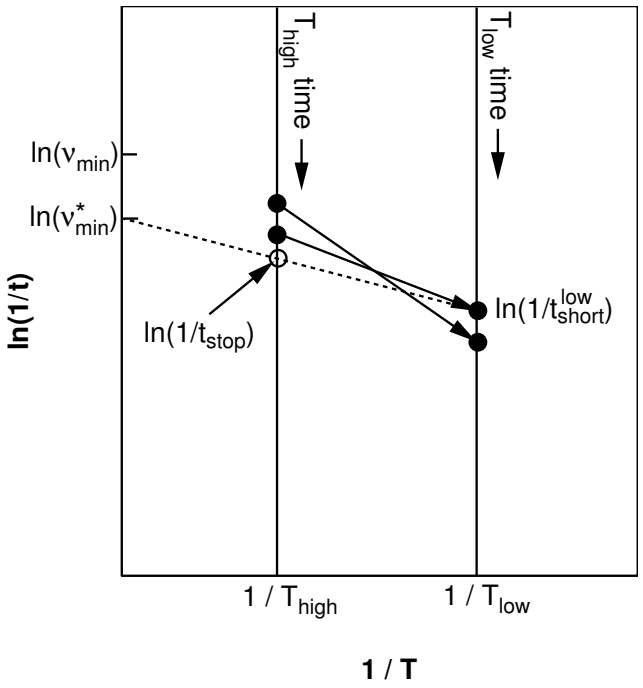


Figure 5. Schematic illustration of the temperature accelerated dynamics method. Progress of the high-temperature trajectory can be thought of as moving down the vertical time line at $1/T_{high}$. For each transition detected during the run, the trajectory is reflected back into the basin, the saddle point is found, and the time of the transition (solid dot on left time line) is transformed (arrow) into a time on the low-temperature time line. Plotted in this Arrhenius-like form, this transformation is a simple extrapolation along a line whose slope is the negative of the barrier height for the event. The dashed termination line connects the shortest-time transition recorded so far on the low temperature time line with the confidence-modified minimum preexponential ($\nu^{\star}_{min} = \nu_{min}/\ln(1/\delta)$) on the y axis. The intersection of this line with the high-T time line gives the time (t_{stop}, open circle) at which the trajectory can be terminated. With confidence 1-δ, we can say that any transition observed after t_{stop} could only extrapolate to a shorter time on the low-T time line if it had a preexponential lower than ν_{min}.

Because the extrapolation can in general cause a reordering of the escape times, a new shorter-time event may be discovered as the BCMD is continued at T_{high}. If we make the additional assumption that there is a minimum preexponential factor, ν_{min}, which bounds from below all the preexponential factors in the system, we can define a time at which the BCMD trajectory can be stopped, knowing that the probability that any transition observed after that time would replace the first transition at T_{low} is less than δ. This "stop" time is given by

$$t_{high,stop} \equiv \frac{\ln(1/\delta)}{\nu_{min}} \left(\frac{\nu_{min} t_{low,short}}{\ln(1/\delta)} \right)^{T_{low}/T_{high}}, \tag{16}$$

where $t_{low,short}$ is the shortest transition time at T_{low}. Once this stop time is reached, the system clock is advanced by $t_{low,short}$, the transition corresponding to $t_{low,short}$ is accepted, and the TAD procedure is started again in the new basin. Thus, in TAD, two parameters govern the accuracy of the simulation: δ and ν_{min}.

The average boost in TAD can be dramatic when barriers are high and T_{high}/T_{low} is large. However, any anharmonicity error at T_{high} transfers to T_{low}; a rate that is twice the Vineyard harmonic rate due to anharmonicity at T_{high} will cause the transition times at T_{high} for that pathway to be 50% shorter, which in turn extrapolate to transition times that are 50% shorter at T_{low}. If the Vineyard approximation is perfect at T_{low}, these events will occur at twice the rate they should. This anharmonicity error can be controlled by choosing a T_{high} that is not too high.

As in the other methods, the boost is limited by the lowest barrier, although this effect can be mitigated somewhat by treating repeated transitions in a "synthetic" mode[20]. This is in essence a kinetic Monte Carlo treatment of the low-barrier transitions, in which the rate is estimated accurately from the observed transitions at T_{high}, and the subsequent low-barrier escapes observed during BCMD are excluded from the extrapolation analysis.

Recently, enhancements to TAD, beyond the "synthetic mode" mentioned above, have been developed that can increase the efficiency of the simulation. For systems that revisit states, the time required to accept an event can be reduced for each revisit by taking advantage of the time accumulated in previous visits[21]. This procedure is exact; no assumptions beyond the ones required by the original TAD method are needed. After many visits, the procedure converges. The minimum barrier for escape from that state (E_{min}) is then known to within uncertainty δ. In this converged mode (ETAD), the average time at T_{high} required to accept an event no longer depends on δ, and the average boost factor becomes simply

$$\text{boost(ETAD)} = \frac{\bar{t}_{low,short}}{\bar{t}_{high,stop}} = \exp\left[E_{min} \left(\frac{1}{k_B T_{low}} - \frac{1}{k_B T_{high}} \right) \right] \tag{17}$$

for that state. The additional boost (when converged) compared to the original TAD can be an order of magnitude or more.

For systems that seldom (or never) revisit the same state, it is still possible to exploit this extra boost by running in ETAD mode with E_{min} supplied externally. One way of doing this is to combine TAD with the dimer method[22]. In this combined dimer-TAD approach, first proposed by Montalenti and Voter[21], upon entering a new state, a number of dimer searches are used to find the minimum barrier for escape, after which ETAD is employed to quickly find a dynamically appropriate escape path. This exploits the power

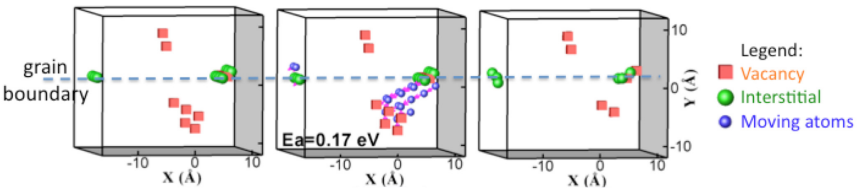


Figure 6. Interstitial emission process at a grain boundary loaded with interstitials in copper: interstitials emitted from the boundary annihilate nearby vacancies.

of the dimer method to quickly find low-barrier pathways, while eliminating the danger associated with the possibility that it might miss important escape paths. Although the dimer method might fail to find the lowest barrier correctly, this is a much weaker demand on the dimer method than trying to find all relevant barriers. In addition, the ETAD phase has some chance of correcting the simulation during the BCMD if the dimer searches did not find E_{min}.

TAD has been used to study a wide variety of systems, and in many cases, it revealed unexpected pathways by which materials evolve. For example, during an investigation of the interaction mechanisms between defects and grain boundaries in copper[23], it was found that during a collision cascade, in which both vacancies and interstitials are created, interstitials are quickly loaded into the boundary. In fact, so many interstitials are trapped at the boundary that the number of vacancies left in the material is typically much greater than if the boundary were not present. On longer timescales, TAD simulations revealed that the boundary acts as a source, emitting those trapped interstitials back into the material to annihilate the vacancies. This unexpected recombination mechanism, illustrated in Fig. 6 has a much lower energy barrier than conventional vacancy diffusion, resulting in enhanced self-healing of the radiation-induced damage and hence to an enhanced radiation tolerance of the material.

TAD has also proved to be particularly useful for studying the long-time behavior of defects produced in collision cascades. In a study using pairwise Coulombic potentials for MgO, the room-temperature annealing of defects generated by MD simulations of cascade collisions was investigated using TAD[24]. In this system, surprisingly high mobilities have been observed for a metastable form of interstitial clusters. In particular, the fastest diffusing species found was a long-lived metastable hexamer that formed during dimer-tetramer encounters. These clusters would not be present at equilibrium because of their high energy, but they form naturally from the aggregation of radiation-induced defects. A reaction-diffusion equation based upon these atomistic results showed that the presence of these metastable clusters would have a significant impact on the size and density of interstitial dislocation loops in the material. This is a good example of complex kinetics that can reveal themselves when long time dynamics are directly simulated. Because the barriers in this system were typically high (relative to $T = 300$ K), TAD yielded substantial boost factors, allowing simulations on very long time scales. This was also made possible through the use of the dimer-TAD approach, which allowed for faster acceptance of processes that had a high barrier but were still the lowest barrier to leave the state.

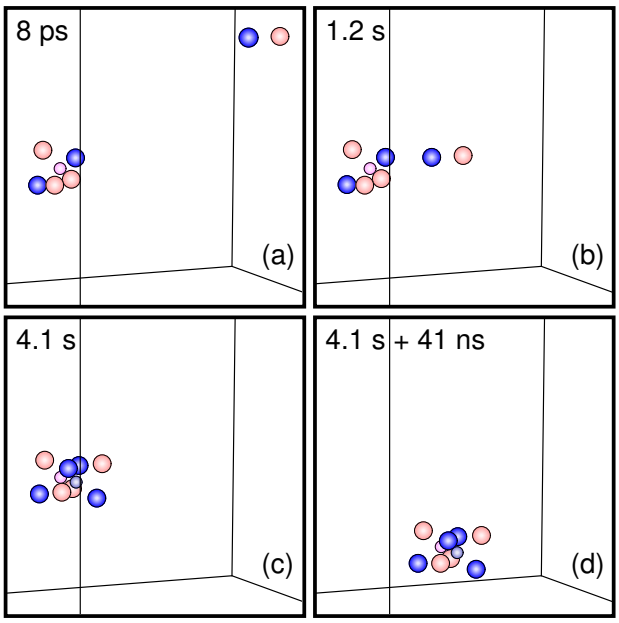


Figure 7. TAD simulation of the formation of I_6 at 300 K. Only defects in the lattice (spheres represent interstitials and cubes represent vacancies) are shown. (a) An I_2 and I_4 begin about 1.2 nm apart. (b) By $t = 1.2$ s, the I_2 approaches the immobile I_4. (c) By $t = 4.1$ s, the combined cluster anneals to form the metastable I_6, (d) which diffuses on the ns time scale with a barrier of 0.24 eV (after Ref. 24).

5 Choosing the Right AMD Method

As these accelerated dynamics methods become more widely used and further developed (including the possible emergence of new methods), their application to important problems in materials science will continue to grow. We conclude this article by comparing and contrasting the three methods presented here, with some guidelines for deciding which method may be most appropriate for a given problem. We point out some important limitations of the methods, areas in which further development may significantly increase their usefulness. Finally, we discuss the prospects for these methods in the immediate future.

The key feature of all of the accelerated dynamics methods is that they collapse the waiting time between successive transitions from its natural time (τ_{rxn}) to (at best) a small number of vibrational periods. Each method accomplishes this in a different way. TAD exploits the enhanced rate at higher temperature, hyperdynamics effectively lowers the barriers to escape by filling in the basin, and parallel-replica dynamics spreads the work across many processors.

The choice of which accelerated dynamics method to apply to a problem will typically depend on three factors. The first is the desired level of accuracy in following the exact dynamics of the system. As described previously, parallel replica is the most accurate of the three methods; the only assumption is that the kinetics are first order. Not even TST is assumed, as correlated dynamical events are treated correctly in the method. This is not true with hyperdynamics, which does rely upon the assumptions of TST, in particular

the absence of correlated events. Finally, temperature accelerated dynamics makes the further assumptions inherent in the harmonic approximation to TST, and is thus the most approximate of the three methods. If complete accuracy is the main goal of the simulation, parallel replica is the superior choice.

The second consideration is the potential gain in accessible time scales that the accelerated dynamics method can achieve for the system. Traditionally, TAD was the method of choice when considering this factor. While in all three methods the boost for escaping from each state will be limited by the smallest barrier, if the barriers are high relative to the temperature of interest, TAD typically achieves large boost factors. In principle, hyperdynamics can also achieve very significant boosts, but, in practice, the design of suitable bias potentials can be difficult. However, coupled with a good bias potential, hyperdynamics can also provide substantial boosts. Finally, if parallel computing resources are available, parallel replica dynamics can provide significant speedups; up to the number of replicas used. With the continued increase in parallel computing power, the future of parallel replica dynamics is bright, and it should take an increasingly important role in the modern computational toolbox.

The last main factor determining which method is best suited to a problem is the shape of the potential energy surface (PES). Both TAD and hyperdynamics require that the PES be relatively smooth. In the case of TAD, this is because saddle points must be found and standard techniques for finding them often perform poorly for rough landscapes. The same is true for the hyperdynamics bias potentials that require information about the shape of the PES. Parallel replica, however, only requires a method for detecting transitions. No further analysis of the potential energy surface is needed. Thus, if the PES describing the system of interest is relatively rough, parallel replica dynamics may be the only method that can be applied effectively.

6 Conclusion

Since their introduction about 15 years ago, the AMD methods have proven useful in a variety of situations where the timescales of interest are out of reach of direct MD and where the kinetics are too rich to be adequately described with a limited list of predetermined pathways. When the activation barriers between the different states are high relative to the thermal energy, any of the AMD methods can yield colossal accelerations, providing a view of atomistic dynamics over unprecedented timescales. Further, by leveraging the particular strength of each of the methods a wide variety of situations can be efficiently simulated. If the methods have enjoyed considerable successes, they have also sometimes failed to provide significant acceleration. In most, if not all, of the problematic cases, this failure is related to the presence of large numbers of states connected by very low barriers where there is no separation of timescale between vibration and escape out of single potential energy basins. While some strategies have been put forward to mitigate this issue (e.g., superstate parallel-replica dynamics[14], synthetic TAD[20], state-bridging hyperdynamics[25]), more work is required before victory can be claimed. For example, an on-the-fly state definition algorithm that automatically identifies an exploitably large separation of timescales would tremendously extend the reach of parallel-replica dynamics, enabling it to address notoriously difficult problems like protein folding, where the energy landscape is extremely rough. Statistical analysis tools could also be used to identify dynamically

"irrelevant" states that could be ignored or lumped with others without affecting the long-time dynamics. Many of these ideas are now being explored and will hopefully lead to more general and robust AMD methods in the next few years.

Acknowledgments

These notes are based on a previous set of notes prepared for the Advanced Study Institute on Radiation Effects in Solids held in Erice, Italy in 2004[26]. Work at Los Alamos National Laboratory (LANL) was supported by the DOE Office of Basic Energy Sciences and by the LANL Laboratory Directed Research and Development program. LANL is operated by Los Alamos National Security, LLC, for the National Nuclear Security Administration of the US DOE under Contract No. DE-AC52-06NA25396.

References

1. H. Eyring, *The activated complex in chemical reactions*, Journal of Chemical Physics, **3**, 107–15, 1935.
2. P. Pechukas, *Transition-State Theory*, Annual Review Of Physical Chemistry, **32**, 159–177, 1981.
3. D. G. Truhlar, B. C. Garrett, and S. J. Klippenstein, *Current status of transition-state theory*, J. Phys. Chem., **100**, 12771–12800, 1996.
4. D. Chandler, *Statistical-Mechanics Of Isomerization Dynamics In Liquids And Transition-State Approximation*, Journal Of Chemical Physics, **68**, no. 6, 2959–2970, 1978.
5. A. F. Voter and J. D. Doll, *Dynamical Corrections to Transition State Theory for Multistate Systems: Surface Self-Diffusion in the Rare-Event Regime*, Journal of Chemical Physics, **82**, 80–92, 1985.
6. C. H. Bennett, *Molecular-Dynamics And Transition-State Theory - Simulation Of Infrequent Events*, ACS Symposium Series, **1977**, no. 46, 63–97, 1977.
7. A. F. Voter, F. Montalenti, and T. C. Germann, *Extending the time scale in atomistic simulation of materials*, Annual Review Of Materials Research, **32**, 321–346, 2002.
8. R. Marcelin, Ann. Physique, **3**, 120–231, 1915.
9. E. P. Wigner, Z. Phys. Chemie B, **19**, 203, 1932.
10. A. F. Voter and J. D. Doll, *Transition-State Theory Description Of Surface Self-Diffusion - Comparison With Classical Trajectory Results*, Journal of Chemical Physics, **80**, 5832–8, 1984.
11. B. J. Berne, M. Borkovec, and J. E. Straub, *Classical And Modern Methods In Reaction-Rate Theory*, J. Phys. Chem., **92**, 3711–25, 1988.
12. G. H. Vineyard, *Frequency Factors And Isotope Effects In Solid State Rate Processes*, J. Phys. Chem. Solids, **3**, 121–7, 1957.
13. A. F. Voter, *Parallel replica method for dynamics of infrequent events*, Physical Review B, **57**, 13985–8, 1998.
14. O. Kum, B. M. Dickson, S. J. Stuart, B. P. Uberuaga, and A. F. Voter, *Parallel replica dynamics with a heterogeneous distribution of barriers: Application to n-hexadecane pyrolysis*, The Journal of Chemical Physics, **121**, no. 20, 9808–9819, 2004.

15. B. P. Uberuaga, S. J. Stuart, and A. F. Voter, *Parallel replica dynamics for driven systems: Derivation and application to strained nanotubes*, Phys. Rev. B, **75**, 014301, Jan 2007.
16. A. F. Voter, *A method for accelerating the molecular dynamics simulation of infrequent events*, Journal of Chemical Physics, **106**, 4665–77, 1997.
17. J. P. Valleau and S. G. Whittington, "A guide to monte carlo for statistical mechanics: 1. highways", in: Statistical Mechanics. A. A Modern Theoretical Chemistry, B. J. Berne, (Ed.), vol. 5, pp. 137–68. Plenum, New York, 1977.
18. B. J. Berne, G. Ciccotti, and D. F. Coker, (Eds.), *Classical and Quantum Dynamics in Condensed Phase Simulations*, World Scientific, Singapore, 1998.
19. R. A. Miron and K. A. Fichthorn, *Accelerated molecular dynamics with the bond-boost method*, The Journal of Chemical Physics, **119**, no. 12, 6210–6216, 2003.
20. M. R. Sørensen and A. F. Voter, *Temperature-accelerated dynamics for simulation of infrequent events*, Journal of Chemical Physics, **112**, 9599–606, 2000.
21. F. Montalenti and A. F. Voter, *Exploiting past visits or minimum-barrier knowledge to gain further boost in the temperature-accelerated dynamics method*, Journal Of Chemical Physics, **116**, no. 12, 4819–4828, MAR 2002.
22. G. Henkelman and H. Jónsson, *A dimer method for finding saddle points on high dimensional potential surfaces using only first derivatives*, Journal Of Chemical Physics, **111**, no. 15, 7010–7022, OCT 1999.
23. X.-M. Bai, A. F. Voter, R. G. Hoagland, M. Nastasi, and B. P. Uberuaga, *Efficient Annealing of Radiation Damage Near Grain Boundaries via Interstitial Emission*, Science, **327**, no. 5973, 1631–1634, 2010.
24. B. P. Uberuaga, R. Smith, A. R. Cleave, F. Montalenti, G. Henkelman, R. W. Grimes, A. F. Voter, and K. E. Sickafus, *Structure and Mobility of Defects Formed from Collision Cascades in MgO*, Phys. Rev. Lett., **92**, 115505, Mar 2004.
25. R. A. Miron and K. A. Fichthorn, *Multiple-Time Scale Accelerated Molecular Dynamics: Addressing the Small-Barrier Problem*, Phys. Rev. Lett., **93**, 128301, Sep 2004.
26. B. P. Uberuaga and A. F. Voter, *Accelerated Molecular Dynamics Methods*, in: Radiation Effects in Solids, Kurt E. Sickafus, Eugene A. Kotomin, and Blas P. Uberuaga, (Eds.), vol. 235 of *NATO Science Series II: Mathematics, Physics and Chemistry*, Springer, Dordrecht, The Netherlands, 2007.

Tracking the Dynamics of Systems Evolving through Infrequent Transitions in a Network of Discrete States

Doros N. Theodorou

School of Chemical Engineering
National Technical University of Athens
9 Heroon Polytechniou Street, Zografou Campus, 157 80 Athens, Greece
E-mail: doros@chemeng.ntua.gr

Many physicochemical, materials, and biological systems whose dynamics is too slow to be addressed via conventional molecular dynamics (MD) simulations can be considered as evolving in time through infrequent transitions in a network of discrete states, each state providing a coarse-grained description of a domain in multidimensional configuration space. We briefly discuss how states can be defined starting from the detailed potential energy hypersurface of such a system and how rate constants for transitions between states can be estimated based on the theory of infrequent events. We then concentrate on tracking the evolution of a system as a succession of transitions between states. Two general approaches are introduced for this: Kinetic Monte Carlo simulation, and analytical solution of the master equation for the time-dependent probabilities of occupancy of the states. For the latter approach we outline how time autocorrelation functions can be computed under equilibrium and nonequilibrium conditions. We present examples from the computation of diffusivities of gases in zeolites and in glassy amorphous polymers. We then introduce the method of Dynamic Integration of a Markovian Web (DIMW), designed to track relaxation towards equilibrium from a narrow initial distribution among states by solving the master equation in a network of explored states that is progressively augmented on the fly. We present an application of the DIMW method to physical ageing in a glassy polymer. Finally, we outline how computation of the long-time evolution in a network of states can be simplified by "lumping" states into clusters of states.

1 Introduction

The dynamics of many physical, chemical, materials, and biological systems is slow because it proceeds as a succession of infrequent transitions between domains in their configuration space, which we shall call "states". The states constitute "basins" of low potential energy with respect to the generalized coordinates spanning configuration space, or of low free energy with respect to a set of order parameters providing a coarse-grained description of the system. Each state contains one or more local minima of the the free energy. Transitions between states are infrequent events, in the sense that the mean waiting time for transition out of a state is long in comparison to the time required for the system to establish a restricted equilibrium distribution among configurations in the state. The entire configuration space can be tessellated into states. Representing each state in a coarse-grained sense by a point in configuration or in order parameter space and connecting all pairs of states between which a transition is possible, one obtains a graph, or network of states. Examples of phenomena that can be modelled as occurring through a succession of transitions in a network of states include diffusion of defects and impurities in metals and semiconductors[1]; of gas molecules in amorphous polymers[2]; of bulky hydrocarbons in microporous solids, such as zeolites[3]; structural relaxation and plastic deformation in glasses[4]; phase transitions in molecular and atomic clusters[5]; surface diffusion[6]; protein folding[7]; and chemical reactions[8].

The possibility of coarse-graining dynamics into a sequence of transitions in a network of states is of strategic importance for understanding and predicting macroscopic time-dependent properties from atomic-level structure and interactions. The longest times that can be simulated with atomistic MD on conventional computational means are microseconds. (Note, however, that millisecond-long MD runs on specialized hardware have been reported recently[9]). This is too short by many orders of magnitude in comparison with the experimental time scales of most phenomena of interest. A more efficient strategy than "brute-force" MD is to construct a network of states i and compute the rate constants $k_{i\to j}$ between them from atomic-level information. By definition, the rate constant $k_{i\to j}$ is a conditional probability per unit time that a transition to state j will occur, provided the system is in state i.

Once states and interstate rate constants are known, the system evolution at the state level can be tracked by solving the master equation:

$$\frac{\partial P_i(t)}{\partial t} = \sum_{j\neq i} P_j(t)k_{j\to i} - P_i(t)\sum_{j\neq i} k_{i\to j} \quad \text{, or} \quad \frac{\partial \mathbf{P}(t)}{\partial t} = \mathbf{K}\mathbf{P}(t) \tag{1}$$

The transition rate constant $k_{i\to j}$ is independent of time, thanks to the time scale separation which makes the transition an infrequent event[10,11]. The evolution of the system in state space is a Poisson process[12]. $P_i(t)$ is the probability of occupancy of state i at time t. According to Eq. 1 this changes as a result of influx of probability from other states and efflux of probability to other states. State occupancy probabilities are normalized over all n states of the system. The time-dependent vector $\mathbf{P}$ in the matrix representation of Eq. 1 has all the $P_i(t)$ as elements. The $n \times n$ rate constant matrix is defined by $K_{ij} = k_{j\to i}$, $K_{ii} = -\sum_{j\neq i} k_{i\to j}$. At very long times, the system will adopt its equilibrium probability distribution among states, $\mathbf{P}(\infty)$. This is a stationary solution of the master equation, Eq. 1, by virtue of the condition of microscopic reversibility satisfied by the rate constants:

$$k_{i\to j}P_i(\infty) = k_{j\to i}P_j(\infty) \tag{2}$$

These notes address the problem of how to solve the master equation, Eq. 1, and learn about the long-time dynamics of a system evolving through a succession of infrequent transitions between discrete states. Sections 2 and 3 briefly discuss how states can be identified and rate constants for transitions between states can be computed, given the potential energy as a function of atomic coordinates and the masses of all atoms in the system. Sec. 4 reviews the basics of Kinetic Monte Carlo (KMC) simulation for generating stochastic trajectories consisting of long successions of jumps between states. Sec. 5 outlines a method for analytical solution of the master equation and computation of time autocorrelation functions therefrom. Example applications of the KMC and master equation solution strategies to diffusion problems are presented in Sections 6 (for xenon in the zeolite silicalite) and 7 (for CO_2 in a glassy poly(amide imide)). Sec. 8 addresses the more complex problem of nonequilibrium relaxation of a system that is initially confined to a small subset of states. States are not known a priori, but have to be charted out as the system relaxes. We introduce the "Dynamic Integration of a Markovian Web" (DIMW) method for solving the master equation in a network of states that is progressively augmented "on the fly". We apply DIMW to the very challenging problem of tracking structural relaxation in a polymer glass. Finally, in Sec. 9 we discuss a systematic approach for "lumping" groups of states that communicate with each other through relatively fast transitions into single "metas-

tates" and thereby reducing the number of states needed for the description of dynamics at long times.

2 Identifying States

States are regions of configuration space where the system is trapped for long periods of time. Let f be the number of degrees of freedom needed to specify the microscopic configuration of a system. For a classical system of N particles with periodic boundary conditions described in full detail, $f = 3N - 3$. We will use the f-dimensional vector $\mathbf{r}$ to denote the configuration of a system. We will also use $\mathbf{x}$ to denote the f-dimensional vector of mass-weighted coordinates, with elements $m_l^{1/2} r_l^\alpha$ with m_l being the mass of particle l ($l = 1, 2, \ldots, N$) and r_l^α being the position coordinate of particle l along direction α ($\alpha = 1, 2, 3$). Let $\mathcal{V}(\mathbf{x})$ be the potential energy of the system as a function of the mass-weighted coordinates. A state is a domain in $\mathbf{x}$-space surrounding a local minimum of $\mathcal{V}(\mathbf{x})$.

For small f, an exhaustive determination of all minima and consequent identification of all states and dividing surfaces between them is possible. For example, in the case of low-occupancy diffusion of a monatomic sorbate in a zeolite represented as a rigid framework[13], $f = 3$ (the three translational degrees of freedom of the sorbate within the rigid zeolite). The volume of the asymmetric unit of the zeolite unit cell was discretized into voxels of edge length approximately 0.2 Å. A steepest descent trajectory was initiated at the center of each voxel, terminating in a local minimum of $\mathcal{V}(\mathbf{x})$. The minimization was refined using a quasi-Newton algorithm. In this way, a "drainage pattern" was constructed in three-dimensional space, leading to the local minima. The set of all voxels from which the steepest descent construction terminated at a certain minimum was assigned to the state of that minimum. Similarly, the dividing surface between two states i and j was defined as the set of all faces (squares) shared by two voxels such that the steepest descent construction from one of the voxels leads to minimum i, while that from the other voxel leads to minimum j. An exhaustive identification of all states was similarly undertaken in the work of Snurr et al.[14] on the diffusion of benzene in the zeolite silicalite, where both the zeolite framework and the sorbate molecule were represented as rigid. In this case, $f = 6$ degrees of freedom (three translational and three orientational of the benzene relative to the framework) come into play. A very large number of insertions of the benzene at random positions and orientations within the asymmetric unit was used as a first step. From each configuration resulting from insertion that did not exceed a certain energy threshold, a quasi-Newton minimization was initiated, leading to an energy minimum in $\mathcal{V}(\mathbf{x})$ in six-dimensional configuration space, representing a sorption state. Increasing the number of random insertions for the initial guess configuration did not lead to any other minima; this, and the symmetry of determined minima, indicated that the calculation was exhaustive.

In more complex situations, where f is larger, the identification of states can be greatly facilitated by geometric analysis. An example is provided by Greenfield's study of methane diffusion in glassy atactic polypropylene[15]. Static configurations of the amorphous polymer, constituting local minima of its potential energy, were used as a starting point. Within each static configuration, the volume accessible to spherical probes of various radii smaller than the van der Waals radius of the penetrant of interest (methane) was analyzed using a Delaunay tessellation and clustering algorithm[16]. For large probe radius the accessible vol-

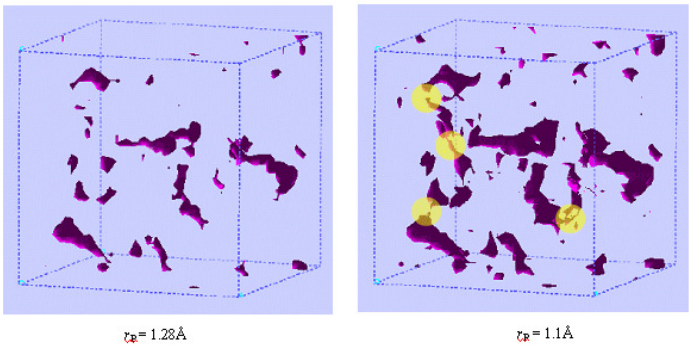


Figure 1. Geometric analysis of accessible volume in an amorphous poly(amide imide) configuration aimed at the identification of states and transition paths for diffusion of CO_2 at infinite dilution within the polymer. Analysis with a spherical probe of radius $r_P = 1.28$ Å reveals disjoint. elongated clusters of accessible volume. Analysis with a smaller probe radius $r_P = 1.1$ Å reveals "necks" of accessible volume connecting the original clusters. The positions of the necks (encircled in the figure) are used as initial guesses for the center of mass position of the penetrant at the saddle point along an elementary transition path.

ume consists of relatively small disjoint clusters. As the probe radius decreases, accessible volume clusters grow in size and some clusters merge at narrow "necks" of accessible volume. The position of each of these necks between a pair of clusters is used as an initial guess for the position of the penetrant at the saddle point of the energy along the transition from a (meta)state of occupancy of one cluster to a (meta)state of occupancy of the other. A saddle point of $\mathcal{V}(\mathbf{x})$ is computed from the geometrically obtained neck position as follows: The center of the penetrant is placed at the neck position and a saddle point is first calculated with respect to the three translational degrees of freedom of the penetrant, keeping the configuration of the polymer fixed. Using this three-dimensional saddle point as an initial guess, the number of system degrees of freedom with respect to which the saddle point is calculated is progressively increased, by including more and more atoms of the polymer in concentric spheres around the penetrant. This calculation goes on until the saddle point energy becomes asymptotic with respect to inclusion of additional polymer degrees of freedom[15]. The saddle point searches can be performed using the Cerjan-Miller type algorithm of Baker[17]. Having obtained a multidimensional saddle point in both penetrant and matrix degrees of freedom, an entire transition path is constructed using Fukui's intrinsic reaction coordinate approach[18]: Starting at the saddle point, the system is displaced by a small step along the eigenvector corresponding to the negative eigenvalue of the Hessian matrix of second derivatives $\partial\mathcal{V}/(\partial\mathbf{x}\partial\mathbf{x}^{\mathrm{T}})$. Subsequently, a steepest descent construction in $\mathcal{V}(\mathbf{x})$ is undertaken using small steps in $\mathbf{x}$, until a local minimum of $\mathcal{V}(\mathbf{x})$ is reached. Completing this construction on either side of the saddle point, i.e. with the initial displacement first along the positive and then along the negative direction of the eigenvector, yields an entire reaction path between two (meta)states, in which different adjacent clusters of accessible volume are occupied by the penetrant. This calculation has been extended by Vergadou to more complex multiatom penetrants, such as CO_2 in a poly(amide imide) (see Fig. 1)[19].

When no guidance is provided by geometry or crystal symmetry, the identification of states is considerably more involved. Kopsias[20] and Boulougouris[21] addressed the problem of finding connected minima in the full configuration space ($f = 3N - 3$) in order to track structural relaxation in a glass. Given a minimum of $\mathcal{V}(\mathbf{x})$, they strove to find as many as possible other minima connected to it via transition paths passing through a single first-order saddle point of $\mathcal{V}(\mathbf{x})$. For this purpose, they undertook saddle point searches in f-dimensional space, starting off along the lowest-curvature eigendirections of the Hessian at the original minimum. Beyond a certain number of searches, no new saddle points were located (the algorithm returned saddle points that had already been found); this was taken as an indication that all relevant transitions out of the initial minimum (i.e., transitions taking the system over reasonably low energy barriers), had been found. In Ref. 20 the saddle point searches were conducted using the Baker algorithm, while Ref. 21 employed the dimer method of Henkelman and Jónsson, which does not require second derivatives[22]. From each saddle point located in this way, a pair of steepest descent constructions was undertaken in full configuration space using Fukui's intrinsic reaction coordinate approach, as described above. On one side the original minimum was recovered, while on the other side the steepest descent construction led to a new minimum adjacent to the original one. The procedure was repeated from each new minimum, in order to map out a network of minima, or "states".

In many problems it is a good approximation to assume that the reaction coordinate taking the system from a state to another state is shaped by a relatively small subset of "primary" degrees of freedom, the remaining degrees of freedom fluctuating rapidly and achieving a constrained equilibrium distribution subject to the values of the primary set. Then, system "states" can be defined as local minima of the *potential of mean force* with respect to the primary subset of degrees of freedom. Although calculating the potential of mean force is generally a challenge for molecular simulations, the reduction of dimensionality in passing from the full configuration space to the subspace of primary degrees of freedom greatly facilitates the definition of states and transitions between them. An example of such an approach based on the potential of mean force is provided by Forester and Smith's[23] calculations on the diffusion of benzene in silicalite. These authors used a unidimensional reaction coordinate, corresponding to the projection of the center of mass position of the sorbed benzene on the axes of straight or sinusoidal channel segments in the zeolite. The latter axes were taken as rectilinear, for simplicity. All other degrees of freedom (translational of the benzene in directions transverse to the channel axis, orientational of the benzene, and vibrational of the surrounding zeolite framework) were integrated over at each position along an axis. The potential of mean force was computed by dragging the benzene along the channels, through the "blue moon ensemble" MD method. States were readily identified as local minima of the potential of mean force (see also Sec. 3).

3 Calculating Rate Constants

Once states have been defined, the transition rate constants $k_{i\to j}$ can be computed by a variety of methods. We briefly outline some of these methods here. For a more thorough treatment, the reader is referred to standard texts on molecular simulation[24].

If transitions are subject to relatively low barriers (say, up to 7 $k_{\mathrm{B}}T$), such that rate constants $k_{i\to j}$ are relatively high (say, up to ns^{-1}), then rate constants can be estimated

by MD simulation. All one needs is a technique to map every configuration recorded in the course of a MD trajectory onto a state. Very often, when states are defined as regions around local minima in configuration space, this mapping is accomplished by direct energy minimization leading to the closest energy minimum or "inherent structure"[25]. A reduced trajectory of states visited is thus accumulated in parallel with the MD trajectory. Switches between states can readily be identified along this reduced trajectory. Rate constants can be computed by statistical analysis of the reduced trajectory, capitalizing on the exponential distribution of waiting times that characterizes Poisson processes. A simple method that can be used for this purpose is "hazard plot analysis", outlined in the following paragraphs[26].

We first introduce some definitions that are generally applicable to any stochastic process involving infrequent transitions. The particular example of stochastic process we will have in mind is that of exiting a specific state i in the network of states we have introduced in Sec. 1, once the system has entered that state. The rate constant for this process is $k_{i\rightarrow} = \sum_{j\neq i} k_{i\rightarrow j}$. For the stochastic process considered, let $\hat{P}(t)$ be the probability of having undergone a transition at time t. In our particular example, $\hat{P}(t)$ can be interpreted as the cumulative distribution function of residence (or "waiting") times within state i. The hazard rate, $\hat{h}(t)$, is defined such that $\hat{h}(t)dt$ equals the (conditional) probability that a system (in an ensemble of systems governed by the stochastic process) which has not undergone a transition until time t, will undergo a transition at time t. From the definitions of $\hat{P}(t)$ and $\hat{h}(t)$, the following differential equation is satisfied:

$$\hat{P}(t+dt) = \hat{P}(t) + \left[1 - \hat{P}(t)\right] \hat{h}(t)dt \tag{3}$$

or

$$d\hat{P}/dt = \left[1 - \hat{P}(t)\right] \hat{h}(t) \tag{4}$$

Eq. 4 must be solved with initial condition $\hat{P}(0) = 0$. The solution is

$$\hat{P}(t) = 1 - \exp\left[-\int_0^t \hat{h}(t')dt' = \right] = 1 - \exp\left[-\hat{H}(t)\right] \tag{5}$$

where we have defined the cumulative hazard $\hat{H}(t)$ as

$$\hat{H}(t) = \int_0^t \hat{h}(t')dt'. \tag{6}$$

For a Poisson process, the hazard rate $\hat{h}(t)$ is a constant, independent of time. In our example of exiting state i, $\hat{h}(t) = k_{i\rightarrow}$, a constant at sufficiently long times. This is because, once the system enters state i which is in a region surrounded by high energy barriers, it will quickly thermalize (distribute itself according to the requirements of a restricted equilibrium) within state i and forget how it came there. Exit from state i is an infrequent event because of the time scale separation between the correlation time for thermalizing within state i and the mean waiting time for escaping state i. Note the Markovian character imparted to the process by this time scale separation. For a Poisson process, then, the

cumulative distribution function of waiting times has the form:

$$\hat{P}(t) = 1 - \exp\left(-k_{i\rightarrow}t\right) \tag{7}$$

and the probability density of waiting times is exponential:

$$\hat{\rho}(t) = k_{i\rightarrow} \exp\left(-k_{i\rightarrow}t\right). \tag{8}$$

The mean waiting time in state i is readily computed from Eq. 8 as $k_{i\rightarrow}^{-1}$.

In view of these definitions and properties of Poisson processes, the following computational procedure emerges for computing the rate constant $k_{i\rightarrow}$ from the reduced trajectory (sequence of visited states) onto which a MD run has been mapped. The MD run must be long enough to sample a large number of transitions out of state i. One goes through the reduced trajectory and measures all time intervals t_l between an entry into state i and the immediately following exit from i to any other state. One orders these residence times as $t_1 \leq t_2 \leq \ldots \leq t_n$, where n is the total number of visits to state i observed in the reduced trajectory. Clearly, based on the reduced trajectory, the quantity $\hat{P}(t_l) = l/n, 1 \leq l \leq n$, provides an estimate of the probability that the residence time in state i will not exceed t_l, i.e. an estimate of the cumulative probability distribution of waiting times at t_l. One forms an estimate of the cumulative hazard at t_l, $\hat{H}(t_l)$, as

$$\hat{H}(t_l) = \frac{1}{n} + \frac{1}{n-1} + \ldots + \frac{1}{n-l+1} \tag{9}$$

One then plots $\hat{H}(t_l)$ as a function of t_l for $l = 1, 2, \ldots n$. At short times the resulting hazard plot may display some curvature, associated with fast recrossing events of the dividing surfaces between state i and its surrounding states. At long times, however, if time scale separation holds, the hazard plot becomes linear. The slope at long times is the sought rate constant $k_{i\rightarrow}$. Individual rate constants $k_{i\rightarrow j}$ can readily be obtained from $k_{i\rightarrow}$ as

$$k_{i\rightarrow j} = k_{i\rightarrow} \frac{\text{Number of times exit from } i \text{ occurred to } j}{n} \tag{10}$$

The rationale behind Eq. 9 is that, for a Poisson process, the cumulative hazard $\hat{H}(t)$ is related to the cumulative probability distribution of residence times $\hat{P}_i(t)$ via Eqs. 5 and 6, hence $\hat{H}(t) = -\ln\left[1 - \hat{P}(t)\right]$. The reader can readily verify that the right-hand side of Eq. 9 is an estimate of $-\ln(1 - l/n) \simeq \int_0^{l/n} \frac{1}{1-x} dx$.

It is advisable to make sure that rate constants extracted from hazard plot analysis are invariant to the frequency of conducting minimizations along the MD trajectory to form the reduced trajectory; to ensure that no transitions are missed, the latter frequency, as well as the frequency of recording configurations along the MD trajectory, should be considerably higher than the rate constant of the fastest transition taking place in the system.

Fig. 2 displays an example of a hazard plot for transition out of a state (basin of the potential energy) of a glassy binary Lennard-Jones mixture at low temperature[27].

When energy barriers between states are high in relation to $k_{\mathrm{B}}T$ and rate constants are correspondingly low, transitions between states cannot be sampled adequately by straightforward MD. One way to get around this problem is to resort to temperature-accelerated

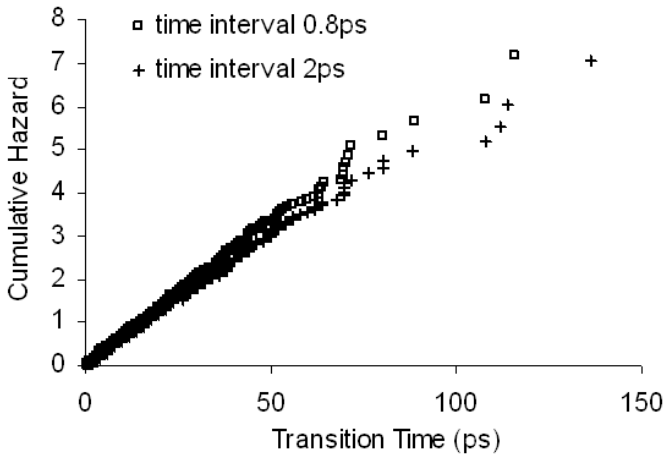


Figure 2. Hazard plot for transitions out of a given state (basin of the potential energy) for a glassy binary Lennard-Jones mixture at reduced density 1.1908 and temperature 9 K, as computed from a canonical MD simulation. Two sets of calculations are presented, one using a 0.8 ps interval between minimizations (squares) and another one using a 2 ps interval between minimizations (crosses) in forming the reduced trajectory (sequence of states visited as a function of time).

dynamics (TAD) simulations, as originally proposed by Voter and collaborators[28]. MD simulation at a higher temperature is used to access transition pathways. Waiting times obtained at the higher temperature are extrapolated down to the temperature of interest using the Arrhenius dependence of rate constants on temperature. The method has been used to great advantage in surface diffusion problems[28]. Tsalikis et al.[29] have combined microcanonical simulations at various energy levels with the histogram reweighting method to obtain rate constants in the spirit of TAD for transitions between basins in configuration space in the course of structural relaxation of a glassy binary Lennard-Jones mixture.

Infrequent event analyses based on dynamically corrected transition-state theory have found widespread use in the computation of rate constants from simulations. These analyses are based on the theory of Bennett[30] and Chandler[10], which was extended to multistate systems by Voter and Doll[6]. Let us assume that the boundary of state i in configuration space is described by an equation $C_i(\mathbf{x}) = 0$, where C_i is a continuous, differentiable function of the mass-weighted coordinates $\mathbf{x}$. $C_i(\mathbf{x}) < 0$ for all points in state i, while $C_i(\mathbf{x}) > 0$ for all points outside state i. Then, $\mathbf{n}_i = \nabla C_i(\mathbf{x}) / |\nabla C_i(\mathbf{x})|$ is a unit vector normal to the boundary surface of state i at point $\mathbf{x}$ pointing towards the outside of the state. Furthermore, the function $h_i(\mathbf{x}) = 1 - H\left(C_i(\mathbf{x})\right)$, with $H(x)$ being the Heaviside step function, equals 1 if $\mathbf{x}$ belongs to state i and zero otherwise. The rate constant for transitions from i to any other state j can be expressed as

$$k_{i\to j}(t) = \frac{\left\langle \mathbf{n}_i\left(\mathbf{x}(0)\right) \cdot \dot{\mathbf{x}}(0) \delta\left(C_i\left(\mathbf{x}(0)\right)\right) \left|\nabla C_i\left(\mathbf{x}(0)\right)\right| h_j\left(\mathbf{x}(t)\right) \right\rangle}{P_i(\infty)} \tag{11}$$

In Eq. 11 the average is taken over all equilibrium dynamical trajectories of the system. The numerator has nonzero contributions from those trajectories which cross the boundary (hyper)surface of state i at time 0 and find themselves in state j after time t. The averaged quantity in the numerator is the component of the mass-weighted velocity $\dot{\mathbf{x}}$ at time 0 normal to the boundary surface of state i times a delta function along the component of $\mathbf{x}$ normal to the boundary surface which requires that the system be on that surface at time 0. The denominator is the equilibrium probability of occupancy of state i (compare Eq. 2). Clearly, the right-hand side of Eq. 11 has dimensions of inverse time, as expected of a rate constant. As discussed by Chandler[10] and Voter and Doll[6], thanks to the time scale separation making exit from state i an infrequent event, $k_{i\to j}$ will practically reach a time-independent plateau value at times sufficiently longer than the time required for internal equilibration within state i.

It is useful to consider the rate constant $k_{i\to j}$ given by Eq. 11 as a product of a transition-state theory estimate of the rate constant for exiting state i times a dynamical correction factor:

$$k_{i\to j}(t) = k_{i\to}^{\mathrm{TST}} f_{\mathrm{d},i\to j} \tag{12}$$

Transition state theory rests on an approximation: It assumes that, whenever the system finds itself on the boundary surface of state i with momentum directed towards the outside of state i, then a successful transition out of state i will occur. In reality, this is not necessarily the case because of fast recrossings of the boundary surface at short times. Mathematically, $k_{i\to}^{\mathrm{TST}}$ is obtained by replacing $h_j\left(\mathbf{x}(t)\right)$ in the numerator of Eq. 11 with $1 - h_i(\mathbf{x}(0^+)) = H\left(\mathbf{n}_i(\mathbf{x}(0)) \cdot \dot{\mathbf{x}}(0)\right)$. The averaging over configuration and momentum space can be separated, the momentum-space average reducing to a Boltzmann-weighted mean of the component of the mass-weighted velocity vector normal to the boundary surface over the positive semiaxis. The result is:

$$k_{i\to}^{\mathrm{TST}} = \frac{1}{(2\beta\pi)^{1/2}} \frac{\displaystyle\int_{\text{bound. surf. of state } i} d^{f-1}x \exp\left[-\beta\mathcal{V}(\mathbf{x})\right]}{\displaystyle\int_{\text{state i}} d^{f}x \exp\left[-\beta\mathcal{V}(\mathbf{x})\right]} \tag{13}$$

The reader is reminded that $\mathbf{x}$ is the vector of mass-weighted coordinates of the system. The dynamical correction factor $f_{\mathrm{d},i\to j}$, on the other hand, emerges as the ratio:

$$f_{\mathrm{d},i\to j} = \frac{\langle \mathbf{n}_i(\mathbf{x}(0)) \cdot \dot{\mathbf{x}}(0) \delta\left[C_i(\mathbf{x}(0))\right] \left|\nabla C_i(\mathbf{x}(0))\right| h_j(\mathbf{x}(t)) \rangle}{\langle \mathbf{n}_i(\mathbf{x}(0)) \cdot \dot{\mathbf{x}}(0) \delta\left[C_i(\mathbf{x}(0))\right] \left|\nabla C_i(\mathbf{x}(0))\right| \left[1 - h_i(\mathbf{x}(0^+))\right] \rangle} \tag{14}$$

which can be simplified to

$$f_{\mathrm{d},i\to j} = \frac{\langle \mathbf{n}_i(\mathbf{x}(0)) \cdot \dot{\mathbf{x}}(0) \delta\left[C_i(\mathbf{x}(0))\right] \left|\nabla C_i(\mathbf{x}(0))\right| h_j(\mathbf{x}(t)) \rangle}{\frac{1}{2} \langle \left|\mathbf{n}_i(\mathbf{x}(0)) \cdot \dot{\mathbf{x}}(0)\right| \delta\left[C_i(\mathbf{x}(0))\right] \left|\nabla C_i(\mathbf{x}(0))\right| \rangle} \tag{15}$$

The numerator in Eqs. 14 and 15 for $f_{\mathrm{d},i\to j}$ is an average over all dynamical trajectories crossing the boundary of state i which ultimately thermalize in state j. The denominator in Eq. 14 is an average over all dynamical trajectories crossing the boundary surface of state i in an outward direction. The factor $1/2$ and the absolute value of the component

of velocity along the normal to the boundary surface in Eq. 15 stem from the fact that the latter component is symmetrically distributed around zero. Trajectories initiated on the boundary surface thermalize in a destination state within a correlation time that is much smaller than $(k_{i\rightarrow}^{\mathrm{TST}})^{-1}$ and therefore their sampling entails modest computational cost. A simple sampling scheme for implementing Eq. 15 is discussed in Ref. 6.

Interestingly, in this multistate formulation for the calculation of rate constants, due to Voter and Doll[6], transition state theory is applied to the total efflux from origin state i (see Eq. 13). The destination state j enters only through the dynamical correction factor $f_{\mathrm{d},i\rightarrow j}$, computed from short dynamical trajectories initiated on the dividing surface, via Eqs. 14 or 15. For adjacent states i and j that share parts of their boundary surfaces, $f_{\mathrm{d},i\rightarrow j}$ starts off high (equal to the Boltzmann-weighted fraction of the boundary surface of i that is shared with j) and quickly decays with time to an asymptotic value due to dynamical recrossing and fast correlated multistate jumps. For nonadjacent states i, j the dynamical correction factor $f_{\mathrm{d},i\rightarrow j}$ starts off at 0 and quickly rises to an asymptotic value. This describes transitions where the system crosses the boundary surface of i, spends a short time in one or more intermediate states without thermalizing in them, then enters j, which is nonadjacent to i, and ultimately thermalizes there. Such events are referred to as fast correlated multistate jumps.

The transition-state theory expression for the rate constant for exiting state i, $k_{i\rightarrow}^{\mathrm{TST}}$, Eq. 13, emerges as the product of half the mean absolute value of a component of the (mass-weighted) velocity along one direction in configuration space times a ratio of two configurational integrals: one taken over the boundary surface of the origin state i, and another one taken over the entire state i. Clearly, the ratio of configurational integrals has the physical meaning of a conditional probability that the system will find itself on the boundary surface, *provided* it is allowed to sample state i according to its equilibrium distribution. Instead of configurational integrals, one may consider the *partition function* Q_i of the system confined in the origin state i, as an integral over f-dimensional configuration space within state i and over f-dimensional momentum space; and the partition function $Q_i^{\dagger}$ of the system confined to the boundary surface of state i, as an integral over the $f-1$ dimensions of that surface in configuration space and over the $f-1$ dimensions of momentum space corresponding to moving within the surface, but not normal to it. Then, the expression for $k_{i\rightarrow}$ can be rewritten as

$$k_{i\rightarrow}^{\mathrm{TST}} = \frac{k_{\mathrm{B}}T}{h}\frac{Q_i^{\dagger}}{Q_i} \tag{16}$$

where the factor h takes care of the different dimensionalities of the phase spaces to which the two partition functions refer. Eq. 16 is applicable beyond the classical analysis adopted here, in systems where quantum mechanical effects are important. For a system under constant pressure, where volume fluctuations are important in effecting transitions out of state i, Q_i and $Q_i^{\dagger}$ must be interpreted as isothermal-isobaric partition functions. Recalling the connection between Gibbs energy and isothermal-isobaric partition function, Eq. 16 can be recast in the form

$$k_{i\rightarrow}^{\mathrm{TST}} = \frac{k_{\mathrm{B}}T}{h}\exp\left[-\left(\frac{G_i^{\dagger}-G_i}{k_{\mathrm{B}}T}\right)\right] \tag{17}$$

An example application of Eqs. 13 and 15 to the calculation of dynamically corrected

rate constants can be found in Ref. 13. There, elementary transitions of Xe and SF_6 in the pores of the zeolite Silicalite-1 were analyzed with the purpose of computing the self-diffusivity of these molecules at low occupancy. An inflexible model was invoked for the zeolite, allowing all calculations to be carried out in three dimensions ($f = 3$). States and boundary surfaces were mapped out explicitly as sets of voxels and pixels, respectively, after discretization of the intracrystalline space in the zeolite (see Sec. 2). The configurational integrals in Eq. 13 were computed by Monte Carlo integration in these voxels and pixels.

When state i is surrounded by high potential energy ridges relative to $k_\mathrm{B}T$ all along its boundary surface, transitions between nonadjacent states are improbable. A transition state estimate between adjacent states i and j can be obtained by analogy to Eqs. 13 and 17 as

$$k_{i\to j}^{\mathrm{TST}} = \frac{1}{(2\beta\pi)^{1/2}} \frac{\displaystyle\int_{\text{sep. surf. between states } i \text{ and } j} d^{f-1}x \exp\left[-\beta\mathcal{V}(\mathbf{x})\right]}{\displaystyle\int_{\text{state } i} d^{f}x \exp\left[-\beta\mathcal{V}(\mathbf{x})\right]} \tag{18}$$

$$k_{i\to j}^{\mathrm{TST}} = \frac{k_\mathrm{B}T}{h} \exp\left[-\left(\frac{G_{ij}^{\dagger} - G_i}{k_\mathrm{B}T}\right)\right] \tag{19}$$

In Eq. 18, the configurational integral in the numerator is taken over the part of the boundary surface of i that is common with the boundary surface of j, which we will call the separating surface between i and j. In Eq. 19, $G_{ij}^{\dagger}$ symbolizes the Gibbs energy of the system confined to that separating surface.

In many solid-state problems, transition between i and j is possible only through a narrow passage in the dividing surface, surrounding the first-order saddle point $(\mathbf{x}_{ij}^{\dagger}, \boldsymbol{\epsilon}_{ij}^{\dagger})$ between the configurations $(\mathbf{x}_i, \boldsymbol{\epsilon}_i)$ and $(\mathbf{x}_j, \boldsymbol{\epsilon}_j)$ of the two local energy minima, the energy being too high outside this narrow passage. Here $\boldsymbol{\epsilon}$ symbolizes the strain tensor with respect to a reference spatial extent of the system, usually taken as that characterizing the origin state i. Under given applied stress tensor $\boldsymbol{\sigma}$, this strain tensor may well be different between the origin state, the destination state, and the saddle point. When all the probability flux of the transition is directed through such a narrow, high-energy passage, for the purpose of computing the configurational integrals appearing in Eq. 18 one can invoke a quasiharmonic approximation, i.e. replace the potential energy with its Taylor expansion to second order with respect to $\mathbf{x}$ around a stationary point (saddle point for the numerator, minimum for the denominator) under the current volume of the system. The Gibbs energies in Eq. 19 are then estimated as

$$G_i \simeq \mathcal{V}_i + A_i^{\mathrm{vib}} - V_i\boldsymbol{\sigma}{:}\boldsymbol{\epsilon}_i \tag{20}$$

$$G_{ij}^{\dagger} = \mathcal{V}_{ij}^{\dagger} + A_{ij}^{\dagger\mathrm{vib}} - V_i\boldsymbol{\sigma}{:}\boldsymbol{\epsilon}_{ij}^{\dagger} \tag{21}$$

Here $\mathcal{V}_i$ is the potential energy at the minimum corresponding to state i and $\mathcal{V}_{ij}^{\dagger}$ is the potential energy at the saddle point corresponding to the transition state. V_i is the volume

at the reference configuration used for measuring strain, usually taken as that of the origin state i, $\boldsymbol{\epsilon}_i$ is the strain tensor at the origin state and $\boldsymbol{\epsilon}_{ij}^{\dagger}$ is the strain tensor at the saddle point. A_i^{vib} is a vibrational Helmholtz energy calculated from the angular frequencies $\omega_i^{(l)}$ of the normal modes of the system at the energy minimum of the origin state, while $A_{ij}^{\dagger\mathrm{vib}}$ is a vibrational Helmholtz energy calculated from the angular frequencies of the normal modes $\omega_{ij}^{\dagger(l)}$ at the saddle point:

$$A_i^{\mathrm{vib}} = -k_{\mathrm{B}}T \ln \left[\prod_{l=1}^{f} \frac{\exp(-\hbar\omega_i^{(l)}/(k_{\mathrm{B}}T))}{1-\exp(-\hbar\omega_i^{(l)}/(k_{\mathrm{B}}T))} \right] \tag{22}$$

$$A_{ij}^{\dagger\mathrm{vib}} = -k_{\mathrm{B}}T \ln \left[\prod_{l=1}^{f-1} \frac{\exp(-\hbar\omega_{ij}^{\dagger(l)}/(k_{\mathrm{B}}T))}{1-\exp(-\hbar\omega_{ij}^{\dagger(l)}/(k_{\mathrm{B}}T))} \right] \tag{23}$$

The spatial extent of the system at the minimum corresponding to the origin state is set based on the condition that G_i, as defined in Eq. 20, have a minimum with respect to the system dimensions under the applied stress $\boldsymbol{\sigma}$. Similarly, the spatial extent of the system at the saddle point is set based on the condition that $G_{ij}^{\dagger}$, as defined in Eq. 21, have a minimum with respect to the system dimensions under the applied stress $\boldsymbol{\sigma}$[20]. Kopsias[20] and Boulougouris[21] have invoked the quasiharmonic approximation approach to compute rate constants for elementary transitions in configuration space corresponding to structural relaxation of a Lennard-Jones and of an atactic polystyrene glass.

When all normal mode angular frequencies are very low relative to $k_{\mathrm{B}}T/\hbar$ and volume changes are negligible between the origin state and the transition state, the expression for the rate constant obtained from Eqs. 19 - 23 reduces to

$$k_{i\to j}^{\mathrm{TST}} = \frac{1}{2\pi} \frac{\prod_{l=1}^{f} \omega_i^{(l)}}{\prod_{l=1}^{f-1} \omega_{ij}^{\dagger(l)}} \exp\left[-\frac{\mathcal{V}_{ij}^{\dagger} - \mathcal{V}_i}{k_{\mathrm{B}}T} \right] \tag{24}$$

Eq. 24 has been proposed originally by Vineyard[1] in connection with the elementary jumps executed by an isotopic atom in the course of its self-diffusion in a solid lattice.

As pointed out in Sec. 2, in many problems it suffices to define states, transition paths, and dividing surfaces in the space of a few, slowly evolving degrees of freedom (coarse-grained variables or "order parameters"), rather than in the full $3N-3$-dimensional configuration space of the model system (assumed here to be characterized by periodic boundary conditions). In these cases, the transition-state theory estimate of the rate constant $k_{i\to j}^{\mathrm{TST}}$ is obtainable from Eq. 18 with f being a small number, $\mathbf{x}$ being the vector of (mass-weighted) coarse-grained variables and $\mathcal{V}$ being a potential of mean force with respect to these variables. For $f \leq 3$ it is feasible to map out this potential of mean force as a function of the coarse-grained variables. This provides a free energy profile (for $f = 1$) or landscape (for $f > 1$) that is useful for visualizing the transition.

In such lower-dimensional formulations, the configurational part of the Gibbs (or Helmholtz, in cases where volume changes are not important for the transition) energy

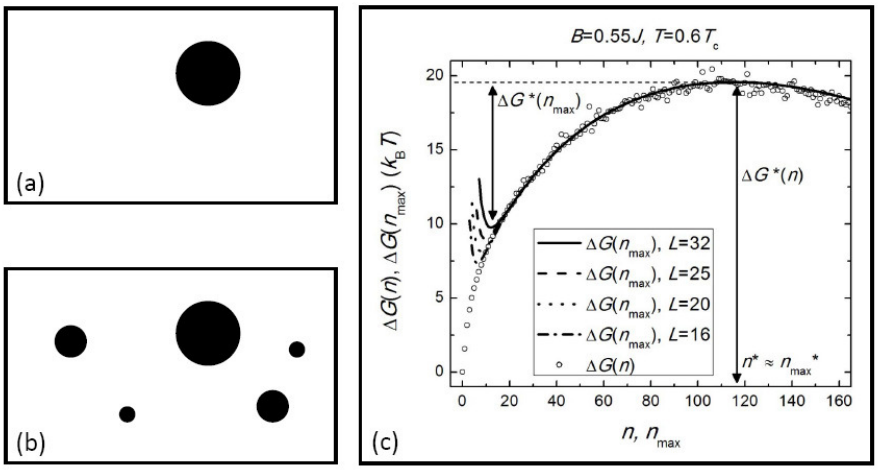


Figure 3. Gibbs energy profile for nucleation in a three-dimensional Ising model system consisting of $L \times L \times L$ sites arranged on a cubic lattice, as computed from umbrella sampling Monte Carlo simulations using spin inversions as the only moves. (a) A system containing only one nucleus. (b) A system containing multiple nuclei, the largest of which has size $n_{\max}$. (c) Gibbs energies $\Delta G(n)$ and $\Delta G(n_{\max})$ as functions of n and $n_{\max}$, respectively. The barrier heights encountered in these functions are indicated by an asterisk. See text for details.

difference $G_{ij}^{\dagger} - G_i$ appearing in Eq. 19 can be obtained through any statistical mechanics-based method designed for the computation of free energy differences. Free energy perturbation methods[31,24,32] offer themselves for this purpose. As the free energy barriers involved are typically large relative to $k_{\mathrm{B}}T$ (otherwise the phenomenon studied would not be an infrequent event), biased sampling techniques have to be invoked. A general strategy is umbrella sampling, wherein histograms of the relative free energy are accumulated through Boltzmann inversion of the probability density of coarse-grained variables within small overlapping windows in the space of coarse-grained variables, and different histograms are patched together to obtain the entire free energy landscape.

An example calculation of a Gibbs energy profile via umbrella sampling Monte Carlo simulation, based on work by K. Binder et al., is shown in Fig. 3. The model system is an Ising model with coupling constant J between neighboring spins, consisting of $L \times L \times L$ spins arranged on a simple cubic lattice in three dimensions. Initially, the system is in a phase with all spins "down" at a temperature of $T = 0.6T_c$, lower than the critical temperature $T_c \simeq 4.51J/k_{\mathrm{B}}$ for order- disorder transition. Then, a magnetic field $B = 0.55J$ is applied, rendering the initial phase metastable with respect to its counterpart with all spins "up". A first order phase transition ensues, which takes place via a nucleation and growth mechanism. Nuclei appear in the initial phase, each nucleus consisting of a cluster of "up" spins connected through nearest neighbor interactions. The Gibbs energy $\Delta G(n)$ for the formation of a nucleus of size (number of spins) n was accumulated by Boltzmann inversion of the size distribution of the nuclei. In addition, the Gibbs energy $\Delta G(n_{\max})$ for the largest nucleus in the system to be of size $n_{\max}$ was accumulated. The two functions are shown in Fig. 3. $\Delta G(n)$ is system-size independent, while the

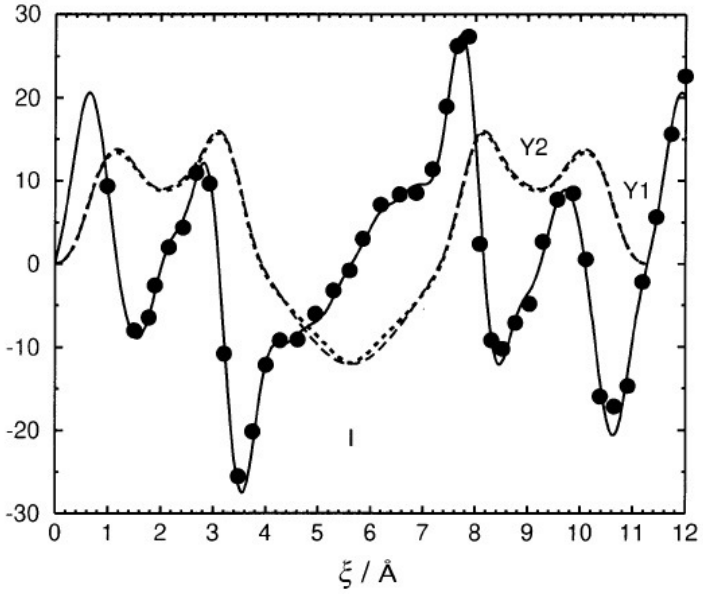


Figure 4. Configurational Helmholtz energy (potential of mean force) profile for a sorbed benzene molecule along the straight channel of the zeolite Silicalite-1, as computed by blue moon ensemble MD simulations, shown as a broken line[23]. The potential energy is measured in kJ mol^{-1}. The horizontal axis (reaction coordinate) measures the position of the center of mass of the benzene molecule projected along the axis of the channel, in Å, at room temperature. This calculation is based on a flexible model, incorporating the vibrational degrees of freedom of the zeolite. The global minimum near the center of the graph corresponds to the molecule residing within an intersection of the straight channel with a zigzag channel. Shallower minima are observed in the interior of the straight channel segments. Note that barriers in the potential of mean force are on the order of tens of kJ/mol, indicating that translational motion along the channel will proceed as a sequence of infrequent jump events. The continuous line with the points displays the derivative of the potential of mean force with respect to the reaction coordinate. This is the force needed to hold the system at a specific value of the reaction coordinate, computed via the blue moon ensemble method. The Helmholtz energy profile was obtained via numerical integration of this force.

barrier in $\Delta G(n_{\max})$ is reduced with increasing system size and would be expected to become very small for very large systems. This means that the new phase would nucleate very fast in a very large system. The barriers $\Delta G^*(n)$ and $\Delta G^*(n_{\max})$ are related via $\Delta G^*(n_{\max}) = \Delta G^*(n) - k_{\mathrm{B}}T\ln(L^3)$[33].

A related strategy is blue moon ensemble simulation, invoked by Forester and Smith[23] in their calculations of diffusion of benzene in the zeolite silicalite-1, as mentioned in Sec. 2 (see Fig. 4).

In recent years, a variety of advanced methods have been proposed for calculating free energy profiles along a coarse-grained variable or reaction coordinate. One such method is flux-tempered metadynamics[34], based on the metadynamics method introduced by Laio and Parrinello[35]. Metadynamics entails molecular dynamics simulation in which a repulsive Gaussian potential in a few selected coarse-grained variables is periodically added to the potential energy function of a system, to encourage its escape from the vicinity of local free energy minima with respect to these coarse-grained variables. If uniform sampling of

the space of coarse-grained variables is achieved, the free energy can be estimated from the sum of added Gaussian potentials.

A general, essentially exact, but computationally intensive method for computing transition rate constants between two known states in configuration space is transition path sampling, developed by Chandler and collaborators. This method is particularly useful in complex fluid systems, where the variables participating in the reaction coordinate are difficult to anticipate. The method samples dynamical trajectories connecting the two states. These trajectories are generated and manipulated using importance sampling techniques. We will not dwell on this method here, as detailed information can be found in a number of excellent reviews[36,37].

4 Kinetic Monte Carlo Simulation

We now turn to the question of how to track the temporal evolution of a system evolving through a sequence of infrequent events, once we know the states i, the transitions between them, and the interstate rate constants $k_{i\to j}$.

A widely used strategy is to generate a large number of stochastic trajectories of the system, conforming to the master Eq. 1. Each trajectory consists of a sequence of transitions between states. The transitions take place at times which are chosen by generation of pseudorandom numbers. The method is known as Kinetic Monte Carlo (KMC) simulation. The earliest application of KMC is thought to be Beeler's 1966 simulation of radiation damage annealing, although the term "kinetic Monte Carlo" was not widely adopted before 1990[38].

The usual implementation of KMC relies on the following properties of Poisson processes:

- If a number of Poisson processes occur in parallel in the same system with rate constants k_i, they comprise a Poisson process with rate constant $k = \sum_i k_i$.

- The waiting time of a Poisson process with rate constant k is exponentially distributed, with mean k^{-1} (see Eq. 8 and associated discussion).

- If ξ is a continuous random variable that is uniformly distributed in [0,1), then the random variable $\Delta t = -\ln(1-\xi)/k$ follows the exponential distribution with probability density $\hat{\rho}(\Delta t) = k\exp(-k\Delta t)$.

To begin the KMC simulation, a large number $\mathcal{N} >> n$ of independent walkers are deployed among the states of the system, according to a prescribed initial probability distribution among states, $P_i(0), i = 1, 2, \ldots, n$. For a system in equilibrium, $P_i(0) = P_i(\infty)$. (An easy way to generate a sample of a discrete or continuous random variable with prescribed probability distribution is to sample uniformly distributed pseudorandom values $\in [0, 1)$ for the cumulative distribution function and then find the inverse of this function at each of the sampled values. The prescription given above for sampling an exponentially distributed variable relies on the same principle.) We will use the symbol $\mathcal{N}_i(t)$ to denote the number of walkers that find themselves in state i at time t. Initially, $\mathcal{N}_i(0)/\mathcal{N} \simeq P_i(0)$. After initialization ($t = 0$), the KMC simulation proceeds according to the following steps:

(i) For each state i that is occupied at the current time t, calculate the expected fluxes $R_{i\to j}(t) = \mathcal{N}_i(t)k_{i\to j}$ to all states j to which state i is connected. Also, compute the overall flux $R(t) = \sum_i \sum_j R_{i\to j}(t)$ and the probabilities $q_{i\to j}(t) = R_{i\to j}(t)/R(t)$.

(ii) Generate a uniformly distributed pseudorandom number[39] $\xi \in [0, 1)$. Choose the time for occurrence of the next transition in the network of states as $\Delta t = -\ln(1-\xi)/R(t)$. Choose the type of the next transition by picking one of the possible transitions $i \to j$ according to the probabilities $q_{i\to j}(t)$.

(iii) Of the $\mathcal{N}_i(t)$ walkers present in state i, pick one with probability $1/\mathcal{N}_i(t)$ and move it to state j.

(iv) Advance the simulation time by Δt. Update the array, keeping track of the current positions of all walkers to reflect the implemented transition. Update the occupancy numbers $\mathcal{N}_i(t+\Delta t) = \mathcal{N}_i(t) - 1$ and $\mathcal{N}_j(t+\Delta t) = \mathcal{N}_j(t) + 1$.

(v) Return to step (i) to implement the next transition.

The outcome from performing this stochastic simulation over a large number of steps is a set of trajectories for all $\mathcal{N}$ walkers. Each trajectory consists of a long sequence of transitions between states of the network. Time-dependent system properties are estimated as ensemble averages over all trajectories at specific times. For example, if states correspond to sites in a three-dimensional network where a molecule can reside, one can calculate the mean square displacement along each one of the three coordinate directions as a function of time by averaging over the trajectories, and hence obtain the self-diffusivity tensor via the Einstein relation[13,14].

When all rate constants $k_{i\to j}$ are small, KMC will take large strides Δt on the time axis. Thus, times on the order of milliseconds, seconds, or even hours can be accessed, which are prohibitive for "brute force" MD.

5 Analytical Solution of the Master Equation

When the rate constants $k_{i\to j}$ are very broadly distributed, KMC simulation may become inefficient. This is because time steps Δt must be short enough to track the fastest processes occurring in the system. With such a short Δt, processes whose rate constants are several orders of magnitude lower than those of the fastest processes can hardly be sampled. Thus, one is faced with the same long-time problem as in MD.

In such cases of great dynamical heterogeneity, it may be better to resort to a direct solution of the master equation, Eq. 1, for the time-dependent state probabilities $\{P_i(t)\}$, under prescribed initial conditions $\{P_i(0)\}$. Remarkably, this solution can be developed analytically, as discussed in Wei and Prater's classic work on the kinetics of a network of reversible chemical reactions[8], and as detailed in recent work by Buchete and Hummer[40] and by Boulougouris[41]. We briefly outline this mathematical development here.

We start from the master equation in its matrix form, as written in Eq. 1. We transform the state probability vector $\mathbf{P}(t)$ into a reduced state probability vector $\tilde{\mathbf{P}}(t)$ with elements

$$\tilde{P}_i(t) = P_i(t)/\sqrt{P_i(\infty)} \tag{25}$$

$\tilde{\mathbf{P}}(t)$ satisfies the reduced master equation

$$\frac{\partial \tilde{\mathbf{P}}(t)}{\partial t} = \tilde{\mathbf{K}}\tilde{\mathbf{P}}(t) \tag{26}$$

with $\tilde{K}_{ij} = K_{ij}\sqrt{P_j(\infty)}/\sqrt{P_i(\infty)}$. The matrix $\tilde{\mathbf{K}}$ is symmetric by virtue of the microscopic reversibility condition, Eq. 2. One can readily show that $\tilde{\mathbf{K}}$ has the same eigenvalues as $\mathbf{K}$. These eigenvalues are real, since $\tilde{\mathbf{K}}$ is symmetric. Of these eigenvalues, one (corresponding to the establishment of the equilibrium distribution among states) is zero, and the remaining are negative. This is because $\tilde{\mathbf{K}}$ is a negative semidefinite matrix. The latter statement can be proved as follows: Let $\mathbf{y}$ be an arbitrary $n-$ dimensional vector of real elements. Then,

$$\begin{aligned}
\mathbf{y}^{\mathrm{T}} \cdot \tilde{\mathbf{K}} \cdot \mathbf{y} &= \sum_{i=1}^{n}\sum_{j=1}^{n} \tilde{K}_{ij} y_i y_j = \sum_{i=1}^{n} \tilde{K}_{ii} y_i^2 + \sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} \tilde{K}_{ij} y_i y_j \\
&= \sum_{i=1}^{n} \left(-\sum_{\substack{j=1 \\ j\neq i}}^{n} k_{i\to j} \right) y_i^2 + \sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} k_{j\to i} \left(\frac{P_j^{\mathrm{eq}}}{P_i^{\mathrm{eq}}} \right)^{1/2} y_i y_j \\
&= -\sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} k_{i\to j} y_i^2 + \sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} k_{i\to j} \left(\frac{P_i^{\mathrm{eq}}}{P_j^{\mathrm{eq}}} \right)^{1/2} y_i y_j \\
&= -\frac{1}{2}\sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} k_{i\to j} P_i^{\mathrm{eq}} \left[\frac{y_i}{(P_j^{\mathrm{eq}})^{1/2}} - \frac{y_j}{(P_i^{\mathrm{eq}})^{1/2}} \right]^2 \leq 0
\end{aligned} \tag{27}$$

Eq. 27 establishes $\tilde{\mathbf{K}}$ as a negative semidefinite matrix. The proof seems to have been given for the first time by Shuler[42]. Now, if λ is one of the real eigenvalues of $\tilde{\mathbf{K}}$ with corresponding real eigenvector $\tilde{\mathbf{u}}$, then $\tilde{\mathbf{K}} \cdot \tilde{\mathbf{u}} = \lambda\tilde{\mathbf{u}}$ and therefore $\tilde{\mathbf{u}}^{\mathrm{T}} \cdot \tilde{\mathbf{K}} \cdot \tilde{\mathbf{u}} = \lambda|\tilde{\mathbf{u}}|^2$. Because $\tilde{\mathbf{K}}$ is negative semidefinite, the left-hand side of the latter equation is negative or zero, hence $\lambda \leq 0$.

Let us denote the eigenvalues of $\tilde{\mathbf{K}}$ by $\lambda_0 = 0 \geq \lambda_1 \geq \ldots \geq \lambda_{n-1}$. We symbolize by $\tilde{\mathbf{u}}_m = (\tilde{u}_{1,m}, \tilde{u}_{2,m}, \ldots, \tilde{u}_{i,m}, \ldots \tilde{u}_{n,m})$ the eigenvector of $\tilde{\mathbf{K}}$ corresponding to eigenvalue $\lambda_m, 0 \leq m \leq n-1$. The eigenvector $\tilde{\mathbf{u}}_0$ has elements $\tilde{u}_{i,0} = \tilde{P}_i(\infty) = \sqrt{P_i(\infty)}$, corresponding to the equilibrium distribution among states. The Euclidean norm of $\tilde{\mathbf{u}}_0$ is unity by the normalization of $P_i(\infty)$.

The solution to the reduced master equation can be written as:

$$\tilde{\mathbf{P}}(t) = \sum_{m=0}^{n-1} \left[\tilde{\mathbf{u}}_m \cdot \tilde{\mathbf{P}}(0) \right] \exp(\lambda_m t) \tilde{\mathbf{u}}_m = \tilde{\mathbf{P}}(\infty) + \sum_{m=1}^{n-1} \left[\tilde{\mathbf{u}}_m \cdot \tilde{\mathbf{P}}(0) \right] \exp(\lambda_m t) \tilde{\mathbf{u}}_m \tag{28}$$

where the normalization condition $\sum_{j=1}^{n} P_j(0) = 1$ has been used in separating out the equilibrium contribution ($\lambda_0 = 0$). The eigenvectors $\tilde{\mathbf{u}}_m$ form an orthonormal basis set:

$$\tilde{\mathbf{u}}_m \cdot \tilde{\mathbf{u}}_l = \delta_{ml}, \quad 0 \leq m, l \leq n-1 \tag{29}$$

They also satisfy $\sum_{m=0}^{n-1} \tilde{u}_{i,m}\tilde{u}_{j,m} = \delta_{ij}$. Once $\tilde{\mathbf{P}}(t)$ has been determined, the state probabilities $\mathbf{P}(t)$ can be calculated via $P_i(t) = \tilde{P}_i(t)\sqrt{P_i(\infty)}$.

Eq. 28 has an interesting geometric interpretation, which is discussed at length by Boulougouris in the context of a general formulation for analytical solution of the master equation and calculation of time-dependent averages and autocorrelation functions which was dubbed "EROPHILE," for "Eigenvalue Representation of Observables and Probabilities in a HIgh-Dimensional Euclidean space"[41]. In the n-dimensional Euclidean space spanned by the reduced state probabilities $\tilde{P}_i$, the point $\tilde{\mathbf{P}}(t)$ moves in a hyperplane that is normal to the eigenvector $\tilde{\mathbf{u}}_0 = \left(\sqrt{P_1(\infty)}, \sqrt{P_2(\infty)}, \ldots, \sqrt{P_n(\infty)}\right)$ and contains point $\tilde{\mathbf{P}}(0)$. This plane is, of course, spanned by the remaining eigenvectors $\tilde{\mathbf{u}}_1, \tilde{\mathbf{u}}_2, \ldots, \tilde{\mathbf{u}}_{n-1}$. It intersects each of the $\tilde{P}_i$ axes at $1/\sqrt{P_i(\infty)}$. As time goes by, $\tilde{\mathbf{P}}(t)$ traces a curved trajectory on this hyperplane from $\tilde{\mathbf{P}}(0)$ to the equilibrium distribution $\tilde{\mathbf{P}}(\infty)$.

Let us consider any observable, $\mathcal{A}$, which has well-defined values $\mathcal{A}_i$ within each of the states i. The (nonequilibrium) ensemble average $\langle \mathcal{A}(t) \rangle$ at any time t is

$$\langle \mathcal{A}(t) \rangle = \sum_{i=1}^{n} P_i(t)\mathcal{A}_i = \langle \mathcal{A}(\infty) \rangle + \sum_{m=1}^{n-1} a_m \beta_m \exp(\lambda_m t) \tag{30}$$

where

$$a_m = \tilde{\mathbf{u}}_m \cdot \tilde{\mathbf{P}}(0) \tag{31}$$

and

$$\beta_m = \tilde{\mathbf{u}}_m \cdot \tilde{\boldsymbol{\mathcal{A}}} \tag{32}$$

In Eq. 32, $\tilde{\boldsymbol{\mathcal{A}}}$ is an n-dimensional vector with elements $\tilde{\mathcal{A}}_i = \mathcal{A}_i\sqrt{P_i(\infty)}$, formed from the values $\mathcal{A}_i$ of the observable in each state and the equilibrium probabilities $P_i(\infty)$ of the states.

Eq. 30 expresses the time-dependent ensemble average of the observable, $\langle \mathcal{A}(t) \rangle$, as a sum of its value $\langle \mathcal{A}(\infty) \rangle$ when equilibrium among all n states has been established, plus a sum of exponentially decaying functions. The sum is taken over all relaxation modes, with characteristic time constants $-1/\lambda_1 \geq -1/\lambda_2 \geq \ldots \geq -1/\lambda_{n-1}$.

In the space spanned by $\tilde{P}_i$, considered above, one can draw the vector $\tilde{\boldsymbol{\mathcal{A}}}$ with components $\tilde{\mathcal{A}}_i = \mathcal{A}_i\sqrt{P_i(\infty)}$ along each $\tilde{P}_i$ axis. The equilibrium average $\langle \mathcal{A}(\infty) \rangle$ is interpreted geometrically as the projection of this vector on the eigenvector $\tilde{\mathbf{u}}_0 = \tilde{\mathbf{P}}(\infty)$. The time-dependent average $\langle \mathcal{A}(t) \rangle$, on the other hand, is interpreted as a projection of the same vector on the reduced probability vector $\tilde{\mathbf{P}}(t)$. As the tip of $\tilde{\mathbf{P}}(t)$ moves from $\tilde{\mathbf{P}}(0)$ toward the equilibrium point $\tilde{\mathbf{P}}(\infty)$, $\langle \mathcal{A}(t) \rangle$ moves to $\langle \mathcal{A}(\infty) \rangle$. The exponentially decaying components of $\langle \mathcal{A} \rangle$ along the modes are proportional to the projections β_m of $\tilde{\boldsymbol{\mathcal{A}}}$ on the eigenvectors[41].

One can readily express time autocorrelation functions for observables using the analytical solution to the reduced master equation. For any observable $\mathcal{A}$ defined in the states,

$$\langle \mathcal{A}(0)\mathcal{A}(t) \rangle - \langle \mathcal{A}(0) \rangle \langle \mathcal{A}(\infty) \rangle = \sum_{m=1}^{n-1} \beta_m^2 \exp(\lambda_m t) + \langle \mathcal{A}(\infty) \rangle \sum_{m=1}^{n-1} a_m \beta_m \exp(\lambda_m t) + \sum_{m=1}^{n-1} \beta_m \exp(\lambda_m t) \sum_{l=1}^{n-1} \beta_l \sum_{k=1}^{n-1} \sum_{i=1}^{n} \left[\frac{a_k \tilde{u}_{i,l} \tilde{u}_{i,m} \tilde{u}_{i,k}}{\tilde{P}_i(\infty)} \right] \tag{33}$$

In the special case where the system is initially distributed among states according to equilibrium, $\tilde{\mathbf{P}}(0) = \tilde{\mathbf{P}}(\infty)$, by virtue of the orthonormality of eigenvectors we have $a_m = 0, m = 1, 2, \ldots, n-1$ and Eq. 33 simplifies to

$$\langle \mathcal{A}(0)\mathcal{A}(t)\rangle - \langle \mathcal{A}(\infty)\rangle^2 = \sum_{m=1}^{n-1} \beta_m^2 \exp(\lambda_m t) \tag{34}$$

In this special case, $\left\langle (\delta\mathcal{A})^2 \right\rangle^{1/2} = \left[\langle \mathcal{A}(0)\mathcal{A}(t)\rangle - \langle \mathcal{A}(\infty)\rangle^2\right]^{1/2}$ can be interpreted geometrically as the length of the projection of vector $\tilde{\boldsymbol{\mathcal{A}}}$ on the on the $n-1$-dimensional hyperplane on which $\tilde{\mathbf{P}}(t)$ moves[41].

Implementation of this analytical solution scheme requires that the equilibrium state probabilities $P_i(\infty)$ be found at the beginning of the calculation. An easy strategy for accomplishing this without diagonalizing matrix $\mathbf{K}$ is to use the iterative successive substitution scheme[43]:

$$P_i^{(l+1)}(t) = \frac{\sum_{j\neq i} P_j^{(l)}(t) k_{j\to i}}{\sum_{j\neq i} k_{i\to j}} \tag{35}$$

Implementation of Eq. 28 requires diagonalization of the singular symmetric $n \times n$ matrix $\tilde{\mathbf{K}}$.

For spatially periodic systems, in which the set of states is obtainable by replication of a "unit cell" of states in one, two, or three dimensions, Kolokathis[44] has developed a method for calculating the eigenvalues and eigenvectors of the reduced rate constant matrix of the whole system by diagonalizing matrices of dimension corresponding to a single unit cell. This Master Equation Solution by Recursive Reduction of Dimensionality (MESoRReD) in diagonalizing the rate constant matrix method greatly reduces the computational effort required for diagonalization and is valuable in addressing problems of diffusion in crystalline solids.

6 Example: Diffusion of Xenon in Silicalite

Zeolites are crystalline aluminosilicates whose crystal structure is characterized by the presence of regular cavities and pores of diameter commensurate with the sizes of common gas or solvent molecules. This structure imparts to zeolites a unique ability to distinguish among molecules sorbed in their pores in terms of their size, shape, and charge distribution and forms the basis for a large number of technological applications of zeolites as industrial separation media, catalysts, and ion exchange agents.

Diffusivities in zeolites are commonly computed via MD simulations. In many systems of practical relevance, however, diffusion is too slow to be computed reliably by MD. For example, as can be seen in Fig. 4, benzene experiences a tight fit in the pores of silicalite-1, such that moving from an intersection region to the interior of a straight channel requires overcoming a free energy barrier of approximately 27 kJ/mol. A MD simulation of benzene sorbed at low occupancy in silicalite at room temperature would exhaust itself tracking local motions of the benzene within a sorption site and would hardly sample any jumps into other sorption sites, which contribute to translational diffusion. A reasonable prediction of

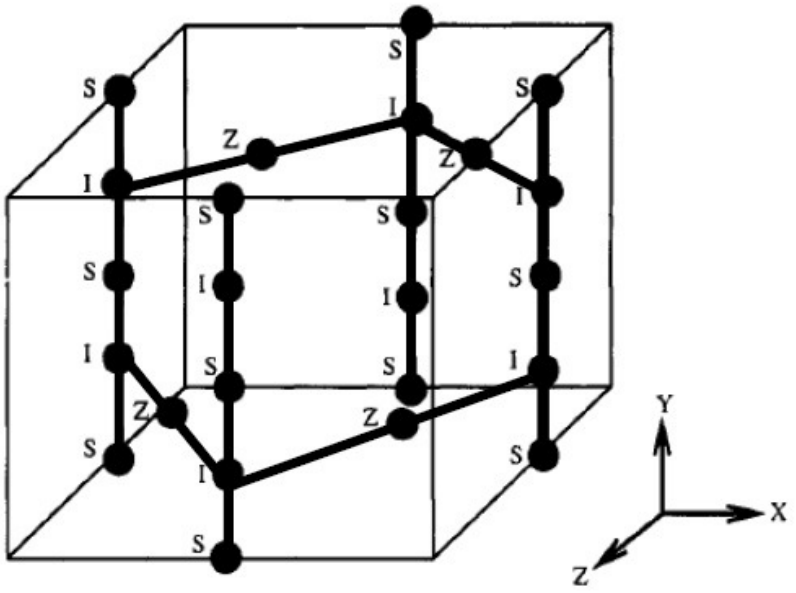


Figure 5. Schematic outline of the pore structure of a unit cell of silicalite. Spheres represent the three types of sorption states (Z = sinusoidal channel state, S = straight channel state, and I = intersection state) on the zeolite-sorbate potential hypersurface. The thick lines provide a rough depiction of the axes of straight and sinusoidal (zig-zag) channels.

the diffusivity can only be obtained through computation of rate constants for jumping from site to site via infrequent event analysis and solution of the master equation in the network of sorption sites (states)[14,23].

A simple sorbate/zeolite system on which infrequent event-based calculations appear to have been conducted for the first time is xenon (Xe) in silicalite-1 at low temperatures and occupancies. Here we review briefly some calculations on this system at 150 K, coming from the early work of June et al.[13] and the very recent work of Kolokathis[44]. The unit cell of silicalite has the chemical constitution $Si_{96}O_{192}$. Calculations were conducted with its orthorhombic form, which has lattice parameters $a = 20.07$ Å, $b = 19.92$ Å, $c = 13.42$ Å along the x, y, and z directions, respectively. The zeolite possesses two intersecting systems of channels, both of diameter around 5.5 Å: Straight channels, which run along the b crystallographic axis, and sinusoidal, or zig-zag, channels, which run along the a crystallographic axis. The channel systems come together at intersections, which are more spacious (diameter around 9 Å).

In the modeling work of June et al.[45,13], silicalite was considered as rigid and its interaction with Xe was described as a sum of Lennard-Jones potentials between each oxygen in its framework and the Xe molecule. An efficient potential pretabulation and interpolation scheme in three dimensions was developed for this potential in simulations[45]. June et al.[13] conducted a thorough analysis of the potential energy hypersurface experienced by

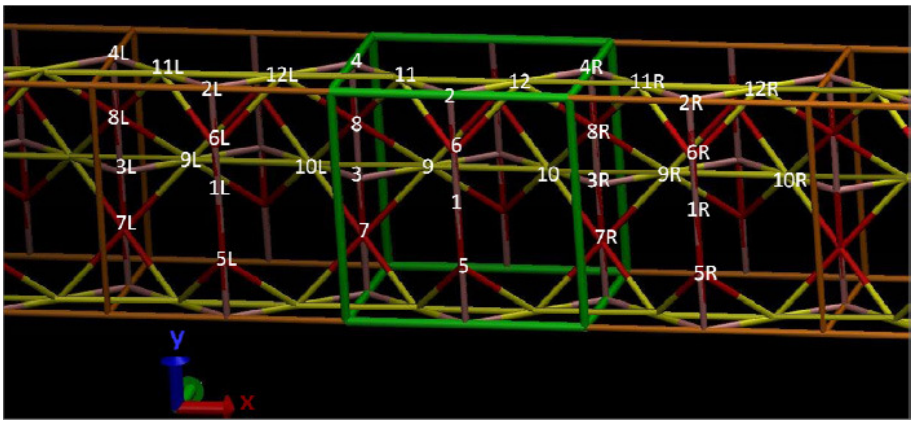


Figure 6. Transitions of Xe in silicalite-1, depicted as straight lines in three-dimensional space. Red color shows S (straight-channel) states, yellow color shows Z (zig-zag channel) states and pink color shows I (intersection) states. Green box defines the borders of one unit cell. Orange color shows the borders of cells along the x axis.

Xe in silicalite as a function of its three translational degrees of freedom, identified states and transitions between them, and computed rate constants $k_{i\to j}$ using Transition State Theory (TST) with or without dynamical corrections. This analysis led to the identification of 12 states per unit cell for Xe in silicalite at very low loadings. There are four states per unit cell in the interior of straight channel segments (S), four states per unit cell in the interior of zig-zag channel segments (Z) and four states per unit cell in intersections (I). Of these, Z and S states are more favorable, while I, where the dispersive attraction of Xe with the surrounding zeolite lattice is weaker, is less favorable. At 150 K the equilibrium probabilities of occupancy of these states, normalized within one fourth of the unit cell, are $P_Z^{\rm eq} = 0.572$, $P_S^{\rm eq} = 0.414$, $P_I^{\rm eq} = 0.014$. The spatial arrangement of these states within one unit cell of silicalite is shown in Fig. 5.

There is a rich connectivity among the states for Xe in silicalite. Apart from I to S and I to Z transitions, June et al.[13] identified direct transitions between S and Z states which circumvent the intersection regions. There are eleven distinct types of transitions. These types and their associated rate constants, as calculated by Transition State Theory without dynamical corrections [Eq. 18], are shown in Tab. 1.

Fig. 6 provides a pictorial depiction of the spatial arrangement of sorption states 1-12 in a central unit cell (outlined with green borders) and of the periodic images of these states located to the right (R) and left (L) of the central unit cell. States 1-4 are I states; states 5-8 are S states; and states 9-12 are Z states. Fig. 6 also shows the network of transitions as a set of straight line segments connecting the states. Each I, S, and Z state is connected to another 4, 6, and 8 states, respectively. There are 72 transition pathways lying within or crossing the boundaries of a unit cell, where forward and reverse pathways are counted separately. These transitions are summarized in the third column of Tab. 1; to each of these transitions a rate and a type are assigned in the first two columns of the same table.

Tab. 2 shows estimates of the diffusivities D_{xx}, D_{yy}, D_{zz}, as well as of the orientationally averaged diffusivity $D = (D_{xx} + D_{yy} + D_{zz})/3$ at 150 K, obtained from the states,

Type of Transition	Rate constant $k_{i\to j}$ (s^{-1})	Transitions
I $\to$ S	1.309×10^{11}	$1 \to 5, 1 \to 6, 2 \to 5, 2 \to 6,$ $3 \to 7, 3 \to 8, 4 \to 7, 4 \to 8$
S $\to$ I	4.444×10^{9}	$5 \to 1, 5 \to 2, 6 \to 1, 6 \to 2,$ $7 \to 3, 7 \to 4, 8 \to 3, 8 \to 4$
I $\underset{a}{\to}$ Z	2.958×10^{10}	$1 \to 9, 2 \to 12, 3 \underset{\text{L}}{\to} 10, 4 \to 11$
Z $\underset{a}{\to}$ I	7.241×10^{8}	$9 \to 1, 10 \underset{\text{R}}{\to} 3, 11 \to 4, 12 \to 2$
I $\underset{b}{\to}$ Z	1.501×10^{10}	$1 \to 10, 2 \to 11, 3 \to 9, 4 \underset{\text{L}}{\to} 12$
Z $\underset{b}{\to}$ I	3.673×10^{8}	$9 \to 3, 10 \to 1, 11 \to 2, 12 \underset{\text{R}}{\to} 4$
S $\underset{a}{\to}$ Z	3.974×10^{8}	$5 \to 9, 6 \to 9, 7 \underset{\text{L}}{\to} 10, 8 \underset{\text{L}}{\to} 10,$ $7 \to 11, 8 \to 11, 5 \to 12, 6 \to 12$
Z $\underset{a}{\to}$ S	2.853×10^{8}	$9 \to 5, 9 \to 6, 10 \underset{\text{R}}{\to} 7, 10 \underset{\text{R}}{\to} 8,$ $11 \to 7, 11 \to 8, 12 \to 5, 12 \to 6$
S $\underset{b}{\to}$ Z	8.567×10^{8}	$5 \to 10, 5 \to 11, 6 \to 10, 6 \to 11,$ $7 \to 9, 7 \underset{\text{L}}{\to} 12, 8 \to 9, 8 \underset{\text{L}}{\to} 12$
Z $\underset{b}{\to}$ S	6.150×10^{8}	$10 \to 5, 11 \to 5, 10 \to 6, 11 \to 6,$ $9 \to 7, 12 \underset{\text{R}}{\to} 7, 9 \to 8, 12 \underset{\text{R}}{\to} 8$
Z $\to$ Z	9.737×10^{8}	$9 \to 10, 9 \underset{\text{L}}{\to} 10, 10 \to 9, 10 \underset{\text{R}}{\to} 9,$ $11 \to 12, 11 \underset{\text{L}}{\to} 12, 12 \to 11, 12 \underset{\text{R}}{\to} 11$

Table 1. Rate constants[13] for interstate transitions of xenon in silicalite at 150 K as calculated from Transition-State Theory in tree dimensions, without dynamical corrections. I, S, and Z represent an intersection, straight channel state and sinusoidal channel state, respectively. The indices under the arrows distinguish between different transitions starting at the same origin state and ending at different images of the destination state.

connectivity, and rate constant information of Figs. 5, 6 and Tab. 1. No distinction is made between self- and transport diffusivities, as the two are equal at the very low occupancies considered here. Diffusivities have been calculated by three methods:

- Kinetic Monte Carlo simulation: Here one deploys a large number (e.g. 4000) of noninteracting Xe molecules ("walkers") among the states of a large (e.g. $10\times10\times10$ unit cells) network with periodic boundary conditions, according to the equilibrium occupancy probabilities of the states. One then generates a long (e.g., at least 27000 steps, corresponding to roughly 18 ns for the Xe/silicalite-1 system) KMC trajectory by the procedure discussed in Sec. 4. The diffusivity is calculated via the Einstein relation, e.g.

$$D_{xx} = \lim_{t\to\infty} \frac{\left\langle [x(t) - x(0)]^2 \right\rangle}{2t} \tag{36}$$

  and similarly for y and z. Tab. 2 presents KMC results from both the original work of June et al.[13] and the very recent calculations of Kolokathis[44]. The two sets of KMC are indistinguishable, within simulation error.

- Numerical solution of the master equation. Here, the master equation, Eq. 1, was solved numerically as an initial value problem with the Euler method to determine the state occupancy probabilities as functions of time. The calculation was performed on a system of $50 \times 50 \times 50$ unit cells with periodic boundary conditions. Initially, a probability of 1 was assigned to an S state at the center of the system, all other states being empty. The integration time step in the Euler method was 10^{-12} s. State probabilities from the numerical solution were summed at the level of unit cells and divided by the unit cell volume to obtain the probability density $\rho_{\mathrm{cell}}(x, y, z, t)$. The marginal probability densities along the three directions, $\rho_{\mathrm{cell},x}(x, t)$, $\rho_{\mathrm{cell},y}(y, t)$, $\rho_{\mathrm{cell},z}(z, t)$ were then calculated. The diffusivities D_{xx}, D_{yy}, D_{zz} were obtained by matching these time-dependent probability densities to the solution of the corresponding continuum diffusion problem. For the maximum time used in the Euler integration, 10 ns, this is indistinguishable from the Gaussian

$$\rho_{\mathrm{cell},x}(x, t) = \frac{1}{\sqrt{4\pi D_{xx} t}} \exp\left[-\frac{(x - x_0)^2}{4D_{xx}t}\right] \tag{37}$$

  and similarly for y and z.

- Analytical solution of the master equation. Model systems consisting of $2^7 = 128$ adjacent unit cells arranged in a linear array along the x, y, or z directions, with periodic boundary conditions at the ends, were considered. The symmetrized rate constant matrix $\tilde{\mathbf{K}}_{2^7}$ for each of these systems was formed and diagonalized. Initially, all probability was distributed in the central two unit cells of the array. The time-dependent probability of occupancy of all states in the system was calculated as a sum of exponentially decaying functions of time using the eigenvectors and eigenvalues of matrix $\tilde{\mathbf{K}}_{2^7}$, according to Eqs. 25 and 28. To avoid the time- consuming diagonalization of the 1536×1536-dimensional matrix $\tilde{\mathbf{K}}_{2^7}$, a recursive reduction scheme was devised[44], which ultimately expresses the eigenvalues and eigenvectors of $\tilde{\mathbf{K}}_{2^7}$ in terms of the eigenvalues and eigenvectors of the symmetrized rate constant

matrix for a single unit cell, $\tilde{\mathbf{K}}_1$ and other 12×12 matrices that can be formed readily from the set of rate constants. This MESoRReD scheme[44] affords great savings in CPU time. The calculation of diffusivities from the time-dependent state probability profiles is again accomplished by fitting the solution to the corresponding continuum diffusion equation to the master equation results.

Method	D_{xx} (m^2s^{-1})	D_{yy} (m^2s^{-1})	D_{zz} (m^2s^{-1})	D (m^2s^{-1})
June et al. MD[13]	4.3×10^{-10}	1.0×10^{-9}	0.99×10^{-10}	5.1×10^{-10}
June et al. KMC-DC TST[13]	5.1×10^{-10}	7.3×10^{-10}	0.83×10^{-10}	4.41×10^{-10}
June et al. KMC-TST[13]	1×10^{-9}	1.2×10^{-9}	1.7×10^{-10}	7.9×10^{-10}
KMC-TST[44]	9.75×10^{-10}	1.21×10^{-9}	1.71×10^{-10}	7.85×10^{-10}
Euler Method TST[44]	9.70×10^{-10}	1.25×10^{-9}	1.83×10^{-10}	8.01×10^{-10}
Master Eq. Soln. by Recursive Reduction of Dimensionality[44]	9.71×10^{-10}	1.17×10^{-9}	1.75×10^{-10}	7.71×10^{-10}
PFG-NMR[53,54]	-	-	-	1.633×10^{-10}

Table 2. Diffusion coefficients for xenon in silicalite-1 at 150 K as computed by different methods and as measured experimentally

As seen in Tab. 2, estimates of D_{xx}, D_{yy}, D_{zz} and D obtained by different TST-based methods are within 3% of each other. Estimates based on rate constants computed via *dynamically corrected* TST, i.e., using Eqs. 12, 13 and 15, obtained by June et al.[13] are also included in the table, for comparison. Consideration of dynamical corrections gives lower rate constants for interstate transitions (mainly due to recrossings of the dividing surfaces) and therefore lower diffusivities. Estimates from the dynamically corrected TST are very close to those obtained by direct MD simulation, which can be considered as the "exact results" for the force field employed. In Tab. 2 is also shown the single experimental value of the orientationally averaged self-diffusivity D available for Xe in silicalite-1 at 150 K via pulsed field gradient nuclear magnetic resonance (PFG-NMR) experiments using ^{129}Xe. The experimental value is of the same order as, but considerably lower than,

Method	CPU time (s)	Memory (MB)
June et al. MD[13]	309173.00	
Euler method[44]	23760.00	401
KMC[44]	6183.46	6
Master Eq. Soln. by Recursive Reduction of dimensionality[44]	2.96	42

Table 3. CPU time and memory (RAM) requirements for calculating the diffusivity of Xe in silicalite-1 at 150 K with a relative error of 3% by various methods. All times except that for MD were measured[44] on an Intel Celeron CPU E200 system with 1.99 GB RAM, running at 2.40 GHz. The time for MD is an estimate, based on the work of June et al.[13] The CPU time required for the method of analytical solution of the master equation by recursive reduction of dimensionality of the rate constant matrix is partitioned as follows: (a) Determination of the time-dependent state probabilities 0.27 s; (b) Determination of diffusivity by fitting the profile of state probabilities with the solution to the continuum diffusion equation 2.69 s.

the best simulation estimates from MD and DC-TST. This is partly due to the fact that the intracrystalline occupancy was finite in the experiments, rather than tending to zero, as considered in the simulations. Imperfections in the zeolite crystals employed in the experiment and in the force field employed in the simulations and the fact that the high-temperature, orthorhombic form of the crystal was used in the simulations at 150 K no doubt contribute to the difference between experimental and predicted values.

The computational requirements of MD and of the TST-based methods for computing the diffusivity of Xe in silicalite-1 to the same level of accuracy are compared in Tab. 3. Clearly, analytical solution of the master equation for a periodic model system, based on recursive reduction of the rate constant matrix, is the most efficient among the methods examined; its CPU time requirement is smaller than that of MD, numerical solution of the master equation by the Euler method, and Kinetic Monte Carlo by factors of 100000, 8000, and 2100, respectively. The widely practiced KMC comes next. For penetrants experiencing a close fit in zeolite pores, such as benzene in silicalite, MD is incapable of tracking diffusional progress and infrequent event-based methods remain as the only viable alternative[14,23].

7 Example: Diffusion of CO_2 in Poly(amide imide)

Knowing the diffusion coefficient of small (gas, solvent) molecules in glassy polymers is of great importance to the design of packaging materials with controlled barrier properties, as well as of separation membranes with tailored permeability and selectivity[19]. While the

problem of diffusion in molten and rubbery polymer matrices at temperatures sufficiently above the glass temperature T_{g} can be addressed successfully via MD simulation, diffusion in polymer glasses is too slow to be predictable by direct MD. The self-diffusivities of gases dissolved at low concentration in glassy polymers are typically on the order of $10^{-12}\mathrm{m}^2/\mathrm{s}$ and would require simulation times longer than μs in order to be predicted by MD from the mean square displacement $\left\langle [\mathbf{r}(t) - \mathbf{r}(0)]^2 \right\rangle$ via the Einstein relation:

$$D_{\mathrm{s}} = \lim_{t\to\infty} \frac{\left\langle [\mathbf{r}(t) - \mathbf{r}(0)]^2 \right\rangle}{6t} \tag{38}$$

The presence of an "anomalous diffusion" regime at short times, where $\left\langle [\mathbf{r}(t) - \mathbf{r}(0)]^2 \right\rangle$ rises sublinearly with time (see below) makes the reliable calculation of D_{s} even more demanding.

MD simulations have established that the diffusion of a small molecule in a glassy polymer takes place as a sequence of infrequent jumps between accessible volume clusters within the polymer. Thus, the problem of calculating the self-diffusivity in an amorphous glassy polymer is similar to that in a zeolite, with the following important differences: (a) Simulating the structure of the amorphous polymer is a challenge in itself, which has stimulated significant methodological development. Currently, a satisfactory strategy for generating glassy polymer configurations is to coarse-grain an atomistic model into one involving fewer degrees of freedom, equilibrate the coarse-grained model at all length scales using connectivity-altering Monte Carlo algorithms, reverse-map back to the atomistic level to obtain well-equilibrated melt configurations, and finally quench to the glassy state[55]. (b) Infrequent-event analyses of elementary jumps only in the penetrant degrees of freedom, assuming an inflexible polymer matrix, are of very limited utility; the motion of polymer degrees of freedom in the course of a diffusive jump must be taken into account in calculating rate constants for the elementary diffusive jumps in order to obtain a realistic estimate of D_{s}.

The first serious calculation of diffusivities in an amorphous polymer matrix based on TST concepts was performed by Gusev and Suter[56]. This calculation is based on the idea that atoms of the polymer matrix execute harmonic vibrations around their equilibrium positions in the minimum energy configuration of the penetrant-free polymer. For a spherical penetrant, this leads to a three-dimensional free energy field that can be expressed in terms of additive contributions depending on the distances of the center of the penetrant from the equilibrium positions of the polymer atoms. All (three-dimensional) states and (two-dimensional) dividing surfaces for translational motion of the penetrant in the polymer matrix are determined via steepest descent constructions in this free energy function, in a similar way as in rigid zeolite models (compare Sections 2 and 6) and transition rate constants for all elementary jumps were determined via Eq. 18 with $f = 3$ and the free energy field including vibrational contributions from polymer atoms playing the role of $\mathcal{V}(\mathbf{x})$. The amplitude of polymer atom vibrations, Δ, is usually treated as an adjustable parameter. A self-consistent method has been proposed for its determination from short-time MD simulations of the polymer matrix[57]. This is a useful and computationally efficient approach if the penetrant is small enough to justify the assumption of harmonic ("elastic") motion of matrix atoms.

Greenfield[15,58,59] developed a multidimensional TST approach for diffusion in a glassy

polymer, where polymer degrees of freedom are taken into account explicitly in the reaction coordinate of the infrequent events whereby diffusion takes place. For the identification of states and dividing surfaces, Greenfield introduced a method based on geometric analysis of accessible volume within penetrant-free minimum energy configurations of the glassy polymer, which has been outlined briefly in Sec. 2. This calculation goes from geometrically identified "necks" between accessible volume clusters to saddle points in the multidimensional configuration space of the penetrant plus polymer system, to transition paths in that configuration space. Each transition path connects two basins (regions around local minima) i and j in multidimensional configuration space, with the center of mass of the penetrant residing in one cluster of accessible volume in basin i and in another cluster of accessible volume in basin j. The rate constant $k_{i\to j}$ for the jump between i and j is calculated in the harmonic approximation via Eqs. 19 - 23 with the stress set to zero and volume changes neglected. In general, there are many basins corresponding to the penetrant residing in the same cluster of accessible volume as in basin i; these basins communicate with each other via facile transitions and are envisioned as constituting a "macrostate" or "metabasin" I. Similarly, basin j belongs to a larger "metabasin" J. The rate constant for transition between metabasins I and J is estimated as

$$k_{I\to J} = \sum_{i\in I}\sum_{j\in J} k_{i\to j}\,\frac{P_i(\infty)}{P_I(\infty)} \tag{39}$$

The ratio $P_i(\infty)/P_I(\infty)$ is estimated from a short MD simulation of the polymer plus penetrant system with the penetrant confined in the accessible volume of metabasin I; it is the ratio of time spent in basin i to that spent in the entire metabasin I. The rate constants $k_{I\to J}$ constitute a rate constant matrix $\mathbf{K}$ providing a stochastic description of the motion of the penetrant at the level of metabasins, or clusters of accessible volume. They may have to be adjusted to ensure that microscopic reversibility, Eq. 2, is satisfied.

Vergadou[60] extended and applied Greenfield's method to study permeation of CO_2 in a glassy poly(amide imide) of complex repeat unit constitution [-NH-C_6H_4-$C(CF_3)_2$-C_6H_4-NH-CO-$C_6H_4(CH_3)$-$N(CO)_2C_6H_3$-CO-]$_n$. All multidimensional TST calculations were performed in atomic Cartesian coordinates. The distribution of rate constants for elementary jumps $k_{i\to j}$ was found to be very broad, covering the range 10^{-14} to 10^{-1} s^{-1}, and skewed towards low values, the most probable value being around 10^{-6} s^{-1}. The distribution of elementary jump lengths of the penetrant, on the other hand, was found to be relatively narrow, covering the range 2 to 10 Å, with a most probable value around 4 Å. Fig. 7 displays three characteristic snapshots in the course of an elementary jump of a CO_2 molecule. The initial and final configurations constitute local minima of the potential energy of the polymer plus penetrant system, while the middle configuration (transition state) is a saddle point of the potential energy function. Molecular configurations are shown in part a of the figure, while part b displays the accessible volume distribution at these three characteristic points along the transition path of the elementary jump. Clearly, in the initial and final states the CO_2 molecule lies in the interior of accessible volume clusters formed among the atoms of the glassy polymer. In the transition state a "neck" of accessible volume has developed which momentarily connects the origin and destination clusters, letting the penetrant go through. At the transition state the penetrant is oriented roughly parallel to this neck. Evidently, the degrees of freedom of the polymer and the orientational degrees of freedom of the penetrant play a significant role in shaping the transition path

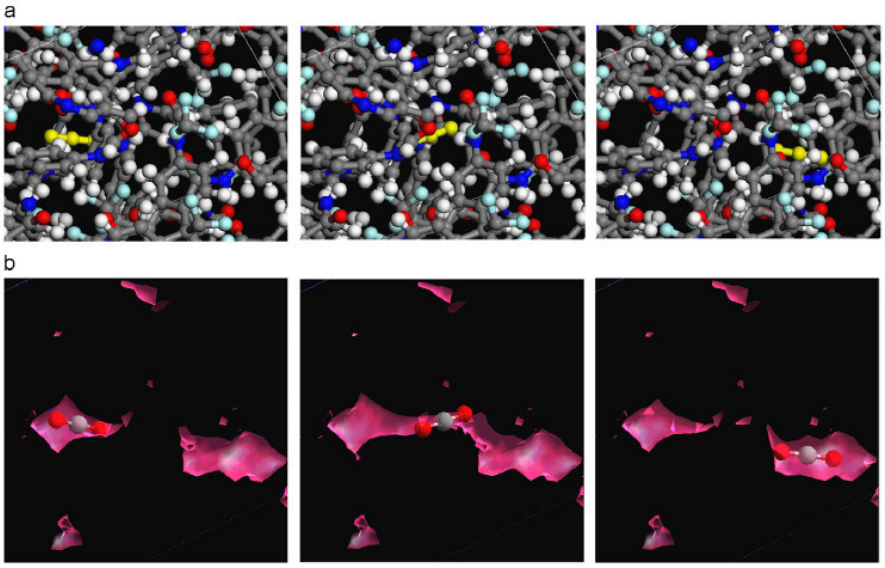


Figure 7. (a) Snapshots along the transition path of an elementary jump of a CO_2 molecule within an amorphous poly(amide imide) (PAI) matrix. The configurations on the left and right correspond to local minima of the potential energy of the CO_2 + PAI with respect to all atomic coordinates. The configuration in the middle corresponds to a saddle point of the potential energy. (b) Visualization of the accessible volume of the polymer, as determined using a spherical probe of radius 1.3 Å in the same three snapshots along the transition path. The CO_2 penetrant is also shown. In the saddle point configuration, polymer degrees of freedom have moved in such a way as to form a "neck" connecting the accessible volume clusters in the initial and final states. The orientation of the CO_2 at the saddle point is more or less parallel to this neck of accessible volume.

and hence the rate constant of the elementary jump.

After calculating all relevant rate constants $k_{I\to J}$ by multidimensional TST, the diffusive progress of CO_2 in the PAI matrix was tracked via Kinetic Monte Carlo simulation, applying periodic boundary conditions at the simulation cell boundaries (see Sec. 4). Fig. 8 displays the mean square displacement $\left\langle [\mathbf{r}(t)-\mathbf{r}(0)]^2 \right\rangle$ from KMC trajectories as a function of elapsed time in log-log (left) and linear (right) coordinates. A strongly anomalous regime ($\left\langle [\mathbf{r}(t)-\mathbf{r}(0)]^2 \right\rangle \propto t^n$ with $n<1$) is observed at short times. Beyond 1 μs, however, where the root mean square displacement exceeds the dimension L of the periodic simulation box, the dependence becomes linear, allowing one to extract the self-diffusion coefficient as one sixth the slope of the right-hand side plot, in linear coordinates [compare Eq. 38].

The presence of an anomalous regime at short times has by now been well established from simulations of transport in amorphous polymers. Anomalous diffusion is due to long-lived structural correlations in the polymer, which cause the diffusant to encounter a locally heterogeneous environment. From a practical point of view, anomalous diffusion increases the computational cost of simulations required for the prediction of D_s, since such simulations must be long enough for the Einstein (exponent $n=1$) regime to be adequately sampled. In glassy polymer matrices, the crossover from anomalous to normal diffusion

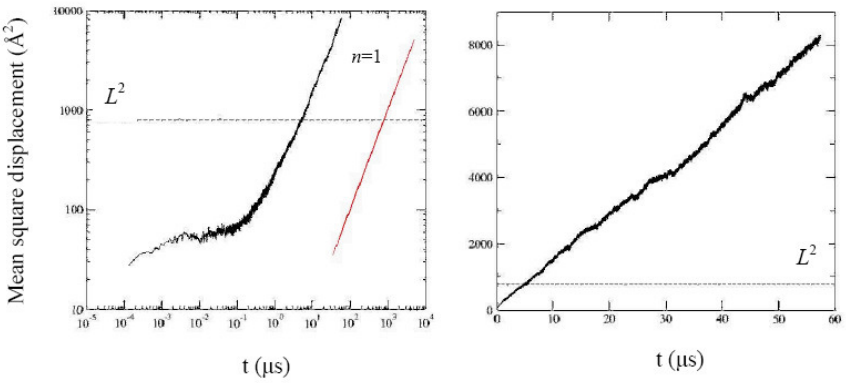


Figure 8. Mean square displacement of CO_2 penetrant in a glassy poly(amide imide) matrix as a function of time from kinetic Monte Carlo simulations of Vergadou[60] based on atomistically calculated sorption states and jump rate constants between them. (a, left): log-log coordinates; (b, right): Linear coordinates. The straight line marked $n = 1$ on the left-hand side plot indicates the expected slope for diffusion [Einstein equation, Eq. 38]. The dotted lines labelled L^2 mark the edge length of the primary simulation cell, on which periodic boundary conditions are applied. The self- diffusivity D_s is computed from the slope of the right-hand side plot.

is often observed at root mean squared penetrant displacements roughly equal to the simulation box size. This is a system size effect. At length scales larger than the simulation box size, the model matrix looks like a regular lattice to the penetrant; structural heterogeneities leading to anomalous diffusion are suppressed, precipitating a premature onset of the Einstein regime. Based on the work of Karayiannis[61], despite this premature onset, the estimate of D_s extracted from the linear part of the mean square displacement versus time curves is not significantly affected by system size, provided the model structures employed in the simulation are large enough and numerous enough. Karayiannis[61] has conducted a systematic KMC study and Effective Medium Theory analysis of the relation between the duration of the anomalous diffusion regime and the heterogeneity in the distribution of elementary jump rate constants.

Based on Fig. 8, the duration of the anomalous regime for diffusion of CO_2 in PAI is at least 1 μs. State-of-the-art measurements of CO_2 diffusion in glassy polymers with carbon-13 Pulsed Field Gradient NMR indicate that it may take 10 ms for motion of the penetrant to become fully isotropic and the Einstein regime to be reached[62].

From the slope of the Einstein regime of Fig. 8 we extract a diffusivity value for the diffusion of CO_2 in PAI at low concentration equal to $D_s = 0.25 \times 10^{-12} m^2 s^{-1}$. An experimental estimate is[63] $D_s = 0.81 \times 10^{-12} m^2 s^{-1}$. The solubility coefficient of CO_2 in the PAI, estimated by the Widom test particle insertion method[31] based on the same atomistic model, is $S = 0.42$ cm^3(STP)/(cm^3 polymer cmHg). The permeability $\mathcal{P} = D.S$ of CO_2 through the PAI is thus estimated as $\mathcal{P} = 10.5$ cm^3(STP) cm/(cm^2 s cmHg) $\times 10^{-10}$, or 10.5 barrer. This compares with experimental estimates of $\mathcal{P} = 9.54$ barrer[63] and $\mathcal{P} = 15.01$ barrer[64] from the literature. The comparison between predicted and experimental values is quite favorable, given the uncertainties in the force field employed, in the structure of the model polymer, but also in the measured permeabilities.

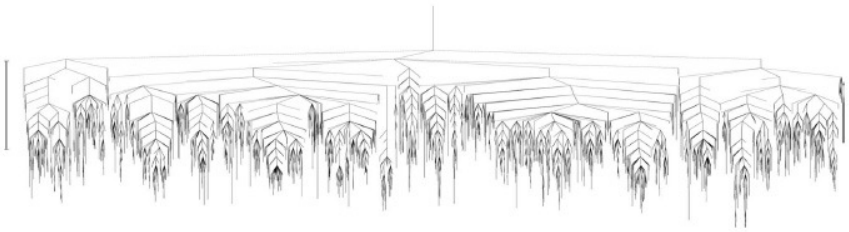


Figure 9. Disconnectivity graph for a liquid mixture of 48 A and 12 B type Lennard-Jones particles with $\epsilon_{\mathrm{AB}} = 1.5\epsilon_{\mathrm{AA}}$, $\epsilon_{\mathrm{BB}} = 0.5\epsilon_{\mathrm{AA}}$, $\sigma_{\mathrm{AB}} = 0.8\sigma_{\mathrm{AA}}$, $\sigma_{\mathrm{BB}} = 0.88\sigma_{\mathrm{AA}}$ at a particle number density $\rho = 1.3\sigma_{\mathrm{AA}}^{-3}$, as computed from MD simulations at a temperature of 0.71 $\epsilon_{\mathrm{AA}}/k_{\mathrm{B}}$. The glass temperature for this system is approximately $T_{\mathrm{g}} = 0.32\epsilon_{\mathrm{AA}}/k_{\mathrm{B}}$. The length of the scale bar on the left corresponds to a total system energy change of 10 ϵ_{AA}[49].

8 Dynamic Integration of a Markovian Web and its Application to Structural Relaxation in Glasses

Glassy materials play an important role in our life and have therefore constituted an object of extensive research, both at basic and applied levels. Glasses are nonequilibrium materials, their properties depending on their formation history. Furthermore, their properties change very slowly with time in the course of "physical ageing," whose characteristic times exceed common macroscopic observation times below the glass temperature T_{g}. The study of glassy materials by means of molecular simulation faces serious challenges, because one needs to bridge time scales spanning some 20 orders of magnitude, from the period of fast atomic vibrations (10^{-14} s) up to the longest time for structural, volume, and enthalpy relaxation (on the order of years 20° C or so below T_{g}).

State-of-the-art theories of the supercooled liquid state include mode coupling theory[46] and theories for enumerating stationary points[47] on the multidimensional energy hypersurface of the system. Analyses of the potential energy landscape have been reviewed[48].

"Fragile" glass-forming liquids, whose viscosity exhibits a strongly non-Arrhenius dependence on temperature, are characterized by very rugged potential energy landscapes. This is seen characteristically in the "disconnectivity graphs" computed by D. Wales and collaborators[49] (see Fig. 9). All branches of the inverted tree in a disconnectivity graph terminate at a local minimum of the energy (inherent structure). Relative energies can be read off on the vertical axis. The node (branch point) through which two inherent structures communicate corresponds to the lowest lying first-order saddle point between these structures. From the "willow tree" appearance of the graph, it is clear that there are sets of basins ("metabasins") communicating through relatively fast transitions, sets of metabasins communicating through slower transitions etc., i.e., the potential energy landscape exhibits a hierarchical structure.

The complexity of the energy landscape of a binary Lennard-Jones glass of the same composition as that studied in Ref. 49 is also seen in Fig. 10, taken from the work of Tsalikis et al.[29]. Here a system consisting of $N = 641$ particles is considered, at a constant number density $\rho = 1.1908\sigma_{\mathrm{AA}}^{-3}$. To analyze the dynamics in real time units, the properties of Argon have been attributed to component A ($m_{\mathrm{A}} = m_{\mathrm{B}} = 6.634 \times 10^{-26}$kg,

$\epsilon_{\rm AA} = 1.65678 \times 10^{-21}$J, $\sigma_{\rm AA} = 3.4 \times 10^{-10}$m). With these assignments, the glass temperature of the system is $T_{\rm g} = 38.4K$. Fig. 10 refers to a set of 290 basins that were identified as belonging to a metabasin through MD simulation at 37K. By "belonging to a metabasin" here we mean that the time required for the system to escape from this particular set of basins is significantly longer than the time needed for the system to establish a restricted equilibrium among the basins in the set. In a plot of the number of distinct basins visited versus time in the course of a MD simulation, this reflects itself as a plateau[50]. The number of identified distinct transitions between pairs of these 290 basins is plotted as a function of the rate constant of the transitions in Fig. 10. The long-dashed line shows results from a 3 ns-long NVT MD simulation at 37 K, which was trapped within the metabasin (trajectories were turned back as soon as they were found to exit the metabasin); a total of 3910 distinct transitions were observed during this simulation. The short-dashed line shows results from a swarm of NVE MD trajectories generated in parallel off of an NVT MD trajectory at 37 K. These were able to provide a more thorough sampling of transitions within the metabasin; a total of 24271 distinct transitions were sampled. The solid line comes from a temperature-accelerated MD (TAD) method, which used as input data from swarms of NVE MD trajectories generated in parallel off of NVT MD trajectories conducted at temperatures from 37 K to 55 K. A histogram reweighting method was invoked to translate all data to 37 K (see Sec. 3 and Ref. 29). This latter sampling method, which identified a total of 51207 distinct transitions, was able to access a rich variety of passages between the basins in the metabasin, including passages that go through high-lying terrain in the rugged potential energy landscape of the system. This explains the "wing" extending to very low rate constants on the left-hand side of Fig. 10. Clearly, the fastest transitions sampled have a rate constant around $\nu_0 \simeq 10^{13}$ s^{-1}. The "nose" around 10^{10} s^{-1} is a consequence of the fact that the studied basins belong to a metabasin, so they communicate through relatively low-lying passages with each other. A time of approximately 10^{-10} s is needed for the system to visit the entire metabasin. The wing extending to very low rate constants (indeed, too low to be physically relevant at the reference temperature of 37 K, see inset) tells us something about the topography of the landscape. The inset of Fig. 10 suggests a power-law distribution of rate constants between basins, of the form:

$$\rho(k_{a\to b}/\nu_0) \simeq B(k_{a\to b}/\nu_0)^{\alpha}, \quad \alpha \simeq 0.01 \tag{40}$$

and hence an exponential distribution of barrier heights $E_{a\to b} = -k_B T \ln(k_{a\to b}/\nu_0)$ Interestingly, this is similar to the form proposed for the distribution of barrier heights by J.P. Bouchaud on theoretical grounds[51].

Tsalikis et al.[29] have correlated the rate constants of transitions sampled via their temperature accelerated dynamics/histogram reweighting scheme with the distance traversed in configuration space, with the cooperativity of the transitions, and with their molecular mechanisms. Fast intrabasin transitions in the binary Lennard-Jones system tend to involve single "cage-breaking" events, wherein more than half of the first neighbors of an atom change, or multiple "cage breaking" events occurring at different points in the system. Slower interbasin transitions tend to involve coordinated displacements of "chains" of atoms, wherein each atom jumps to a position close to that previously occupied by another atom in the chain. Even slower, more cooperative transitions involve extended formation of several interlinked chains or massively coordinated displacements which look like shear bands.

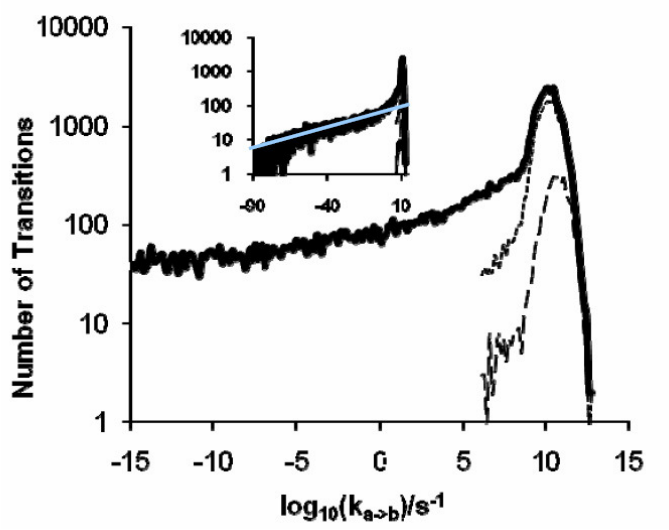


Figure 10. Number of identified distinct transitions between the 290 basins of a metabasin of a Lennard-Jones mixture at 37 K as a function of the rate constant of the transitions, as computed by three sampling methods (see text for details). The inset shows the total range of rate constants sampled by the temperature-accelerated method. The light-colored straight line through the plot in the inset corresponds to Eq. 40.

How do we track structural relaxation of a glass at temperatures below T_g over times relevant to the applications of glasses as structural, optical, packaging, and membrane materials? These time scales (milliseconds to years) are too long to be addressed by direct MD simulation, so reverting to an infrequent event theory-based approach seems appropriate. On the other hand, the rugged potential energy landscape of glass-forming systems gives rise to a very broad distribution of characteristic times for elementary transitions and a complex connectivity among basins. KMC simulation would have to track the fastest of these transitions, and this would limit its ability to sample long-time evolution. An approach based on analytical solution of the master equation, equivalent to averaging over all dynamical trajectories originating from a given initial distribution among basins, would seem more promising. However, it is impossible to build a complete map of all basins and transitions between them in the rugged potential energy landscape of a glassy system even of modest size N. A way out of this difficulty is provided by the fact that, when one studies structural relaxation, one typically starts from an initial distribution among states that is highly localized (e.g. from a single basin in the potential energy landscape, where the system was trapped via the glass formation history that was followed to obtain it). The region of configuration space where the system resides is thus initially very confined, and expands gradually as transitions between basins take place.

This idea led Boulougouris and Theodorou[21] to develop a computational approach for tracking the temporal evolution of the distribution among basins (or "states") via infrequent transitions, starting off from a highly localized initial distribution, which they called "Dynamic Integration of a Markovian Web," or DIMW. DIMW distinguishes states that it

samples into two categories: "explored" and "boundary" states. An "explored" state is a state for which an exhaustive calculation of as many as possible transition pathways leading out of it to neighboring states has been undertaken and rate constants associated with these transitions have been computed. In the application to isothermal - isochoric structural relaxation of a polymer glass, discussed in Ref. 21, this calculation proceeds by computing as many as possible saddle points of the potential energy in $3N-3$-dimensional configuration space around the state under investigation using the dimer method[22] and subsequently constructing a transition path through each of these saddle points to neighboring states via Fukui's intrinsic reaction coordinate approach[18]. Strict energy- and configuration-based criteria for identifying states that have already been visited have been implemented in connection with this exploration process[21]. For each transition pathway, a rate constant is computed. In the application presented in Ref. 21, this computation was based on transition-state theory in the harmonic approximation [compare Eq. 24]. "Boundary" states, on the other hand, are states connected to explored states, which, however, have not been explored themselves. The DIMW algorithm proceeds as follows:

(1) All states populated according to the narrow initial distribution $\mathbf{P}(0)$ are fully explored, as described above, and boundary states connected to these states are identified. Rate constants are computed for all identified transitions emanating from an explored state and for their reverse transitions. Bookkeeping of the explored and boundary states, of the connectivity among them and of associated rate constants, is initialized. Let E and B symbolize the current set of explored and boundary states, respectively.

(2) The evolution of the occupancy probabilities of explored and boundary states for times short enough for the current set of explored states to be adequate is tracked by analytical solution of the master equation in the current explored and boundary states, initial occupancy probabilities for the boundary states being zero and all rate constants not emanating from or terminating in an explored state being taken as zero:

$$\frac{\partial P_i}{\partial t} = \sum_{j \neq i} P_j k_{j \to i} - P_i \sum_{j \neq i} k_{i \to j}, \quad i, j \in E \cup B \tag{41}$$

From the solution to Eq. 41 we compute the total probability of the system residing in the current set of explored states at time t,

$$P_E(t) = \sum_{i \in E} P_i(t). \tag{42}$$

We also compute the efflux of probability from the current set of explored states to each one of the current boundary states,

$$f_j(t) = \sum_{i \in E} P_i(t) k_{i \to j}, \quad j \in B \tag{43}$$

as well as the total efflux of probability from the current set of explored states to the current boundary states,

$$f_B(t) = \sum_{j \in B} f_j(t) \tag{44}$$

(3) For times commensurate with the first passage time for exit of the system from the current set of explored states, the latter set will no longer be adequate. Clearly, for such times the set of explored states must be augmented by including more states. We select a time $t_{\rm select}$ for first passage of the system out of the current set of explored states by sampling the distribution $f_B(t)/\int_0^\infty f_B(t)dt$.

(4) We pick one of the boundary states in set B, $j_{\rm select}$, according to the discrete probabilities $f_j(t_{\rm select})/f_B(t_{\rm select})$. The selected state will be appended to the set of explored states, E.

(5) We update the set E by including state $j_{\rm select}$ in it. Furthermore, we proceed to explore state $j_{\rm select}$ and update set B by removing state $j_{\rm select}$ from it and appending to it all states connected to $j_{\rm select}$ not already belonging to $E \cup B$ that were identified through the exploration of $j_{\rm select}$. Finally, we identify a time $t_{\rm safe}$, beyond which the updated sets E and B have to be used. This time is calculated via the condition $P_E(t_{\rm safe}) = 1 - \delta$, with $P_E(t)$ being the probability of residing in the set of explored states before the update, computed in step 2. A value of $\delta = 10^{-3}$ was used in the application presented in Ref. 21.

(6) We check whether time $t_{\rm safe}$ has exceeded the desired simulation time. If not, we return to step 2 to solve the master equation analytically with the same initial conditions, but in the augmented set of explored states with the updated set of boundary states. For $t < t_{\rm safe}$, the resulting solution should be practically indistinguishable from that obtained so far. For $t \geq t_{\rm safe}$, the solution for the augmented set of explored states should be used.

As described above, DIMW amounts to a series of analytical solutions of the master equation in a set of explored and boundary states that is progressively augmented "on the fly," with rate constants determined from atomistic infrequent event analysis. The progressive augmentation of the set of explored states has a "self-healing" aspect; important connections that were missed at shorter times may be discovered as the network of explored states is expanded. The outcome from performing this calculation out to long times is a set of analytical expressions for the time-dependent probabilities $P_i(t)$ of the explored states.

Fig. 11 displays the result from a DIMW calculation of structural relaxation in a 641 united atom model of glassy atactic polystyrene (aPS) at 250 K, roughly 123 K below the experimental glass temperature $T_{\rm g}$, at a density of 0.951 g/cm^3, equal to the orthobaric density at that temperature[21]. The calculation was performed out to 10^{-5} s with modest computational cost. 240 distinct states were explored and 2880 saddle points were identified in the course of the calculation. Shown in Fig. 11a is a "time-dependent Helmholtz energy" for the system, calculated as

$$A(t) = \sum_i P_i(t) A_i(t) + k_{\rm B}T \sum_i P_i(t) \ln P_i(t), \;\; i \in E \tag{45}$$

with $P_i(t)$ being the time-dependent probability of occupancy of explored state i from the DIMW calculation and $A_i(t)$ being the Helmholtz energy of the system confined in state

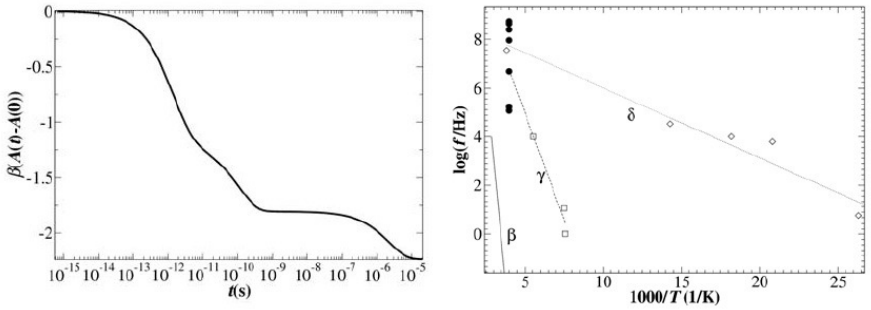


Figure 11. a (left): Helmholtz energy as a function of time for physical ageing of an aPS computer "specimen" at 250 K and initial pressure 1 bar at constant volume, as determined through the DIMW approach. b (right): Characteristic rate constants for the modes from diagonalization of the rate constant matrix at 10^{-5} s (filled symbols) compared to the peak frequencies of loss modulus measurements on aPS at various temperatures (open symbols). The quantity f on the abscissa can be identified with $-\lambda_i$ in the text.

i, computed according to the harmonic approximation [compare Eqs. 20 22]. Note that $A(t)$ consists of an average of the Helmholtz energies $A_i(t)$ of the system confined in each individual state (basin), each state being weighted by its occupancy probability at time t, plus a term of entropic origin that has to do with exchange of probability among the states. At infinite time, when the system would distribute itself according to the Boltzmann distribution in its entire configuration space, $A(t)$ would become the Helmholtz energy of equilibrium thermodynamics. For the relaxing glass, which starts off occupying a single state, $A(t)$ decays with time as the system strives to approach thermodynamic equilibrium. It is interesting that this decay is not featureless, but exhibits characteristic shoulders and plateaux over specific time domains. These features betray the existence of specific relaxation processes. A plateau in $A(t)$ suggests that the system equilibrates locally within a "metabasin" of states that communicate easily with each other and is temporarily trapped there before overcoming the barriers surrounding the metabasin and moving on to states of lower free energy.

One can readily bring out the characteristic rate constants $-\lambda_i$ of modes contributing to relaxation by diagonalizing the rate constant matrix at the longest time accessed, 10^{-5} s. Results from this diagonalization are displayed in Fig. 11b (compare Sec. 5). In the same figure are shown Arrhenius plots for subglass relaxation processes in aPS, determined experimentally by dynamic mechanical spectroscopy[21]. One sees that the characteristic frequencies determined by the DIMW calculation cluster in two frequency ranges, around 10^5 and around 10^9 s^{-1}. These values are quite close to the characteristic frequencies of the so-called γ and δ subglass relaxation processes determined experimentally.

Using the EROPHILE approach (Sec. 5), one can readily compute time autocorrelation functions for specific vectors in the system and analyze the contribution of each mode to the decay of these functions [compare Eq. 34]. Boulougouris and Theodorou[41] have examined the autocorrelation functions of unit vectors normal to the phenyl planes and of unit vectors directed along phenyl stems. Two modes were found to contribute significantly to the

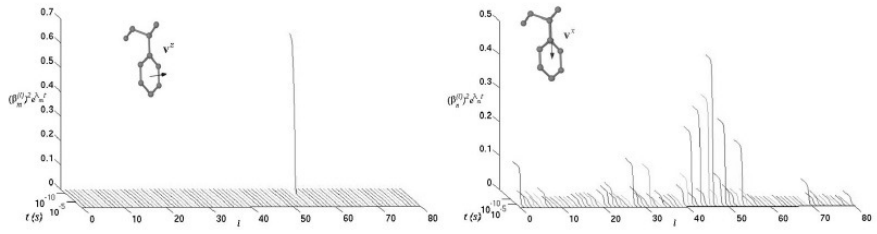


Figure 12. a (left): Fast mode ($\lambda_i = -10^{9.5}\mathrm{s}^{-1}$) contribution to the orientational decorrelation of unit vectors normal to the plane of each phenyl ring in aPS at 250 K. The index l measured along the axis running from left to right enumerates different phenyl rings in the system. b (right): Slower mode ($\lambda_i = -10^{5.2}\mathrm{s}^{-1}$) contribution to the orientational decorrelation of unit vectors along the stem of each phenyl ring in the same system.

decorrelation of these vectors: A fast mode with $\lambda_i = -10^{9.5}\mathrm{s}^{-1}$, which can be associated with the δ subglass relaxation process, and a slower mode with $\lambda_i = -10^{5.2}\mathrm{s}^{-1}$, which can be associated with the γ subglass relaxation process (see also Fig. 11). In Fig. 12, the contributions of these modes to the decorrelation of the characteristic vectors of each phenyl group l in the model glassy aPS system are displayed. The fast mode corresponds to rotation of an isolated, mobile phenyl in the system around its stem. On the other hand, the slower mode corresponds to a cooperative motion involving changes in orientation of several phenyl stems. As regards this latter motion, one can discern relatively long sequences of phenyls along the aPS chain that exhibit very little decorrelation. These sequences tend to be syndiotactic in their stereochemical configuration.

This aPS example shows how mechanictic aspects of dynamics in a system with very complex potential energy landscape can be explored in an unbiased way using a combination of DIMW and EROPHILE methodologies.

9 Lumping

A difficulty with DIMW-type approaches (see Sec. 8) is that the number of states to be tracked becomes prohibitively large at long times. A way out of this problem is to group, or "lump," states communicating via transitions that are fast in relation to the observation time into single clusters of states. If performed judiciously, this lumping does not result in loss of essential information. At long observation times, the system distributes itself among fast-communicating states according to the requirements of a restricted equilibrium (compare plateaux in Fig. 11a), so clusters of such states behave as single "meta-states," for all practical purposes.

From the mathematical point of view, lumping is not a new problem. It has been examined in the context of networks of chemical reactions in the classic work of Wei and Kuo [52] and in several subsequent works. As shown there, lumping calls for the determination of a $\hat{n} \times n$ transformation matrix $\mathbf{M}$, where n is the number of original states and $\hat{n} < n$ is the number of lumped states (or clusters of states). The transformation from the probability distribution among the original states to that among the lumped states at any time t takes

place according to the equation

$$\hat{\mathbf{P}}(t) = \mathbf{M} \cdot \mathbf{P}(t) \tag{46}$$

The lumping matrix $\mathbf{M}$ has the following properties:

i. The elements of matrix $\mathbf{M}$ are either "0" or "1".

ii. Every column of matrix $\mathbf{M}$ contains exactly one "1". The physical meaning behind this is that every state of the original description (assigned to a column of $\mathbf{M}$) belongs to one cluster only (assigned to a row of $\mathbf{M}$).

iii. The position of "1" in every column of $\mathbf{M}$ (i.e., state in the original description) describes to which cluster (row of $\mathbf{M}$) the state of the initial system is being lumped.

Once $\mathbf{M}$ is known, the $\hat{n} \times \hat{n}$ rate constant matrix $\hat{\mathbf{K}}$ to be used at the lumped level is calculated as

$$\hat{\mathbf{K}} = \mathbf{M} \cdot \mathbf{K} \cdot \mathbf{A} \cdot \mathbf{M}^{\mathrm{T}} \cdot \hat{\mathbf{A}}^{-1} \tag{47}$$

where $\mathbf{A}$ is a $n \times n$ diagonal matrix whose diagonal elements equal the elements of the equilibrium probability vector $\mathbf{P}(\infty)$ corresponding to the original rate constant matrix $\mathbf{K}$, the superscripts "T" and "-1" indicate matrix transpose and matrix inverse, respectively, and

$$\hat{\mathbf{A}} = \mathbf{M} \cdot \mathbf{A} \cdot \mathbf{M}^{\mathrm{T}} \tag{48}$$

Lempesis et al.[43] proposed a methodology for the determination of the number of lumped states $\hat{n}$ and the lumping matrix $\mathbf{M}$ in such a way that the long-time dynamics of the original description is reproduced. The strategy is to minimize an objective function of the form

$$z(\hat{n}, \mathbf{M}) = z_1 E + z_2 W + z_3 \hat{n} \tag{49}$$

with z_1, z_2, z_3 being pre-defined real positive constants.

E is the Frobenius norm of the $\hat{n} \times n$ error matrix $\mathbf{E}$[52]:

$$E = ||\mathbf{E}||_F = \sqrt{\sum_{i=1}^{\hat{n}} \sum_{j=1}^{n} |E_{ij}|^2} \tag{50}$$

$$\mathbf{E} = \mathbf{M} \cdot \mathbf{K} - \hat{\mathbf{K}} \cdot \mathbf{M} \tag{51}$$

For exact lumping, the lumping error E would be zero. W, on the other hand, is the Frobenius norm of the lumped matrix $\hat{\mathbf{K}}$:

$$W = ||\hat{\mathbf{K}}||_F = \sqrt{\sum_{i=1}^{\hat{n}} \sum_{j=1}^{\hat{n}} |\hat{K}_{ij}|^2} \tag{52}$$

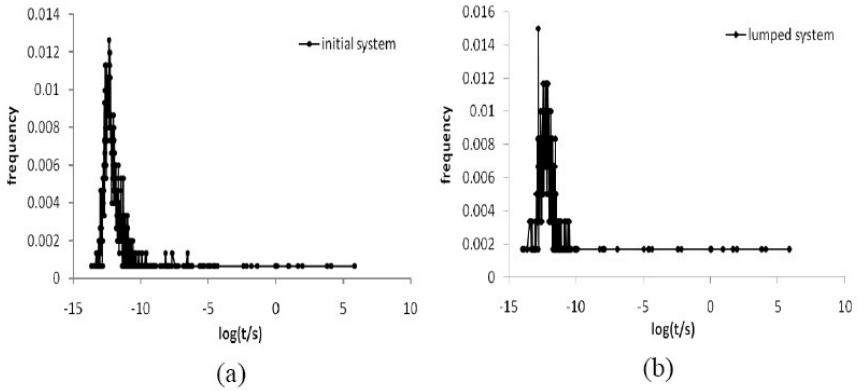


Figure 13. Histogram of the negative inverse eigenvalues for (a) the initial description, (b) the lumped description of the dynamics of a mixture of 641 Lennard-Jones particles just below $T_{\rm g}$ (see text for details). There are $n = 1502$ eigenvalues in (a) and $\hat{n} = 600$ eigenvalues in (b).

Including a term proportional to W in the objective function, Eq. 49, forces the minimization to focus on long times (small W) in matching the dynamics between the original and the lumped system. Including a term proportional to $\hat{n}$ in the objective function, on the other hand, encourages the algorithm to keep the dimensionality of the lumped system as small as possible.

The minimization of the objective function defined in Eq. 49 is performed stochastically, using Monte Carlo moves which change the dimensionality $\hat{n}$ and the form of the lumping matrix $\mathbf{M}$, while respecting the constraints on the form of that matrix stated above. To avoid trapping in local minima of the objective function, a Wang-Landau scheme is invoked to determine the density of $\mathbf{M}$-matrix "states" in the space of variables $(E, W, \hat{n})$ and pick that $\mathbf{M}$, close to the origin of $(E, W, \hat{n})$ space, which minimizes the objective function[43].

Fig. 13 shows results from application of the lumping strategy of Lempesis et al.[43] to a mixture of 641 Lennard-Jones particles with the interaction parameters stated in Sec. 8 and atomic fractions 80% A, 20% B, at a temperature of 37 K, just below $T_{\rm g}$. Shown are histograms of the negative inverse eigenvalues $t_i = -1/\lambda_i$ of the rate constant matrices $\mathbf{K}$ (original description) and $\hat{\mathbf{K}}$ (lumped description). The overall shapes of the histograms are seen to be similar. Furthermore, the eight longest t_i values are seen to agree quantitatively between the original and lumped system, testifying to the success of the lumping method in reproducing the long-time dynamics of the original system.

10 Summary

Addressing long-time ($> 1\mu$s) dynamics in many materials, complex fluid, and biomolecular systems constitutes a great challenge for molecular simulations. In many cases, the temporal evolution of a system is slow because the system spends a long time confined

within regions in configuration space ("states") and only infrequently jumps from state to state by overcoming a (free) energy barrier separating the states. We have briefly discussed ways of probing the time scale separation underlying these infrequent transitions and identifying states, either in terms of all the degrees of freedom or in terms of appropriately chosen slow variables or order parameters. We have also reviewed analytical and simulation techniques, based on the theory of infrequent events, for estimating the rate constants $k_{i\to j}$ for transitions between states.

Emphasis in these notes has been placed on how we predict the long-time dynamical evolution once we have identified a network of states and computed the rate constants between them. We have discussed the principles of two categories of methods for doing this: Kinetic Monte Carlo simulations, which generate long stochastic trajectories for the evolution of the system; and analytical solution of the master equation, which yields expressions for the time-dependent probabilities of occupancy of the states as sums of exponentially decaying functions after diagonalization of an appropriately symmetrized rate constant matrix. We have seen that the analytical solution to the master equation can form the basis for calculating useful time-dependent ensemble averages and correlation functions that quantify the approach to equilibrium and enable the calculation of time-dependent properties in the context of the Eigenvalue Representation of Observables and Probabilities in a HIgh-dimensional Euclidean space[41] (EROPHILE) approach. We have presented applications of both Kinetic Monte Carlo and analytical solution of the master equation to problems of diffusion in zeolites and in amorphous polymers. We have also discussed advantages of the analytical solution in cases where the spectrum of characteristic times for evolution on the network of states, quantified by the eigenvalues of the rate constant matrix, is very broad. In systems characterized by spatial periodicity, such as zeolites, analytical solution of the master equation can be made several orders of magnitude faster than Kinetic Monte Carlo, thanks to a recursive scheme[44] (MESoRReD) that reduces diagonalization of the rate constant matrix for the whole system to diagonalization of much smaller matrices pertaining to a single unit cell.

Nonequilibrium systems with rugged or fractal potential energy hypersurfaces, such as glasses, preclude the a priori determination of all states and transitions between them. One is often interested in the evolution of such systems starting from a narrow, localized distribution in configuration space (e.g., tracking the structural relaxation of a glassy configuration). For addressing this problem, we have introduced Dynamic Integration of a Markovian Web[21] (DIMW), which solves the master equation in a network of states that is progressively augmented as time elapses based on an "on the fly" exploration of configuration space and calculation of rate constants. Application of the DIMW approach to a polymer glass has yielded promising results. To keep the number of states manageable at long times, DIMW can be complemented by a "lumping" algorithm[43] which groups fast-communicating states into single "metastates." This algorithm has been applied successfully to a glassy binary Lennard-Jones mixture.

It is hoped that the concepts and computational tools discussed here may be useful in addressing the long-time properties of systems enountered in the wide range of problems attacked by today's physicists, chemists, chemical engineers, materials scientists, and molecular biologists, starting from fundamental atomic-level information.

Acknowledgments

I am grateful to my collaborators Larry June, Randy Snurr, Mike Greenfield, Nikos Kopsias, Niki Vergadou, George Boulougouris, Dimitris Tsalikis, Nikos Lempesis and Takis Kolokathis for their hard work that helped us develop the concepts and methods presented in these notes.

References

1. G. H. Vineyard, *Frequency factors and isotope effects in solid state rate processes*, J. Phys. Chem. Solids **3**, 121–127, 1957.
2. A. A. Gusev, F. Müller-Plathe, W. F. van Gunsteren, U. W. Suter, *Dynamics of small molecules in bulk polymers*, Advan. Polym. Sci. **116**, 207-247, 1994.
3. D. N. Theodorou, R. Q. Snurr, A. T. Bell, *Molecular dynamics and diffusion in microporous materials*, in G. Alberti, T. Bein (Eds.) *Comprehensive Supramolecular Chemistry* (Elsevier Science, Oxford, 1994), pp 507-548.
4. F. H. Stillinger, *A topographic view of supercooled liquids and glass formation*, Science **267**, 1935-1939, 1995.
5. D. J. Wales, J. P. K. Doye, M. A. Miller, P. N. Mortenson, T. R. Walsh, *Energy landscapes: from clusters to biomolecules*, Advan. Chem. Phys. **115**, 1-111, 2000.
6. A. F. Voter, J. D. Doll, *Dynamical corrections to transition-state theory for multistate systems - Surface self-diffusion in the rare event regime*, J. Chem. Phys. **82**, 80-92, 1985.
7. P. G. Wolynes, *Energy landscapes and solved protein-folding problems* Phil. Trans. Roy. Soc.A **363**, 453-464, 2005.
8. J. Wei, C. D. Prater, *The structure and analysis of complex reaction systems*, Advan. Catal. **13**, 204-390, 1962.
9. D. E. Shaw, *Millisecond-long molecular dynamics simulations of proteins on the Anton machine*, in *Proceedings of FOMMS 2009: Foundations for Innovation*, Blaine, WA, 2009.
10. D. Chandler, *Statistical-mechanics of isomerization dynamics in liquids and transition-state approximation* J. Chem. Phys. **68**, 2959-2970, 1978.
11. D. Chandler, *Throwing ropes over rough mountain passes, in the dark* in B. J. Berne, G. Ciccotti, D. F. Coker (Eds.) *Classical and quantum dynamics in condensed-phase simulations* (World Scientific, Singapore, 1998), pp 55-66.
12. W. Feller, *An introduction to probability theory and its applications: Volume 2* (John Wiley, New York, 1971).
13. R. L. June, A. T. Bell, D. N. Theodorou, *Transition state studies of xenon and SF_6 diffusion in silicalite* J. Phys. Chem. **95**, 8866-8878, 1991.
14. R. Q. Snurr, A. T. Bell, D. N. Theodorou, *Investigation of the dynamics of benzene in silicalite using transition-state theory* J. Phys. Chem. **98**, 11948-11961, 1991.
15. M. L. Greenfield, D. N. Theodorou, *Molecular modeling of methane diffusion in glassy atactic polypropylene via multidimensional transition state theory* Macromolecules **31**, 7068-7090, 1998.
16. M. L. Greenfield, D. N. Theodorou, *Geometric analysis of diffusion pathways in glassy and melt atactic polypropylene* Macromolecules **26**, 5461-5472, 1993.

17. J. Baker *An algorithm for the location of transition states* J. Comput. Chem **7**, 385-395, 1986.
18. K. Fukui, *The path of chemical reactions - the IRC approach* Acc. Chem. Res. **14**, 363-368, 1981.
19. D. N. Theodorou, *Principles of molecular simulation of gas transport in polymers*, in Y. Yampolskii, I. Pinnau, B. Freeman (Eds.) *Materials Science of Membranes for Gas and Vapor Separation* (John Wiley and Sons, New York, 2006), pp 49-94.
20. N. P. Kopsias, D. N. Theodorou, *Elementary structural transitions in the amorphous Lennard-Jones solid using multidimensional transition-state theory* J. Chem. Phys. **109**, 8573-8582, 1998.
21. G. C. Boulougouris, D. N. Theodorou, *Dynamical integration of a Markovian web: a first passage time approach* J. Chem. Phys. **125**, 084903, 2007.
22. G. Henkelman, H. Jónsson, *A dimer method for finding saddle points in high dimensional potential surfaces using only first derivatives* J. Chem. Phys. **111**, 7010-7022, 1999.
23. T. R. Forester, W. Smith, *Bluemoon simulations of benzene in silicalite-1: Prediction of free energies and diffusion coefficients* J. Chem. Soc., Faraday Trans. **93**, 3249-3257, 1997.
24. D. Frenkel, B. Smit, *Understanding molecular simulation: From algorithms to applications*, 2nd Edition, (Academic Press, New York, 2002).
25. F. H. Stillinger, T. A. Weber, *Hidden structure in liquids* Phys. Rev. A **25**, 978-989, 1982.
26. E. Helfand, *Brownian dynamics study of transitions in a polymer chain of bistable oscillators* J. Chem. Phys. **69**, 1010-1018, 1978.
27. D. G. Tsalikis, N. Lempesis, G. C. Boulougouris, D. N. Theodorou, *On the role of inherent structures in glass forming materials: I. The vitrification process* J. Phys. Chem. B **112**, 10619-10627, 2008.
28. M. R. Sørensen, A. F. Voter *Temperature-accelerated dynamics for simulation of infrequent events* J. Chem. Phys. **112**, 9599-9606, 2000.
29. D. G. Tsalikis, N. Lempesis, G. C. Boulougouris, D. N. Theodorou, *Temperature accelerated dynamics in glass-forming materials* J. Phys. Chem. B **114**, 7844-8753, 2010.
30. C. H. Bennett, *Exact defect calculations in model substances* in A. S. Nowick, J. J. Burton (Eds.) *Diffusion in solids: recent developments* (Academic Press: New York, 1975), pp 73-113.
31. M. P. Allen, D. J. Tildesley, *Computer simulation of liquids* (Clarendon, Oxford, 1987).
32. D. A. Kofke, *Free energy methods in molecular simulation* Fluid Phase Equil. **228**, 41-48, 2005.
33. P. Yi and G. C. Rutledge, *Molecular origins of homogeneous crystal nucleation* Annu. Rev. Chem. Biomol. Eng. **3**, in press, 2005.
34. S. Singh, C. C. Chiu, J. de Pablo *Flux Tempered Metadynamics* J. Stat. Phys. **144**, 1-14, 2011.
35. A. Laio, M. Parrinello *Escaping free energy minima* Proc. Natl. Acad. Sci. USA **99**, 12562-12566, 2002.

36. P. G. Bolhuis, D. Chandler, C. Dellago, P.L. Geissler *Transition Path Sampling: Throwing ropes over mountain passes, in the dark* Annu. Rev. Phys. Chem. **53**, 291-318, 2002.
37. C. Dellago, P. G. Bolhuis, P. Geissler *Transition Path Sampling* Adv. Chem. Phys. **123**, 1-78, 2002.
38. A. F. Voter, *Introduction to the Kinetic Monte Carlo Method* in K. E. Sickafus, E. A. Kotomin, B. P. Uberuaga (Eds.), *Radiation Effects in Solids*, NATO Science Series II: Mathematics, Physics, and Chemistry, Vol. 235 (Springer: Dordrecht, The Netherlands, 2007), Chapter 1, pp 1-23.
39. D. E. Knuth, *The art of computer programming* Vol.2, (Addison-Wesley, Reading, MA, 1997), Chapter 3.
40. N. V. Buchete and G.Hummer, *Coarse master equations for peptide folding dynamics*, J. Phys. Chem. B **112**, 6057-6069, 2008
41. G. Boulougouris, D. N. Theodorou, *Probing subglass relaxation in polymers via a geometric representation of probabilities, observables, and relaxation modes for discrete stochastic systems* J. Chem. Phys. **130**, 044905, 2009.
42. K. E. Shuler, *Relaxation processes in multistate systems* Phys. Fluids **2**, 442-448, 1959
43. N. Lempesis, D. G. Tsalikis, G. C. Boulougouris, D. N. Theodorou, *Lumping analysis for the prediction of long-time dynamics: From monomolecular reaction systems to inherent structure dynamics of glassy materials* J. Chem. Phys. **135**, 204507, 2011.
44. P. Kolokathis, D. N. Theodorou, *On solving the master equation in spatially periodic systems*, in preparation
45. R. L. June, A. T. Bell, D. N. Theodorou, *Molecular dynamics study of methane and xenon in silicalite* J. Phys. Chem **94**, 8232-8240, 1990
46. Götze, W., *Aspects of structural glass transitions*, in J.P. Hansen, D. Levesque, J. Zinn-Justin (Eds.) *Les Houches Session LI, 1989 Liquids, Freezing and Glass Transition* (Elsevier Science Publishers B.V., Amsterdam 1991), pp 287-503.
47. S. M. Shell, P. G. Debenedetti, A. Z. Panagiotopoulos *A conformal solution theory for the energy landscape and glass transition of mixtures* Fluid Phase Equilib. **241**, 147-154, 2006.
48. F. Sciortino, *Potential energy description of supercooled liquids and glasses* J. Stat. Mech. , P050515, 2005.
49. V. K. de Souza, D. J. Wales, *Connectivity in the potential energy landscape for binary Lennard-Jones systems* J. Chem. Phys. **130**, 194508, 2009.
50. D. G. Tsalikis, N. Lempesis, G. Boulougouris, D. N. Theodorou, *Efficient parallel decomposition of dynamical sampling in glass-forming materials based on an "on the fly" definition of metabasins* J. Chem. Theory Comp. **6**, 1307-1322, 2010.
51. J. P. Bouchaud, *Weak ergodicity breaking and aging in disordered systems* J. Phys. I France **2**, 1705-1713, 1992.
52. J. Wei, J. C. W. Kuo, *A lumping analysis in monomolecular reaction systems - Analysis of exactly lumpable systems* Ind. Eng. Chem. Fundam. **8**, 114-123, 1969.
53. J. Kärger, H. Pfeifer, F. Stallmach, N. N. Feoktistova, S.P. Zhdanov, *^{129}Xe and ^{13}C PFG-NMR study of the intracrystalline self- diffusion of Xe, CO_2, and CO* Zeolites **13**, 50-55, 1993.

54. H. G. Karge, J. Weitkamp, *Molecular sieves - Characterization II* (Springer, Berlin, 2007).
55. D. N. Theodorou, *Hierarchical modelling of polymeric materials* Chem. Eng. Sci. **62**, 5697-5714, 2007.
56. A. A. Gusev, U. W. Suter, *Dynamics of small molecules in polymers subject to thermal motion* J. Chem. Phys. **99**, 2228-2234, 1993.
57. A. A. Gusev, U. W. Suter, W. R. van Gunsteren, U. W. Suter, *Dynamics of small molecules in bulk polymers* Advan. Polym. Sci. **116**, 207-247, 1994.
58. M. L. Greenfield, D. N. Theodorou, *Coarse-grained molecular simulation of penetrant diffusion in a glassy polymer using reverse and Kinetic Monte Carlo* Macromolecules **34**, 8541-8553, 2001.
59. M. L. Greenfield, *Sorption and diffusion of small molecules using Transition State Theory* in M. J. Kotelyanskii and D. N. Theodorou (Eds.) *Simulation methods for polymers* (Marcel Dekker, New York 2004).
60. N. Vergadou, *Prediction of gas permeability of stiff-chain amorphous polymers through molecular simulation methods* (Ph.D. Thesis, Chemistry Department, University of Athens, 2006).
61. N. Ch. Karayiannis, V. G. Mavrantzas, D. N. Theodorou, *Diffusion of small molecules in disordered media: study of the effect of kinetic and spatial heterogeneities* Chem. Eng. Sci. **56**, 2789-2801, 2001.
62. L. Garrido, M. López-González, E. Riande, *Influence of local chain dynamics on diffusion of gases in polymers as determined by pulsed field gradient NMR* J. Polym. Sci. B: Polym. Phys **48**, 231-235, 2010.
63. D. Fritsch, K. V. Peinemann, *Novel permeselective 6F-poly(amide imide)s as membrane host for nano-sized catalysts* J. Membr. Sci. **99**, 29-38, 1995.
64. C. Nagel, K. Gunther-Schade, D. Fritch, T. Strunskus, F. Faupel, *Free volume and transport properties in highly selective polymer membranes* Macromolecules **35**, 2071-2077, 2002.

Adaptive Resolution Molecular Dynamics: Extension to Quantum Problems

Luigi Delle Site

Institute for Mathematics
Freie Universität Berlin, Arnimallee 6, 14195, Berlin (Germany)
E-mail: luigi.dellesite@fu-berlin.de

Adaptive resolution Molecular Dynamics (MD) is here introduced within the AdResS (Adaptive Resolution Scheme) approach. This is a simulation method for MD that treats different regions with different levels of molecular resolution (i.e. molecular representation). AdResS was originally developed to couple (only) different classical descriptions of the system; instead, coupling a quantum with a classical resolution implies the solution of non trivial conceptual problems. Quantum mechanics is intrinsically a probabilistic theory while classical mechanics is deterministic, thus passing from one description to the other implies not only a change of the number of degrees of freedom (DOF) but also a change regarding the physical principles governing the evolution of the system. In this lecture I will discuss how for (at least) one class of problems the (classical) AdResS method can be extended to the quantum case in an almost straightforward way.

1 Introduction

Bridging scales in condensed matter requires the treatment of a different number (and different kind) of DOF corresponding to each scale. To this aim, in the field of molecular simulation, a large number of techniques have been developed in the last years; many are discussed in the lectures of this school. In these notes the focus is on *concurrent coupling*, that is all the scales (and their corresponding DOF) are treated at the same time within a unified computational approach. Actually here I will discuss an approach that goes beyond the standard concurrent coupling, which usually consists of interfacing regions of different resolution without free exchange of particles, and extends the coupling idea to a truly dynamical zooming in and zooming out on the system. Fig. 1 provides an example with a direct pictorial representation of the idea of zooming for the case of the adsorption of a macromolecule on a surface. When the molecule is far from the surface the relevant physical aspects are those related to the proper sampling of the conformational space of the molecular backbone. In this case a simplified coarse-grained molecular representation that reproduces the backbone properties is sufficient for the conformational sampling. However as the molecule goes closer to the surface the explicit chemical structure becomes important and one needs to zoom in (put the system under a magnification glass) at the contact region and have an explicit atomistic resolution. This process would then allow for the proper description of the chemical recognition between the molecule and the surface and at the same time will properly describe its connection to the conformational rearrangements of the rest of the molecule (at coarser scale). The idea of zooming requires an adaptive resolution simulation approach that allows to change, dynamically, on the fly, during the simulation, the number and/or kind of DOF as the molecule (or part of it) passes from the low resolution region to the high resolution region (region under magnification glass) and vice versa. In the next paragraphs of these notes the adaptive resolution method AdResS

will be briefly introduced and then its extension to quantum problems will be discussed in detail.

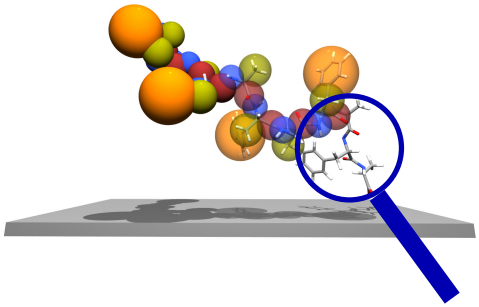


Figure 1. The zooming idea at the contact region. The magnification glass must be intended here as a pictorial representation of a computational tool which introduces explicit chemistry and atomistic structure where this is needed (surface-polymer contact region) while in parallel the evolution of the large scale conformational properties of the polymer takes place.

2 The AdResS Method

A review chapter about the (classical) AdResS method has already appeared in the lecture notes of the school of 2009[1], for this reason here I will report only the basic conceptual and technical aspects to allow the reader to have a sufficient understanding of the method and to proceed without (necessarily) consulting the previous notes (except for specific technical details, if of interest). Moreover, in the last years further developments have been done in the refinement of the approach and they were not reported in Ref. 1; for this reason here I will briefly discuss these improvement to the method and provide the related references.

The essential conceptual/methodological aspect of AdResS are the following:

- It changes the molecular resolution in a subregion of the space while the rest of the system stays at lower resolution.
- It allows for free (i.e. not externally imposed, e.g. Monte Carlo insertion or removal of particles) exchange of molecules from the high resolution to the low resolution region and vice versa.
- The process above occurs under conditions of thermodynamic equilibrium, which means same (average) particle density, temperature and pressure in all regions.
- Density, temperature and pressure must be the same as that of a reference full high resolution simulation. This preserves the *true"* thermodynamic state point.

The general idea is that of having an on-the-fly interchange between atomistic and coarse-grained description with a two stage procedure. The first stage consists of developing an effective, coarse-grained pair potential U^{cm} from the reference all atom simulation.

The second step consists of coupling the atomistic and coarse-grained resolution via an interpolation formula on the forces:

$$\mathbf{F}_{\alpha\beta} = w(X_\alpha)w(X_\beta)\mathbf{F}^{atom}_{\alpha\beta} + [1 - w(X_\alpha)w(X_\beta)]\mathbf{F}^{cm}_{\alpha\beta} \quad (1)$$

Here α and β indicate the two molecules, $\mathbf{F}^{atom}_{\alpha\beta}$ is the force obtained from the atomistic potential, $\mathbf{F}^{cm}_{\alpha\beta}$ is the force derived from the coarse-grained potential. X_α and X_β are the spatial coordinates of the center of mass of respectively the molecule α and β. $w(x)$ is a multiplicative function with value zero in the coarse-grained region and one in the atomistic region; it is then smooth and monotonic in an intermediate region Δ (region of atomistic/coarse-grained hybrid resolution). Fig. 2 provides the pictorial representation of the idea for a test molecule (tetrahedral molecule). On the left the coarse grained region, at the center indicated by Δ, the transition region with spatial-dependent hybrid resolution according to $w(x)$ and on the right the atomistic region. According to this set up, two

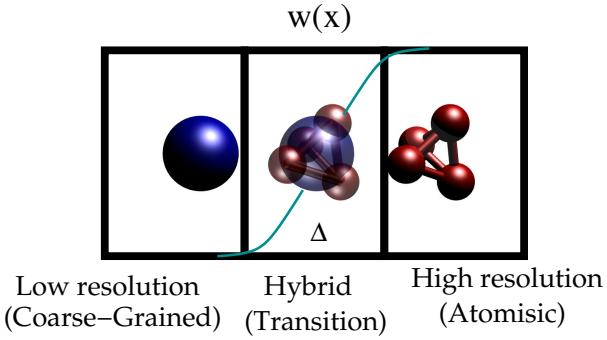


Figure 2. Pictorial representation of the adaptive idea for a tetrahedral molecules. Figure adapted from Ref. 5.

atomistic molecules interact as atomistic, coarse-grained molecules interacts with all the others as coarse-grained pairs (coarse-grained molecules do not have any atomistic degrees of freedom), while for the other cases molecules interact according to their coupled value of $w(X_\alpha)w(X_\beta)$ with hybrid resolutions. This means that a molecule which goes from the atomistic to the coarse-grained region, slowly looses its atomistic degrees of freedom (rotations and vibrations) and becomes an effective sphere going through a continuous stage of hybrid resolutions in Δ. The same process but in opposite direction (the molecule acquires DOF) for a coarse-grained molecule moving towards the atomistic region. In addition, the use of a locally acting thermostat ensures basic thermodynamic equilibrium, so that the reintroduced DOF are thermalized properly[2–4]. As anticipated before, this part of the method with specific technical details is reported in the notes of the previous school[1], instead the part discussed below, concerning a further refinement of the way to obtain thermodynamic equilibrium, has been developed in the last two years. Though the thermodynamic equilibrium provided by the coupling to the thermostat is numerically satisfying, however a small drop of particle density in the transition region cannot be avoided (see e.g.[5,6,4]). A theoretical analysis of the method suggested that the chemical potential which characterizes each resolution, μ_w, is not the same for all values of w and thus it is likely

to produce the drop of density[4,7]. Since the approach is clearly non Hamiltonian[8], here it must be specified what is intended for *chemical potential of each resolution*: **it means the chemical potential, that the system would have if the overall resolution (of the entire box) was a specific, fixed, one** $\bar{w}$. According to this idea let us define: $\phi = \mu_{atom} - \mu(\bar{w})$, and imagine to calculate μ , hypothetically, for each fixed value of $\bar{w}$ between zero and one. In this way we can actually write: $\phi(x) = \mu_{atom} - \mu_w(x)$, that is the difference between the chemical potential of the atomistic resolution and that corresponding to the resolution w at the position x; this quantity in good approximation (see[7]) should compensate the thermodynamic unbalance which produces the drop of density. Indeed, a term proportional to $\nabla_x \phi(x)$ which acts on the center of mass of the molecules, added to the force of Eq. 1, has been shown to remove the problem of drop of density[7]. This addition allows also for a generalization of AdResS as a method that can couple any two (or more) molecular representations; for example two different atomistic force fields which have anyway the same number of DOF. The derivation of $\phi(x)$ has been improved further in terms of compensating pressure and led to the definition of an *effective Grand Canonical* set up for general open systems MD simulations. In practical terms this means that the significantly extended coarse-grained region plays the role of reservoir of molecules for a (usually) smaller atomistic region[9]. Finally, the natural question for AdResS as a molecular dynamics scheme is whether the force of Eq. 1 can be somehow conservative; as anticipated, the method is non Hamiltonian and thus the answer is negative. In fact, despite a different claim[10–13], it has been shown both analytically[8] and numerically[14] that within this scheme there is no possibility of deriving Eq. 1 from **any** potential. The method has been shown to be numerically and conceptually robust in a large number of applications for classical systems (see e.g. Refs. 15–18 and references therein), and the natural question raising at this point regards its applicability to quantum system. This is the subject of next paragraphs.

3 Quantum-Classical Adaptive Resolution: The Conceptual Problem

As anticipated above, while changing number and kind of DOF for classical problems can be achieved by coupling regions governed by the same physical principles and equations, the same cannot be done in a straightforward way when the coupling involves quantum resolution. For the classical case one has to take care that Newton's laws of mechanics and the resulting thermodynamic equilibrium are consistent between the different regions and that the transition region, with the change of resolution, does not perturb this consistency and does not introduce artifacts. Instead, when one couples a quantum region with a classical one, allowing for the free exchange of molecules, the situation is by far more complicated. In fact the meaning of thermodynamic equilibrium can be interpreted differently in the quantum and in the classical region, and, above all, one has Newton's equations in one region and Schrödinger equation (or similar ones, e.g. Kohn-Sham equation in Density Functional Theory) in the quantum region. For the quantum-classical case the dynamical coupling of Eq. 1 is obviously not straightforward anymore. In particular, in case of electronic resolution (for the quantum region), the essential problem can be summarized as: "*how to slowly switch on and off an electron in a physical consistent way*". Electrons are not localized particles, and are characterized by complex (long-range) correlations thus the appearance or disappearance of an electron changes the entire electronic spectrum in the quantum region. The question then is how to slowly introduce or remove an electron so

that the electronic spectrum (of the other electrons) is (at least approximatively) the same as if the whole system was treated at quantum level. From the practical point of view, the question above may be reduced to the search of an equivalent concept as that related to $\phi(x)$ for the classical case, that is a proper statistical way to introduce and remove an electron so that the statistical properties of the electrons in the quantum region are equivalent (or at least close enough) to those one would find in a full quantum treatment of the entire system; solutions along this directions are still missing and would be highly welcome. One must also notice that what is named in current literature a quantum-classical adaptive coupling based on electronic structure[11,19,20], it is actually not a proper quantum-classical concurrent coupling. In fact in such cases the electronic calculations are employed as a separated step to obtain, in the *"quantum region"*, a reasonable, on-the-fly, classical force field which is then interfaced with a standard classical force field. In our case, instead, for quantum-classical adaptive method is intended a method to study in detail the electronic properties in the small quantum region. Thus it refers to problems where electronic properties are of major interest and must be properly reproduced. While for electrons there seems to be no simple solution, for other quantum problems the scheme of AdResS may still be used though within the correct interpretation of the results. This is the case for the treatment via MD of the spatial delocalization of light atoms within the framework of Path Integral (PI) of Richard Feynman[21]. In the next section I will illustrate the basic idea of PIMD representation of atoms and its related principles.

4 Path Integral Molecular Dynamics

When the de Broglie wavelength, $\Lambda = \frac{h}{\sqrt{2\pi m k_B T}}$, with h being Planck's constant, m the mass of the particle (atom), k_B Boltzmann constant and T the temperature, is much smaller than the particle-particle distance, the system can be safely considered classical. Instead when Λ is larger than the particle-particle distance then the quantum character dominates and classical statistical mechanics no more applies; particles, are no more classical localized objects and must be considered according to their symmetry as fermions or bosons. When Λ is of the same order of the particle-particle distance, quantum effects of spatial delocalization play a major role but their nature of fermions or bosons is not necessarily relevant and can be ignored; here I will consider this latter case. Light atoms, as hydrogen, given the small mass, are characterized by sizeable quantum effects, due to the spatial delocalization, even at room temperature. This quantum character can be described via the PIMD approach; here I will report the basic notions and refer for a more complete description to Refs. 22, 23.

Consider the Hamiltonian of N distinguishable particles:

$$H = \sum_{I=1}^{N} \frac{P_I^2}{2M_I} + V(R_1, \ldots, R_N). \tag{2}$$

The related density matrix in the representation of spacial position is:

$$\rho(\mathbf{R}, \mathbf{R'}; \beta) = \langle \mathbf{R} | e^{-\beta H} | \mathbf{R'} \rangle, \tag{3}$$

with $\beta = 1/k_B T$. The quantum mechanical partition function corresponds to the trace of

the density matrix:

$$Z \equiv Tr(e^{-\beta H}) = \int d\mathbf{R} \langle \mathbf{R}|e^{-\beta H}|\mathbf{R}\rangle \tag{4}$$

Next, Trotter theorem[24] is used to factorize $e^{-\beta H}$ in a kinetic and a potential term:

$$e^{-\beta(K+V)} = \lim_{n\to\infty} \left[e^{\frac{-\beta}{2n}V} e^{\frac{-\beta}{n}K} e^{\frac{-\beta}{2n}V} \right]^n \tag{5}$$

here K is the kinetic and V is the potential operator. Substituting the expression above into Eq. 4 and making use of the definition of the identity operator, $\int |\mathbf{R}\rangle\langle\mathbf{R}| d\mathbf{R}$, n-1 times leads to:

$$\begin{aligned} Z = \lim_{n\to\infty} \int d\mathbf{R}^{(1)} \ldots d\mathbf{R}^{(n)} \langle \mathbf{R}^{(1)}| \left[e^{\frac{-\beta}{2n}V} e^{\frac{-\beta}{n}K} e^{\frac{-\beta}{2n}V} \right] |\mathbf{R}^{(2)}\rangle \ldots \\ \langle \mathbf{R}^{(i)}| \left[e^{\frac{-\beta}{2n}V} e^{\frac{-\beta}{n}K} e^{\frac{-\beta}{2n}V} \right] |\mathbf{R}^{(i+1)}\rangle \ldots \langle \mathbf{R}^{(n)}| \left[e^{\frac{-\beta}{2n}V} e^{\frac{-\beta}{n}K} e^{\frac{-\beta}{2n}V} \right] |\mathbf{R}^{(1)}\rangle. \end{aligned} \tag{6}$$

Since the potential is diagonal in the space representation $|\mathbf{R}\rangle$, each matrix element becomes:

$$\langle \mathbf{R}^{(i)}|e^{\frac{-\beta}{2n}V} e^{\frac{-\beta}{n}K} e^{\frac{-\beta}{2n}V}|\mathbf{R}^{(i+1)}\rangle = e^{\frac{-\beta}{2n}V(\mathbf{R}^{(i)})} \langle \mathbf{R}^{(i)}|e^{\frac{-\beta}{n}K}|\mathbf{R}^{(i+1)}\rangle e^{\frac{-\beta}{2n}V(\mathbf{R}^{(i+1)})}. \tag{7}$$

Next one employs the identity operator in momentum space, $\int |\mathbf{P}\rangle\langle\mathbf{P}| d\mathbf{P}$, the remaining matrix elements can be written as:

$$\begin{aligned} \langle \mathbf{R}^{(i)}|e^{\frac{-\beta}{n}K}|\mathbf{R}^{(i+1)}\rangle &= \int d\mathbf{P} \langle \mathbf{R}^{(i)}|\mathbf{P}\rangle\langle\mathbf{P}|e^{\frac{-\beta}{n}K}|\mathbf{R}^{(i+1)}\rangle \\ &= \int d\mathbf{P} \langle \mathbf{R}^{(i)}|\mathbf{P}\rangle\langle\mathbf{P}|\mathbf{R}^{(i+1)}\rangle e^{-\beta\mathbf{P}^2/(2Mn)} \end{aligned} \tag{8}$$

this can be simplified by using the projection of a momentum eigenstate on a position eigenstate

$$\langle \mathbf{R}|\mathbf{P}\rangle = \frac{1}{\sqrt{2\pi\hbar}} e^{i\mathbf{P}\cdot\mathbf{R}/\hbar}. \tag{9}$$

The introduction of Eq. 9 into Eq. 8 leads to:

$$\langle \mathbf{R}^{(i)}|e^{\frac{-\beta}{n}T}|\mathbf{R}^{(i+1)}\rangle = (\frac{Mn}{2\pi\beta\hbar^2})^{1/2} e^{-\frac{Mn}{2\pi\beta\hbar^2}(\mathbf{R}^{(i)}-\mathbf{R}^{(i+1)})^2}. \tag{10}$$

Introducing the results above into the Eq. 6 one obtains:

$$\begin{aligned} Z = \lim_{n\to\infty} \left[\prod_{I=1}^{N} (\frac{Mn}{2\pi\beta\hbar^2})^{n/2} \int dR_I^{(1)} \ldots dR_I^{(n)} \right] \times \\ e^{-\beta \sum_{I=1}^{N} \sum_{s=1}^{n} \frac{1}{2} M_I \omega_n^2 (R_I^{(s)} - R_I^{(s+1)})^2 + \frac{1}{n} V(\{R_I^{(s)}\})}, \end{aligned} \tag{11}$$

where the effective Hamiltonian is given by

$$\mathcal{H}_n = \sum_{I=1}^{N} \sum_{s=1}^{n} \frac{1}{2} M_I \omega_n^2 (R_I^{(s)} - R_I^{(s+1)})^2 + \frac{1}{n} V(\{R_I^{(s)}\}). \tag{12}$$

The Hamiltonian of Eq. 12 is formally equivalent to the Hamiltonian of N classical ring-polymers consisting of n beads each connected by harmonic springs and with a polymer-polymer interaction as illustrated in Fig. 3. The bead-bead interaction between different polymers is attenuated by a factor $1/n$. $\omega_p = m\sqrt{P}(k_B T)/\hbar$ is the frequency of the ring-polymer, n is the Trotter number, T the temperature and M is the mass of the particle (e.g. atom). The higher the Trotter number the better is the quantum description. The

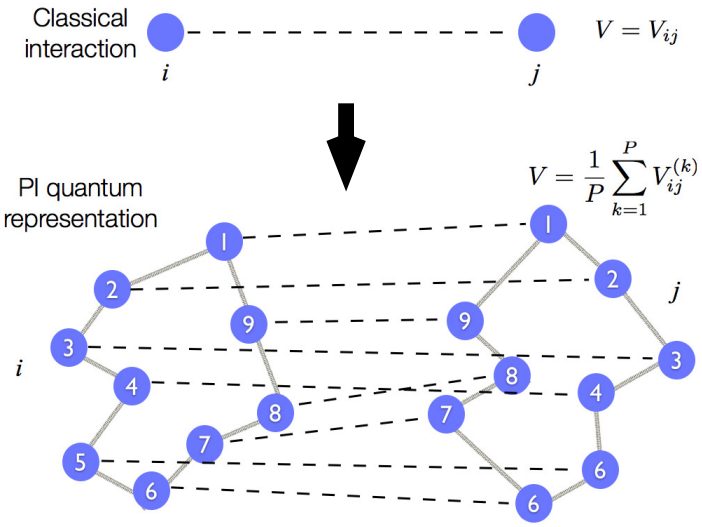


Figure 3. Classical and Path integral representation of atoms. In the quantum treatment the rigid spherical representation (classical) is substituted by fluctuating polymer rings.

calculation of Z via the Hamiltonian of Eq. 12 requires a sampling of the configurational space of the N ring-polymers. In order to devise a molecular dynamics scheme which allows for the sampling, and thus for calculating Z, one has to add n-Gaussian integrals in the momentum space to Eq. 11,

$$Z = \lim_{n\to\infty}\left[\prod_{I=1}^{N} W \int dR_I^{(1)} \dots dR_I^{(n)} \int dP_I^{(1)} \dots dP_I^{(n)}\right] \times$$
$$e^{-\beta \sum_{I=1}^{N}\sum_{s=1}^{n} \frac{\left[P_I^{(s)}\right]^2}{2M_I'} + \frac{1}{2}M_I\omega_n^2 (R_I^{(s)} - R_I^{(s+1)})^2 + \frac{1}{n}V(\{R_I^{(s)}\})}, \qquad (13)$$

where W is a proper normalization factor, M_I' is a fictitious mass of the beads. The momenta $\mathbf{P}_I$ are also fictitious quantities without physical meaning and allow to formally map the static problem of the interacting ring-polymers into a dynamical sampling. The Hamiltonian for the molecular dynamics scheme is then:

$$\mathcal{H}_n(\mathbf{R},\mathbf{P}) = \sum_{I=1}^{N}\sum_{s=1}^{n} \frac{\left[P_I^{(s)}\right]^2}{2M_I'} + \frac{1}{2}M_I\omega_n^2 (R_I^{(s)} - R_I^{(s+1)})^2 + \frac{1}{n}V(\{R_I^{(s)}\}) \qquad (14)$$

The procedure described above is commonly known as the path integral molecular dynamics (PIMD) in the real space. Within PIMD the ring-polymer dynamics can be employed to evaluate the expectation value of any observable A:

$$\langle A \rangle = \lim_{n\to\infty} \left[\prod_{I=1}^{N} W \int dR_I^{(1)} \dots dR_I^{(n)} \int dP_I^{(1)} \dots dP_I^{(n)} \right] \times e^{-\beta \mathcal{H}_n(\mathbf{R},\mathbf{P})} A_n(\mathbf{R}) \quad (15)$$

where A_n is calculated in a statistical sense, by averaging over the ring-polymer trajectories:

$$A_n(\mathbf{R}) = \frac{1}{n} \sum_{s=1}^{n} A(R_1^{(s)}, \dots, R_N^{(s)}) \quad (16)$$

5 Quantum-Classical Adaptive Resolution Simulation via PIMD

According to the paragraph above, the quantum statistical properties of a system can be in an **effective** way described by classical objects (beads of a classical ring-polymer) where the proper statistical sampling can be obtained by averaging over the trajectories of these objects which in turn are governed by Newton's equations. In this perspective, a quantum-classical adaptive is equivalent to couple two regions with a different number of effective **"classical" degrees of freedom** and thus the AdResS method applies straightforwardly. However, before proceeding with few more technical details it must be underlined that conceptually the quantum-classical adaptive is different from the classical one. In fact while in the classical case one may also consider (at least local) dynamical properties of the molecule, for the quantum-classical case the dynamic evolution in the quantum region must be interpreted only as a useful technical tool for sampling; it does not imply any physical meaning about the dynamical evolution of the molecule. In practice the fictitious dynamics of the ring-polymers makes compatible, from the technical point of view, the deterministic character of the classical region with the probabilistic character of the quantum region in calculating static average properties of the system.

Technically the coupling works as in Eq. 1, which in this case reads:

$$\mathbf{F}_{\alpha\beta} = w(X_\alpha) w(X_\beta) \mathbf{F}_{\alpha\beta}^{quant} + [1 - w(X_\alpha) w(X_\beta)] \, \mathbf{F}_{\alpha\beta}^{cg}. \quad (17)$$

$\mathbf{F}_{\alpha\beta}^{quant} = \sum_{i\alpha, i\beta} \mathbf{F}_{i\alpha, i\beta}$ is the total force acting between two polymer rings (thus expressing the quantum character of the particle). This force is derived from a given (classical) potential as illustrated in Fig. 3; the index $i\alpha$ ($i\beta$) identifies the i-th bead of ring α (β). $F_{\alpha\beta}^{cg}$ is instead the force obtained from the atomistic (or even coarse grained) potential and acts on the center of the atom (center of the molecule in case of coarse-grained representation) without being distributed among the beads of the polymer. As in the classical case, a locally acting thermostat takes care of properly thermalizing the reinserted DOF (beads) so that the whole system is in thermodynamic equilibrium, and a thermodynamic, force acting on the center of mass of the polymers, based on the standard calculation of $\phi(x)$, can be introduce to improve the conditions of equilibrium. This idea have been implemented for the test system of a liquid of tetrahedral molecules[25]. In this case, as shown in Fig. 4, we were able to couple a path integral representation of the atoms of the molecules directly to a spherical classical coarse-grained molecular representation; the coarse-grained model

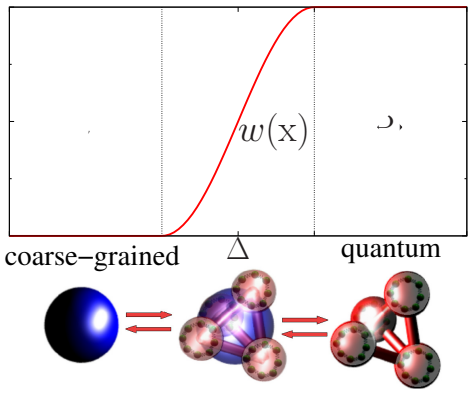


Figure 4. Schematic representation of the adaptive resolution for the tetrahedral molecule. In the quantum region, (right) each atom is represented by a polymer ring with classical beads, in the Δ region the molecule has hybrid coarse-grained/quantum resolution and in the coarse-grained region the molecule is represented by an effective spherical model obtained from the reference full path integral simulation. Figure adapted from Ref. 25.

was derived from a reference full quantum simulation. In Ref. 25 it has been shown that the various radial distribution functions (RDFs) are the same as in the reference full quantum simulation, and in particular the bead-bead RDF in the quantum region agrees reasonably well with the RDF of the equivalent region in the full quantum simulation. This is very important because the bead-bead RDF directly expresses the quantum (spatial) statistical properties of the system and thus the agreement with the results of the reference full quantum simulation shows that the coupling to a classical system (acting as a thermodynamic bath) does not destroy the quantum character of the atoms in the quantum region (at least for spatial properties). A satisfying agreement, within a difference of 5%, has been found for the particle density and has been show that the exchange of molecules between the different regions takes place in a proper way. Later on, a further, more critical test, was provided by the adaptive resolution simulation study of liquid parahydrogen. Here given the extreme thermodynamic conditions (low temperature and zero pressure), the adaptive idea was highly challenged, not only for the coupling scheme but also regarding the technical derivation of reasonable coarse-grained potentials. Moreover, this system has been largely studied employing the PIMD approach and thus relevant reference data for benchmarking the adaptive method are available in literature. The results have shown that the AdResS method works quite well in an extended range of temperature (above 14 Kelvin) and densities[26,27]. At temperatures below 14 Kelvin the bosonic character of the molecule becomes important, the, so called, exchange interactions become relevant and these cannot be described by the current formulation of AdResS[27]. Despite the positive outcome of these studies, there is a conceptual point that still needs to be clarified. In fact as we have seen in the paragraph dedicated to the theoretical derivation of the PIMD formalism, the MD for these systems is a tool to sample a Canonical partition function and thus a Boltzmann factor, that is a quantity which involves a Hamiltonian. It seems to be a conceptual

contradiction between the idea of AdResS, based on non conservative forces (and thus non Hamiltonian) and the idea of PIMD, but at the same time, as a matter of fact, the various RDF of the adaptive simulations agree reasonably well with those of the reference full quantum system and have a form typical for a Canonical or Grand Canonical ensemble. This apparent contradiction disappears if one interprets the quantum region as an effective Grand Canonical ensemble as in Ref. 9, meaning that the classical region acts only as a large (in principle infinite) reservoir of molecules in thermodynamic equilibrium with the smaller region. In this case the Hamiltonian of the quantum region depends on its instantaneous number N of molecules and counts all the corresponding pair interactions between the N molecules. In this perspective, the interaction of the molecules of the quantum region with those of the rest of the system plays, effectively, the role of coupling term to a generic particle reservoir. Since in this case MD is only a tool for a dynamical sampling of the N space, this interaction can be formally ignored in the Hamiltonian of the quantum region. In any case, such a term is in part slowly switched off by the transition region and in part disappears due to the finite (short) range of the interactions, thus it represents (at the worst) only a small perturbation. In this context the adaptive PIMD approach, as a matter of fact, samples the different realizations of N and its corresponding configurational space, thus samples an effective Grand Canonical partition function, for which the PI approach is justified[28].

6 Conclusions and Perspectives

In these notes the extension of the AdResS method to quantum mechanical problems within the PIMD approach have been discussed. The application to the test case of a liquid of tetrahedral molecules has given satisfying results regarding the technical and conceptual aspects. Later on, the application to challenging *real* physical systems, as the molecular liquid of parahydrogen, has provided further evidence of the robustness of the approach. Future work is proceeding along the study of the basic quantum effects of hydrogen (proton) delocalization in liquid water and its consequences for the solvation process of large molecules and their conformational properties. Quantum effects of proton delocalization may be small, however could be crucial in many situations; in this context the AdResS method provides an efficient computational tool to study large systems at reasonable computational cost. At the same time, from the conceptual point of view, since it allows to identify the essential DOF of the system, AdResS can be used to understand the delicate interplay between different scales which usually it is not straightforward for quantum systems.

Acknowledgments

I would like to thank all the collaborators of the AdResS project, in particular my coworkers M. Praprotnik, K. Kremer, C. Junghans, S. Poblete, A. Poma, S. Fritsch, R. Potestio, D. Mukherji, C. Clementi and G. Ciccotti for the many interesting discussions. The funding of the Deutsche Forschungsgemeinschaft (DFG, German Science Foundation) within the Heisenberg Program is also acknowledged.

References

1. C. Junghans, M. Praprotnik and L. Delle Site, *Adaptive Resolution Schemes* in *Multiscale Simulation Methods in Molecular Sciences*, edited by J. Grotendorst, N. Attig, S. Blügel and D. Marx, NIC Series Volume 42, ISBN 978-3-9810843-8-2, Jülich 2009.
2. M. Praprotnik, K. Kremer, and L. Delle Site, *Adaptive molecular resolution via a continuous change of the phase space dimensionality* Phys. Rev. E **75**, 017701, 2007.
3. M. Praprotnik, K. Kremer, and L. Delle Site, *Fractional dimensions of phase space variables: A tool for varying the degrees of freedom of a system in a multiscale treatment* J. Phys. A: Math. Gen. **40**, F281, 2007.
4. M. Praprotnik, L. Delle Site, and K. Kremer, *Multiscale Simulation of Soft Matter: From Scale Bridging to Adaptive Resolution* Annu. Rev. Phys. Chem. **59**, 545, 2008.
5. M. Praprotnik, L. Delle Site, and K. Kremer, *Adaptive resolution molecular-dynamics simulation: Changing the degrees of freedom on the fly* J. Chem. Phys. **123**, 224106, 2005.
6. M. Praprotnik, L. Delle Site, and K. Kremer, *Adaptive Resolution Scheme (AdResS) for Efficient Hybrid Atomistic/Mesoscale Molecular Dynamics Simulations of Dense Liquids* Phys. Rev. E **73**, 066701, 2006.
7. S. Poblete, M. Praprotnik, K. Kremer and L. Delle Site, *Coupling different levels of resolution in molecular simulations* J. Chem. Phys. **132**, 114101, 2010.
8. L. Delle Site, *Some fundamental problems for an energy-conserving adaptive-resolution molecular dynamics scheme* Phys. Rev. E **76**, 047701, 2007.
9. S. Fritsch, S. Poblete, C. Junghans, G. Ciccotti, L. Delle Site and K. Kremer, *Grand canonical Molecular Dynamics Simulations* arXiv:1112.3151v1, 2011.
10. B. Ensing, S. O. Nielsen, P. B. Moore, M. L. Klein, and M. Parrinello, *Energy conservation in adaptive hybrid atomistic/coarse-grain molecular dynamics* J. Chem. Theor. Comp. **3**, 1100, 2007.
11. R. Bulo, B. Ensing, J. Sikkema and L. Visscher, *Energy conservation in adaptive hybrid atomistic/coarse-grain molecular dynamics* J. Chem. Theor. Comp. **5**, 2212, 2009.
12. S. O. Nielsen, P. B. Moore and B. Ensing, *Adaptive Multiscale Molecular Dynamics of Macromolecular Fluids* Phys. Rev. Lett. **105**, 237802, 2010.
13. S. O. Nielsen, P. B. Moore and B. Ensing, *Reply comment to: Adaptive Multiscale Molecular Dynamics of Macromolecular Fluids* Phys. Rev. Lett. **107**, 099801, 2011.
14. M. Praprotnik, S. Poblete, L. Delle Site and K. Kremer, *Comment to: Adaptive Multiscale Molecular Dynamics of Macromolecular Fluids* Phys. Rev. Lett. **107**, 099802, 2011.
15. M. Praprotnik, S. Matysiak, L. Delle Site, K. Kremer and C. Clementi, *Adaptive resolution simulation of liquid water* J. Phys. Cond. Matt. **19**, 292201, 2007.
16. S. Matysiak, C. Clementi, M. Praprotnik, K. Kremer, and L. Delle Site, *Modeling Diffusive Dynamics in Adaptive Resolution Simulation of Liquid Water* J. Chem. Phys. **128**, 024503, 2008.
17. B. P. Lambeth, C. Junghans, K. Kremer, C. Clementi, and L. Delle Site, *On the Locality of Hydrogen Bond Networks at Hydrophobic Interfaces* J. Chem. Phys. **133**, 221101, 2010.

18. D. Mukherji, N. van der Vegt, K. Kremer and L. Delle Site, *Kirkwood-Buff analysis of liquid mixtures in an open boundary simulation* submitted to J.Chem.Theor.Comp.
19. G. Csanyi, T. Albaret, M. C. Payne, and A. DeVita, *"Learn on the Fly": A Hybrid Classical and Quantum-Mechanical Molecular Dynamics Simulation* Phys. Rev. Lett. **93**, 175503, 2004.
20. N. Bernstein, C. Varnai, I. Solt, S .A .Winfield, M. C. Payne, I. Simon, M. Fuxreiter and G. Csanyi, *QM/MM simulation of liquid water with an adaptive quantum region* Phys. Chem. Chem. Phys. **14**, 646, 2012.
21. R. P. Feynman and A. R. Hibbs, *Quantum Mechanics and Path Integrals*, McGraw-Hill, New York, 1965.
22. M. Tuckerman, *Path Integration via Molecular Dynamics* in *Quantum Simulations of Complex Many-Body Systems: From Theory to Algorithms*, edited by Johannes Grotendorst, Dominik Marx, Alejandro Muramatsu, NIC Series Volume 10, Jülich 2002.
23. M. Tuckerman, *Statistical Mechanics: Theory and Molecular Simulation* Oxford University Press, 2010.
24. L. Schulman, in *Techniques and Application of Path Integration* Wiley & Sons, New York, 1965.
25. A. B. Poma and L. Delle Site, *Classical to Path-Integral Adaptive Resolution in Molecular Simulation: Towards a Smooth Quantum-Classical Coupling* Phys. Rev. Lett. **104**, 250201, 2010.
26. A. B. Poma and L. Delle Site, *Adaptive Resolution Simulation of Liquid Para-Hydrogen: Testing the robusteness of the Quantum-Classical Adaptive coupling* Phys. Chem. Chem. Phys. **13**, 10510, 2011.
27. R. Potestio and L. Delle Site, *Quantum Locality and Equilibrium Properties in Low-temperature Parahydrogen: A Multiscale Simulation Study* J. Chem. Phys. 2012 in press.
28. Q. Wang, J. K. Johnson, and J. Q. Broughton, *Path Integral Grand Canonical Monte Carlo* J. Chem. Phys. **107**, 5108, 1997.

Coupling Molecular Dynamics and Lattice Boltzmann to Simulate Brownian Motion with Hydrodynamic Interactions

Burkhard Dünweg

Max Planck Institute for Polymer Research
Ackermannweg 10, 55128 Mainz, Germany
and
Department of Chemical Engineering
Monash University, Melbourne, VIC 3800, Australia
E-mail: duenweg@mpip-mainz.mpg.de

In soft–matter systems where Brownian constituents are immersed in a solvent, both thermal fluctuations and hydrodynamic interactions are important. The article outlines a general scheme to simulate such systems by coupling Molecular Dynamics for the Brownian particles to a lattice Boltzmann algorithm for the solvent. By the example of a polymer chain immersed in solvent, it is explicitly demonstrated that this approach yields (essentially) the same results as Brownian Dynamics.

1 Introduction

Remark: The present contribution intends to just give a very brief overview over the subject matter. It is an updated version of a similar article[1] that the author has written on occasion of the 2009 NIC winter school. For more detailed information, the reader is referred to a longer review article, Ref. 2. —

Many soft–matter systems are comprised of Brownian particles immersed in a solvent. Prototypical examples are colloidal dispersions and polymer solutions, where the latter, in contrast to the former, are characterized by non–trivial internal degrees of freedom (here: the many possible conformations of the macromolecule). Fundamental for these systems is the separation of length and time scales between "large and slow" Brownian particles, and "small and fast" solvent particles. "Mesoscopic" simulations focus on the range of length and time scales which are, on the one hand, too small to allow a description just in terms of continuum mechanics of the overall system, but, on the other hand, large enough to allow the replacement of the solvent by a hydrodynamic continuum. This latter approximation is much less severe than one would assume at first glance; detailed Molecular Dynamics simulations have shown that hydrodynamics works as soon as the length scale exceeds a few particle diameters, and the time scale a few collision times.

To simulate such systems consistently, one has to take into account that the length and time scales are so small that thermal fluctuations cannot be neglected. The "Boltzmann number" Bo (a term invented by us) is a useful parameter for quantifying how important fluctuations are. Given a certain spatial resolution b (for example, the lattice spacing of a grid which is used to simulate the fluid dynamics), we may ask ourselves how many solvent particles N_p correspond to the scale b. On average, this is given by $N_p = \rho b^3/m_p$, where ρ is the mass density and m_p the mass of a solvent particle (and we assume a three–

dimensional system). The relative importance of fluctuations is then given by

$$Bo = N_p^{-1/2} = \left(\frac{m_p}{\rho b^3}\right)^{1/2}. \tag{1}$$

It should be noted that for an ideal gas, where the occupation statistics is Poissonian, Bo is just the relative statistical inaccuracy of the random variable N_p. In soft–matter systems, b is usually small enough such that Bo is no longer negligible.

Furthermore, *hydrodynamic interactions* must be modeled. In essence, this term refers to dynamic correlations between the Brownian particles, mediated by fast momentum transport through the solvent. The separation of time scales can be quantified in terms of the so–called Schmidt number

$$Sc = \frac{\eta_{kin}}{D}, \tag{2}$$

where $\eta_{kin} = \eta/\rho$ is the kinematic viscosity (ratio of dynamic shear viscosity η and mass density ρ) of the fluid, measuring how quickly momentum propagates diffusively through the solvent, and D is the diffusion constant of the particles. Typically, in a dense fluid $Sc \sim 10^2 \ldots 10^3$ for the solvent particles, while for large Brownian particles Sc is even much larger. Finally, we may also often assume that the solvent dynamics is in the creeping–flow regime, i. e. that the Reynolds number

$$Re = \frac{ul}{\eta_{kin}}, \tag{3}$$

where u denotes the velocity of the flow and l its typical size, is small. This is certainly true as long as the system is not driven strongly out of thermal equilibrium.

These considerations lead to the natural (but, in our opinion, not always correct) conclusion that the method of choice to simulate such systems is Brownian Dynamics (BD)[3]. Here the Brownian particles are displaced under the influence of particle–particle forces, hydrodynamic drag forces (calculated from the particle positions), and stochastic forces representing the thermal noise. However, the technical problems to do this efficiently for a large number N of Brownian particles are substantial. The calculation of the drag forces involves the evaluation of the hydrodynamic Green's function, which depends on the boundary conditions, and has an intrinsically long–range nature (such that all particles interact with each other). Furthermore, these drag terms also determine the correlations in the stochastic displacements, such that the generation of the stochastic terms involves the calculation of the matrix square root of a $3N \times 3N$ matrix. Recently, there has been substantial progress in the development of fast algorithms[4]; however, currently there are only few groups who master these advanced and complicated techniques. Apart from this, the applicability is somewhat limited, since the Green's function must be re–calculated for each new boundary condition, and its validity is questionable if the system is put under strong nonequilibrium conditions like, e. g., a turbulent flow – it should be noted that the Green's function is calculated for low–Re hydrodynamics.

Therefore, many soft–matter researchers have rather chosen the alternative approach, which is to simulate the system including the solvent degrees of freedom, with explicit momentum transport. The advantage of this is a simple algorithm, which scales linearly with the number of Brownian particles, and is easily parallelizable, due to its locality. The disadvantage, however, is that one needs to simulate many more degrees of freedom than

those in which one is genuinely interested – *and* to do this on the short inertial time scales in which one is not interested either. It is clear that such an approach involves essentially Molecular Dynamics (MD) for the Brownian particles.

Many ways are possible how to simulate the solvent degrees of freedom, and how to couple them to the MD part. It is just the universality of hydrodynamics that allows us to invent many models which all will produce the correct physics. The requirements are rather weak – the solvent model has to just be compatible with Navier–Stokes hydrodynamics on the macroscopic scale. Particle methods include Dissipative Particle Dynamics (DPD) and Multi–Particle Collision Dynamics (MPCD)[5], while lattice methods involve the direct solution of the Navier–Stokes equation on a lattice, or lattice Boltzmann (LB). The latter is a method with which we have made quite good experience, both in terms of efficiency and versatility. The efficiency comes from the inherent ease of memory management for a lattice model, combined with ease of parallelization, which comes from the high degree of locality: Essentially an LB algorithm just shifts populations on a lattice, combined with collisions, which however only happen locally on a single lattice site. The coupling to the Brownian particles (simulated via MD) can either be done via boundary conditions, or via an interpolation function that introduces a *dissipative* coupling between particles and fluid. In this article, we will focus on the latter method.

2 Coupling Scheme

As long as we view LB as just a solver for the Navier–Stokes equation, we may write down the equations of motion for the coupled system as follows:

$$\frac{d}{dt}\vec{r}_i = \frac{1}{m_i}\vec{p}_i, \tag{4}$$

$$\frac{d}{dt}\vec{p}_i = \vec{F}_i^c + \vec{F}_i^d + \vec{F}_i^f, \tag{5}$$

$$\partial_t \rho + \partial_\alpha j_\alpha = 0, \tag{6}$$

$$\partial_t j_\alpha + \partial_\beta \pi^E_{\alpha\beta} = \partial_\beta \eta_{\alpha\beta\gamma\delta}\partial_\gamma u_\delta + f^h_\alpha + \partial_\beta \sigma^f_{\alpha\beta}. \tag{7}$$

Here, $\vec{r}_i$, $\vec{p}_i$ and m_i are the positions, momenta, and masses of the Brownian particles, respectively. The forces $\vec{F}_i$ acting on the particles are conservative (c, i. e. coming from the interparticle potential), dissipative (d), and fluctuating (f). The equations of motion for the fluid have been written in tensor notation, where Greek indexes denote Cartesian components, and the Einstein summation convention is used. The first equation describes mass conservation; the mass flux $\rho\vec{u}$, where $\vec{u}$ is the flow velocity, is identical to the momentum density $\vec{j}$. The last equation describes the time evolution of the fluid momentum density. In the absence of particles, the fluid momentum is conserved. This part is described via the stress tensor, which in turn is decomposed into the conservative Euler stress $\pi^E_{\alpha\beta}$, the dissipative stress $\eta_{\alpha\beta\gamma\delta}\partial_\gamma u_\delta$, and the fluctuating stress $\sigma^f_{\alpha\beta}$. The influence of the particles is described via an external force density $\vec{f}^h$.

The coupling to a particle i is introduced via an interpolation procedure where first the flow velocities from the surrounding sites are averaged over to yield the flow velocity right

at the position of i. In the continuum limit, this is written as

$$\vec{u}_i \equiv \vec{u}(\vec{r}_i) = \int d^3\vec{r}\,\Delta(\vec{r},\vec{r}_i)\vec{u}(\vec{r}), \tag{8}$$

where $\Delta(\vec{r},\vec{r}_i)$ is a weight function with compact support, satisfying

$$\int d^3\vec{r}\Delta(\vec{r},\vec{r}_i) = 1. \tag{9}$$

Secondly, each particle is assigned a phenomenological friction coefficient Γ_i, and this allows us to calculate the friction force on particle i:

$$\vec{F}_i^d = -\Gamma_i\left(\frac{\vec{p}_i}{m_i} - \vec{u}_i\right). \tag{10}$$

A Langevin noise term $\vec{F}_i^f$ is added to the particle equation of motion, in order to compensate the dissipative losses that come from $\vec{F}_i^d$. $\vec{F}_i^f$ satisfies the standard fluctuation–dissipation relation

$$\left\langle F_{i\alpha}^f \right\rangle = 0, \tag{11}$$

$$\left\langle F_{i\alpha}^f(t)\, F_{j\beta}^f(t')\right\rangle = 2k_BT\Gamma_i\delta_{ij}\delta_{\alpha\beta}\delta(t-t'), \tag{12}$$

where T is is the absolute temperature and k_B the Boltzmann constant. While the conservative forces $\vec{F}_i^c$ conserve the total momentum of the particle system, as a result of Newton's third law, the dissipative and fluctuating terms ($\vec{F}_i^d$ and $\vec{F}_i^f$) do not. The associated momentum transfer must therefore have come from the fluid. The overall momentum must be conserved, however. This means that the force term entering the Navier–Stokes equation must just balance these forces. One easily sees that the choice

$$\vec{f}^h(\vec{r}) = -\sum_i\left(\vec{F}_i^d + \vec{F}_i^f\right)\Delta(\vec{r},\vec{r}_i) \tag{13}$$

satisfies this criterion. It should be noted that we use the *same* weight function to interpolate the forces back onto the fluid; this is necessary to satisfy the fluctuation–dissipation theorem for the overall system, i. e. to simulate a well–defined constant–temperature ensemble. The detailed proof of the thermodynamic consistency of the procedure can be found in Ref. 2.

We still need to specify the remaining terms in the Navier–Stokes equation. The viscosity tensor $\eta_{\alpha\beta\gamma\delta}$ describes an isotropic Newtonian fluid:

$$\eta_{\alpha\beta\gamma\delta} = \eta\left(\delta_{\alpha\gamma}\delta_{\beta\delta} + \delta_{\alpha\delta}\delta_{\beta\gamma} - \frac{2}{3}\delta_{\alpha\beta}\delta_{\gamma\delta}\right) + \eta_b\delta_{\alpha\beta}\delta_{\gamma\delta}, \tag{14}$$

with shear and bulk viscosities η and η_b. This tensor also appears in the covariance matrix of the fluctuating (Langevin) stress $\sigma_{\alpha\beta}^f$:

$$\left\langle \sigma_{\alpha\beta}^f \right\rangle = 0, \tag{15}$$

$$\left\langle \sigma_{\alpha\beta}^f(\vec{r},t)\,\sigma_{\gamma\delta}^f(\vec{r}',t')\right\rangle = 2k_BT\eta_{\alpha\beta\gamma\delta}\delta(\vec{r}-\vec{r}')\,\delta(t-t'). \tag{16}$$

Finally, the Euler stress

$$\pi^E_{\alpha\beta} = p\delta_{\alpha\beta} + \rho u_\alpha u_\beta \tag{17}$$

describes the equation of state of the fluid (p is the thermodynamic pressure), and convective momentum transport.

3 Low Mach Number Physics

At this point an important simplification can be made. The equation of state only matters for flow velocities u that are comparable with the speed of sound c_s, i. e. for which the Mach number

$$Ma = \frac{u}{c_s} \tag{18}$$

is large. In the low Mach number regime, the flow may be considered as effectively incompressible (although no incompressibility constraint is imposed in the algorithm). The Mach number should not be confused with the Reynolds number Re, which rather measures whether inertial effects are important. Now it turns out that essentially all soft–matter applications "live" in the low–Ma regime. Furthermore, large Ma is anyway inaccessible to the LB algorithm, since it provides only a finite set of lattice velocities – and these essentially determine the value of c_s. In other words, the LB algorithm simply cannot realistically represent flows whose velocity is not small compared to c_s. For this reason, the details of the equation of state do not matter, and therefore one chooses the system that is by far the easiest – the ideal gas. Here the equation of state for a system at temperature T may be written as

$$k_B T = m_p c_s^2. \tag{19}$$

In the D3Q19 model (the most popular standard LB model in three dimensions, using nineteen lattice velocities, see below) it turns out that the speed of sound is given by

$$c_s^2 = \frac{1}{3}\frac{b^2}{h^2}, \tag{20}$$

where b is the lattice spacing and h the time step. Therefore the Boltzmann number can also be written as

$$Bo = \left(\frac{m_p}{\rho b^3}\right)^{1/2} = \left(\frac{3k_B T h^2}{\rho b^5}\right)^{1/2}. \tag{21}$$

4 Lattice Boltzmann 1: Statistical Mechanics

The lattice Boltzmann algorithm starts from a regular grid with sites $\vec{r}$ and lattice spacing b, plus a time step h. We then introduce a small set of velocities $\vec{c}_i$ such that $\vec{c}_i h$ connects two nearby lattice sites on the grid. In the D3Q19 model, the lattice is simple cubic, and the nineteen velocities correspond to the six nearest and twelve next–nearest neighbors, plus a zero velocity. On each lattice site $\vec{r}$ at time t, there are nineteen populations $n_i(\vec{r}, t)$.

Each population is interpreted as the mass density corresponding to velocity $\vec{c}_i$. The total mass and momentum density are therefore given by

$$\rho(\vec{r},t) = \sum_i n_i(\vec{r},t), \tag{22}$$

$$\vec{j}(\vec{r},t) = \sum_i n_i(\vec{r},t)\vec{c}_i, \tag{23}$$

such that the flow velocity is obtained via $\vec{u} = \vec{j}/\rho$. The number of "lattice Boltzmann particles" which correspond to n_i is given by

$$\nu_i = \frac{n_i b^3}{m_p} \equiv \frac{n_i}{\mu}, \tag{24}$$

where m_p is the mass of a lattice Boltzmann particle, and μ the corresponding mass density. It should be noted that μ is a measure of the thermal fluctuations in the system, since, according to Eq. 21, one has $Bo^2 = \mu/\rho$.

If we now assume a "velocity bin" i to be in thermal contact with a large reservoir of particles, the probability density for ν_i is Poissonian. Furthermore, if we assume that the "velocity bins" are statistically independent, but take into account that mass and momentum density are fixed (these variables are conserved quantities during an LB collision step and should therefore be handled like conserved quantities in a microcanonical ensemble), we find

$$P\left(\{\nu_i\}\right) \propto \left(\prod_i \frac{\bar{\nu}_i^{\nu_i}}{\nu_i!} e^{-\bar{\nu}_i}\right) \delta\left(\mu \sum_i \nu_i - \rho\right) \delta\left(\mu \sum_i \nu_i \vec{c}_i - \vec{j}\right). \tag{25}$$

for the probability density of the variables ν_i. This must be viewed as the statistics which describes the local (single–site) equilibrium under the condition of fixed values of the hydrodynamic variables ρ and $\vec{j}$. The parameter $\bar{\nu}_i$ is the mean occupation imposed by the reservoir, and we assume that it is given by

$$\bar{\nu}_i = a^{c_i} \frac{\rho}{\mu}, \tag{26}$$

where $a^{c_i} > 0$ is a weight factor corresponding to the neighbor shell with speed c_i.

From normalization and cubic symmetry we know that the low–order velocity moments of the weights must have the form

$$\sum_i a^{c_i} = 1, \tag{27}$$

$$\sum_i a^{c_i} c_{i\alpha} = 0, \tag{28}$$

$$\sum_i a^{c_i} c_{i\alpha} c_{i\beta} = \sigma_2 \, \delta_{\alpha\beta}, \tag{29}$$

$$\sum_i a^{c_i} c_{i\alpha} c_{i\beta} c_{i\gamma} = 0, \tag{30}$$

$$\sum_i a^{c_i} c_{i\alpha} c_{i\beta} c_{i\gamma} c_{i\delta} = \kappa_4 \, \delta_{\alpha\beta\gamma\delta} + \sigma_4 \left(\delta_{\alpha\beta}\delta_{\gamma\delta} + \delta_{\alpha\gamma}\delta_{\beta\delta} + \delta_{\alpha\delta}\delta_{\beta\gamma}\right), \tag{31}$$

where σ_2, σ_4, κ_4 are yet undetermined constants, while $\delta_{\alpha\beta\gamma\delta}$ is unity if all four indexes are the same and zero otherwise.

Employing Stirling's formula for the factorial, it is straightforward to find the set of populations n_i^{eq} which maximizes P under the constraints of given ρ and $\vec{j}$. Up to second order in u (low Mach number!) the solution is given by

$$n_i^{eq} = \rho a^{c_i} \left(1 + \frac{\vec{u} \cdot \vec{c}_i}{\sigma_2} + \frac{(\vec{u} \cdot \vec{c}_i)^2}{2\sigma_2^2} - \frac{u^2}{2\sigma_2} \right). \tag{32}$$

The low–order moments of the equilibrium populations are then given by

$$\sum_i n_i^{eq} = \rho, \tag{33}$$

$$\sum_i n_i^{eq} c_{i\alpha} = j_\alpha, \tag{34}$$

$$\sum_i n_i^{eq} c_{i\alpha} c_{i\beta} = \rho c_s^2 \delta_{\alpha\beta} + \rho u_\alpha u_\beta. \tag{35}$$

The first two equations are just the imposed constraints, while the last one (meaning that the second moment is just the hydrodynamic Euler stress) follows from imposing two additional conditions, which is to choose the weights a^{c_i} such that they satisfy $\kappa_4 = 0$ and $\sigma_4 = \sigma_2^2 (= c_s^4)$. From the Chapman–Enskog analysis of the LB dynamics (see below) it follows that the asymptotic behavior in the limit of large length and time scales is compatible with the Navier–Stokes equation only if Eq. 35 holds, and this in turn is only possible if the abovementioned isotropy conditions are satisfied. Together with the normalization condition, we thus obtain a set of three equations for the a^{c_i}. Therefore at least three neighbor shells are needed to satisfy these conditions, and this is the reason for choosing a nineteen–velocity model. For D3Q19, one thus obtains $a^{c_i} = 1/3$ for the zero velocity, $1/18$ for the nearest neighbors, and $1/36$ for the next–nearest neighbors. Furthermore, one finds $c_s^2 = \sigma_2 = (1/3) b^2/h^2$.

For the fluctuations around the most probable populations n_i^{eq},

$$n_i^{neq} = n_i - n_i^{eq}, \tag{36}$$

we employ a saddle–point approximation and approximate u by zero. This yields

$$P\left(\{n_i^{neq}\}\right) \propto \exp\left(- \sum_i \frac{(n_i^{neq})^2}{2\mu\rho a^{c_i}} \right) \delta\left(\sum_i n_i^{neq} \right) \delta\left(\sum_i \vec{c}_i \, n_i^{neq} \right). \tag{37}$$

We now introduce normalized fluctuations via

$$\hat{n}_i^{neq} = \frac{n_i^{neq}}{\sqrt{\mu\rho a^{c_i}}} \tag{38}$$

and transform to normalized "modes" (symmetry–adapted linear combinations of the n_i, see Ref. 2) $\hat{m}_k^{neq}$ via an orthonormal transformation $\hat{e}_{ki}$:

$$\hat{m}_k^{neq} = \sum_i \hat{e}_{ki} \hat{n}_i^{neq}, \tag{39}$$

$k = 0, \ldots, 18$, and obtain

$$P\left(\{m_k\}\right) \propto \exp\left(-\frac{1}{2}\sum_{k\geq 4} m_k^2\right). \tag{40}$$

It should be noted that the modes number zero to three have been excluded; they are just the conserved mass and momentum densities.

5 Lattice Boltzmann 2: Stochastic Collisions

A collision step consists of re-arranging the set of n_i on a given lattice site such that both mass and momentum are conserved. Since the algorithm should simulate thermal fluctuations, this should be done in a way that is (i) stochastic and (ii) consistent with the developed statistical–mechanical model. This is straightforwardly imposed by requiring that the collision is nothing but a Monte Carlo procedure, where a Monte Carlo step transforms the pre–collisional set of populations, n_i, to the post–collisional one, $n_i^\star$. Consistency with statistical mechanics can be achieved by requiring that the Monte Carlo update satisfies the condition of detailed balance. Most easily this is done in terms of the normalized modes $\hat{m}_k$, which we update according to the rule ($k \geq 4$)

$$\hat{m}_k^\star = \gamma_k \hat{m}_k + \sqrt{1-\gamma_k^2}\, r_k. \tag{41}$$

Here the γ_k are relaxation parameters with $-1 < \gamma_k < 1$, and the r_k are statistically independent Gaussian random numbers with zero mean and unit variance. Mass and momentum are automatically conserved since the corresponding modes are not updated. Comparison with Eq. 40 shows that the procedure indeed does satisfy detailed balance. The parameters γ_k can in principle be chosen at will; however, they should be compatible with symmetry. For example, mode number four corresponds to the bulk stress, with a relaxation parameter γ_b, while modes number five to nine correspond to the five shear stresses, which form a symmetry multiplett. Therefore one must choose $\gamma_5 = \ldots = \gamma_9 = \gamma_s$. For the remaining kinetic modes one often uses $\gamma_k = 0$ for simplicity, but this is not necessary.

6 Lattice Boltzmann 3: Chapman–Enskog Expansion

The actual LB algorithm now consists of alternating collision and streaming steps, as summarized in the LB equation (LBE):

$$n_i(\vec{r} + \vec{c}_i h, t + h) = n_i^\star(\vec{r}, t) = n_i(\vec{r}, t) + \Delta_i\left\{n_i(\vec{r}, t)\right\}. \tag{42}$$

The populations are first re–arranged on the lattice site; this is described by the so–called "collision operator" Δ_i. The resulting post–collisional populations $n_i^\star$ are then propagated to the neighboring sites, as expressed by the left hand side of the equation. After that, the next collision step is done, etc.. The collision step may include momentum transfer as a result of external forces (for details, see Ref. 2); apart from that, it is just given by the update procedure outlined in the previous section.

A convenient way to find the dynamic behavior of the algorithm on large length and time scales is a multi–time–scale analysis. One introduces a "coarse–grained ruler" by transforming from the original coordinates $\vec{r}$ to new coordinates $\vec{r}_1$ via

$$\vec{r}_1 = \epsilon \vec{r}, \tag{43}$$

where ϵ is a dimensionless parameter with $0 < \epsilon \ll 1$. The rationale behind this is the fact that any "reasonable" value for the scale r_1 will automatically force r to be large. In other words: By considering the limit $\epsilon \to 0$ we automatically focus our attention on large length scales. The same is done for the time; however, here we introduce *two* scales via

$$t_1 = \epsilon t \tag{44}$$

and

$$t_2 = \epsilon^2 t. \tag{45}$$

The reason for this is that one needs to consider both wave–like phenomena, which happen on the t_1 time scale (i. e. the real time is moderately large), and diffusive processes (where the real time is *very* large). We now write the LB variables as a function of $\vec{r}_1, t_1, t_2$ instead of $\vec{r}, t$. Since changing ϵ at fixed $\vec{r}_1$ changes $\vec{r}$ and thus n_i, we must take into account that the LB variables depend on ϵ:

$$n_i = n_i^{(0)} + \epsilon n_i^{(1)} + \epsilon^2 n_i^{(2)} + O(\epsilon^3). \tag{46}$$

The same is true for the collision operator:

$$\Delta_i = \Delta_i^{(0)} + \epsilon \Delta_i^{(1)} + \epsilon^2 \Delta_i^{(2)} + O(\epsilon^3). \tag{47}$$

In terms of the new variables, the LBE is written as

$$n_i(\vec{r}_1 + \epsilon \vec{c}_i h, t_1 + \epsilon h, t_2 + \epsilon^2 h) - n_i(\vec{r}_1, t_1, t_2) = \Delta_i. \tag{48}$$

Now, one systematically Taylor–expands the equation up to order ϵ^2. Sorting by order yields a hierarchy of LBEs of which one takes the zeroth, first, and second velocity moment. Systematic analysis of this set of moment equations (for details, see Ref. 2) shows that the LB procedure, as it has been developed in the previous sections, indeed yields the fluctuating Navier–Stokes equations in the asymptotic $\epsilon \to 0$ limit – however only for low Mach numbers; in the high Mach number regime, where terms of order u^3/c_s^3 can no longer be neglected, the dynamics definitely deviates from Navier–Stokes.

In particular, this analysis shows that the zeroth–order populations must be identified with n_i^{eq}, and that it is *necessary* that this "encodes" the Euler stress via suitably chosen weights a^{c_i}. Furthermore, one finds explicit expressions for the viscosities:

$$\eta = \frac{h\rho c_s^2}{2} \frac{1+\gamma_s}{1-\gamma_s}, \tag{49}$$

$$\eta_b = \frac{h\rho c_s^2}{3} \frac{1+\gamma_b}{1-\gamma_b}. \tag{50}$$

7 A Polymer Chain in Solvent

In Ref. 6 we explicitly aimed at a comparison between BD and coupled LB–MD for the *same* system. We chose a well–studied standard benchmark system, a single bead–spring polymer chain of N monomers in good solvent in thermal equilibrium. The BD algorithm is realized via

$$r_{i\alpha}(t+h) = r_{i\alpha}(t) + (k_B T)^{-1} D_{ij\alpha\beta} F_{j\beta} h + \sqrt{2h} B_{ij\alpha\beta} W_{j\beta}, \qquad i = 1, 2, \ldots, N. \quad (51)$$

Here $\vec{r}_i$ is the coordinate of the ith particle, h is the BD time step, $\overleftrightarrow{D}_{ij}$ is the diffusion tensor coupling particles i and j, and $\vec{F}_j$ denotes the deterministic force on particle j (here spring force and excluded–volume force). We assume summation convention with respect to both Cartesian and particle indexes. The tensor $\overleftrightarrow{B}_{ij}$ is the matrix square root of $\overleftrightarrow{D}_{ij}$,

$$D_{ij\alpha\beta} = B_{ik\alpha\gamma} B_{jk\beta\gamma}, \quad (52)$$

while $\vec{W}_i$ is a discretized Wiener process, $\langle W_{i\alpha} \rangle = 0$ and $\langle W_{i\alpha} W_{j\beta} \rangle = \delta_{ij} \delta_{\alpha\beta}$. For the diffusion tensor we used the Rotne–Prager tensor. The computationally most demanding part is the calculation of the matrix square root. The exact numerical solution of this problem via Cholesky decomposition has a computational complexity $O(N^3)$. We therefore rather used Fixman's trick[7,8] to speed up the calculations. This is based on the observation that the "square root" function, if viewed as a function acting on real numbers, needs to be evaluated only within a finite interval, spanning from the smallest to the largest eigenvalue. Since this interval does not contain the singularity at zero, a truncated (Chebyshev) polynomial expansion approximates the function quite well. The same expansion can then also be used to evaluate the matrix square root. The number of terms needed is empirically found to scale as $O(N^{0.25})$. Furthermore, for each term one needs to do a "matrix times vector" operation, which scales as $O(N^2)$, such that the algorithm in total has a computational complexity $O(N^{2.25})$. We did not employ an FFT–based "superfast" BD algorithm[4]; this would have been quite complicated, and also required to assume a simulation box of size L^3 with periodic bondary conditions, such that an extrapolation $L \to \infty$ would have been necessary.

Such a finite box size, combined with an extrapolation, is however precisely what is needed for LB–MD. We therefore ran these simulations for at least three different values of L in order to allow for meaningful extrapolations (and used the total time for these three systems to estimate our CPU effort). The typical box sizes that are needed are given by the requirement that the polymer chain should fit nicely into the box, without much back–folding. Since in a good solvent the polymer radius R scales as $R \propto N^\nu$, where $\nu \approx 0.59$ is the Flory exponent, we find $L^3 \propto N^{3\nu}$. Furthermore, the computational cost is completely dominated by the operations required to run the solvent, and hence the computational complexity is $O(N^{3\nu}) = O(N^{1.8})$. We see that this is slightly better than BD; however, the prefactor is much smaller for BD. In practice, we find that BD is roughly two orders of magnitude faster than LB–MD, for the typical chain lengths used in simulations, see Fig. 1.

The situation is expected to be quite different when one studies a semidilute solution, where the monomer concentration is still quite low (such that the LB–MD CPU effort is still dominated by the solvent), but the chains are so long that they strongly overlap.

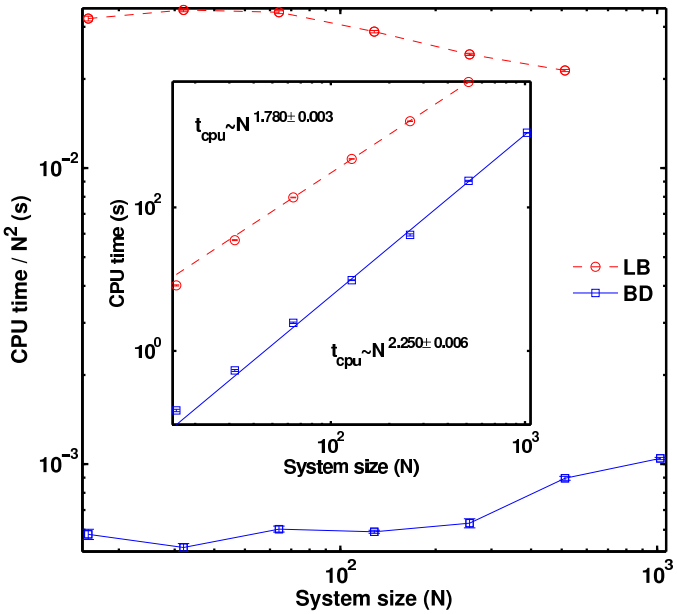


Figure 1. Comparison of the CPU time needed by the LB–MD and BD systems for the equivalent of 1000 LB–MD time steps for various chain lengths N. From Ref. 6.

For example, Ref. 9 studied 50 chains of length $N = 1000$, being well in the semidilute regime. While the additional chains for the LB–MD system pose essentially no computational burden at all (rather on the contrary: Flory screening makes the chains shrink, such that one can afford to run the simulation in a somewhat smaller box), the BD effort (for our algorithm) is expected to increase by a factor of $50^{2.25}$, i. e. more than three orders of magnitude – or even more, since one needs a more complicated scheme to evaluate the hydrodynamic interactions for a periodic system. In other words: For such a system, BD can at best be competitive if the "superfast"[4] version is implemented – and to our knowledge, this has not yet been tested.

In order to allow a meaningful comparison, both systems have to be run for the same system and the same parameters. This implies, firstly, identical interaction potentials between the beads, and the same temperature. From this one concludes (and numerically verifies) that the static properties like gyration radius, static structure factor, etc., must all be identical. For the dynamics, it is important that both simulations are run with the same value for the shear viscosity η, which is easy to achieve, plus with the same value for the monomeric friction coefficient. At this point, one has to take into account that the friction coefficient ζ that appears in the BD algorithm (on the diagonal of the diffusion tensor) is a *long–time* friction coefficient, which describes the asymptotic stationary velocity $\vec{v}$ of a particle that is dragged through the fluid with a force $\vec{F}$, $\vec{F} = \zeta\vec{v}$, while the friction coefficient Γ that appears in the LB–MD algorithm via the coupling prescription $\vec{F} = \Gamma(\vec{v} - \vec{u})$ (see Eq. 10) is a corresponding *short–time* coefficient that does not yet take the backflow effects into account. Indeed, for an experiment in which a particle is dragged through the

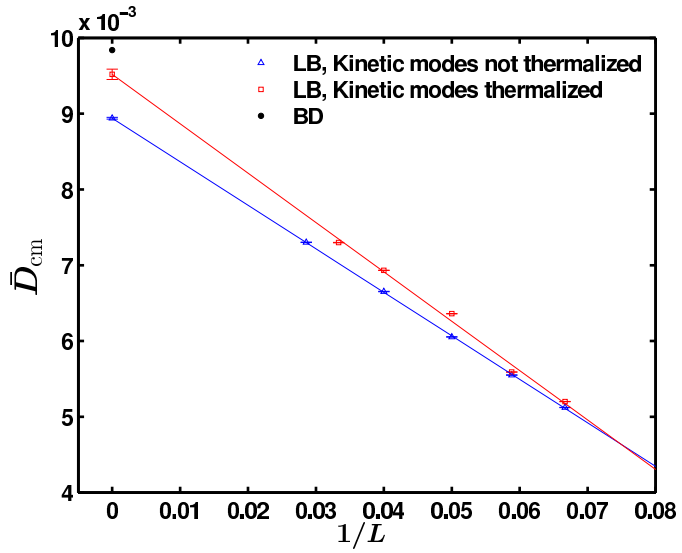


Figure 2. The dimensionless long time diffusion constant for the center of mass at various box lengths L. From Ref. 6

LB fluid, it is clear that the flow velocity $\vec{u}$ will be nonzero, and typically slightly smaller than $\vec{v}$. Hence, $\vec{F} = \zeta\vec{v} = \Gamma(\vec{v} - \vec{u})$, i. e. ζ is smaller than Γ. Since hydrodynamics allows us to estimate $\vec{u}$ up to a numerical prefactor g via a Stokes–like formula, $\vec{F} = g\eta a\vec{u}$, where a is the range of the interpolation scheme, one finds (see also Ref. 2)

$$\zeta^{-1} = \Gamma^{-1} + (g\eta a)^{-1}. \tag{53}$$

For nearest–neighbor linear interpolation, one finds $g \approx 25$ if a is identified with the LB lattice spacing. One hence needs to choose the Γ value in the LB–MD simulations in such a way that it reproduces the BD ζ value.

The diffusion constant of the LB–MD chain depends on the box size, as a result of the hydrodynamic interaction with the periodic images. Since the latter decays like r^{-1}, one concludes an L^{-1} finite size effect, which is nicely borne out by the data of Fig. 2. From these data one sees also that for an accurate description of the dynamics it is necessary to not only thermalize the stress modes in the LB algorithm (only these matter in the strict hydrodynamic limit), but also the kinetic modes, as suggested by the more microscopic theory outlined above. Taking the finite–size effect and the proper thermalization into account, the remaining deviation between BD and LB–MD is only a few percent.

The internal Rouse modes of the chain are defined as ($p = 1, 2, \ldots, N-1$)

$$\vec{X}_p = \frac{1}{N}\sum_{n=1}^{N} \vec{r}_n \cos\left[\frac{p\pi}{N}\left(n - \frac{1}{2}\right)\right]. \tag{54}$$

Fig. 3 shows the decay of the normalized mode autocorrelation function up to $p = 5$. Obviously the agreement with BD is quite good, i. e. the finite size effect is quite weak. The reason is the following: The diffusion constant corresponds to the friction of the chain

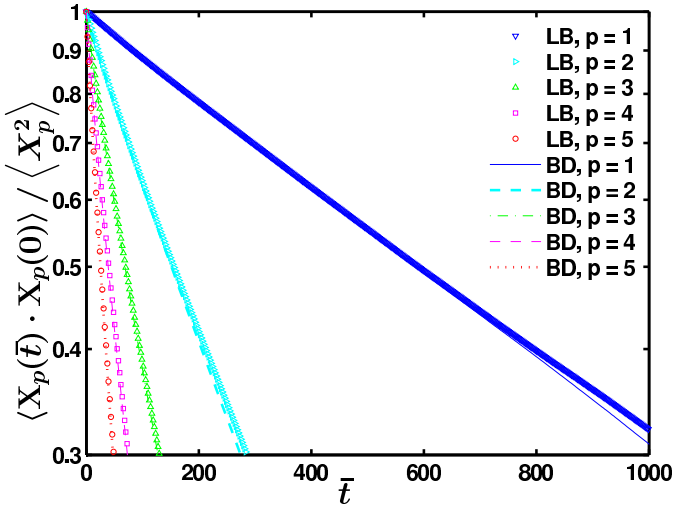


Figure 3. Normalized autocorrelation function of the first 5 Rouse modes $\vec{X}_p$ for LB–MD simulations at fixed $L = 25$ and BD simulations at $L \to \infty$. From Ref. 6.

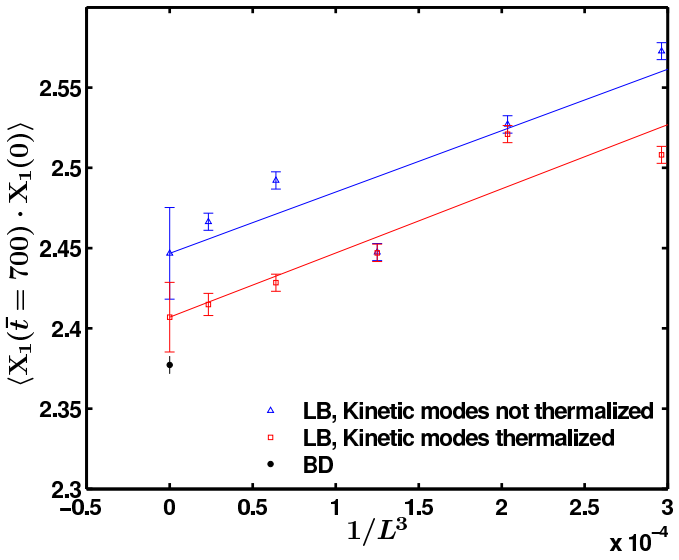


Figure 4. The autocorrelation function for the first Rouse mode $\vec{X}_1$ at a finite time value of $\bar{t} = 700$ for LB–MD simulations at various box lengths L and BD simulations at $L \to \infty$. From Ref. 6.

as a whole, i. e. to an experiment where the chain is being dragged through the fluid with a constant force. This gives rise to a flow field that decays like r^{-1}, and thus an L^{-1} finite size effect. This total force may also be viewed as the monopole moment of a distribution of forces acting on the polymer. The Rouse modes however study the *internal* motion of

the chain, i. e. in the center–of–mass system. Therefore, the monopole contribution of the forces has been subtracted, and only higher–order multipole moments remain. The dipole contribution vanishes for symmetry reasons, i. e. the first higher–order multipole is the quadrupole (this may be vaguely understood by recalling that the mass distribution has a monopole and a quadrupole moment, but not a dipole moment). The quadrupolar flow field decays like r^{-3}, and hence one expects an L^{-3} finite size effect. For a more detailed derivation, see Ref. 10. This finite size effect is indeed observed, see Fig. 4, demonstrating that on the one hand the system is theoretically quite well understood, and that on the other hand such simulations are nowadays so accurate that even rather subtle effects can be analyzed unambiguously.

References

1. B. Dünweg, "Computer simulations of systems with hydrodynamic interactions: The coupled Molecular Dynamics - lattice Boltzmann approach", in: Multiscale Simulation Methods in Molecular Sciences, J. Grotendorst, N. Attig, S. Blügel, and D. Marx, (Eds.). Forschungszentrum Jülich, Jülich, 2009.
2. B. Dünweg and A. J. C. Ladd, *Lattice Boltzmann simulations of soft matter systems*, Advances in Polymer Science, **221**, 89, 2009.
3. G. Nägele, "Brownian dynamics simulations", in: Computational Condensed Matter Physics, S. Blügel, G. Gompper, E. Koch, H. Müller-Krumbhaar, R. Spatschek, and R. G. Winkler, (Eds.). Forschungszentrum Jülich, Jülich, 2006.
4. A. J. Banchio and J. F. Brady, Journal of Chemical Physics, **118**, 10323, 2003.
5. M. Ripoll, "Mesoscale hydrodynamics simulations", in: Computational Condensed Matter Physics, S. Blügel, G. Gompper, E. Koch, H. Müller-Krumbhaar, R. Spatschek, and R. G. Winkler, (Eds.). Forschungszentrum Jülich, Jülich, 2006.
6. Tri T. Pham, Ulf D. Schiller, J. Ravi Prakash, and B. Dünweg, Journal of Chemical Physics, **131**, 164114, 2009.
7. M. Fixman, Macromolecules, **19**, 1204, 1986.
8. R. M. Jendrejack, M. D. Graham, and J. J. De Pablo, Journal of Chemical Physics, **113**, 2894, 2000.
9. P. Ahlrichs, R. Everaers, and B. Dünweg, Physical Review E, **64**, 040501, 2001.
10. P. Ahlrichs and B. Dünweg, Journal of Chemical Physics, **111**, 8225, 1999.

Flow Simulations with Multiparticle Collision Dynamics

Roland G. Winkler

Institute for Advanced Simulation
Forschungszentrum Jülich, 52425 Jülich, Germany
E-mail: r.winkler@fz-juelich.de

1 Introduction

During the last few decades, soft matter has developed into an interdisciplinary research field combing physics, chemistry, chemical engineering, biology, and materials science. This is driven by the specificities of soft matter, which consists of large structural units in the nano- to micrometer range and is sensitive to thermal fluctuations and weak external perturbations[1–3]. Soft matter comprises traditional complex fluids such as amphiphilic mixtures, colloidal suspensions, and polymer solutions, as well as a wide range of phenomena including self-organization, transport in microfluidic devices and biological capillaries, chemically reactive flows, the fluid dynamics of self-propelled objects, and the visco-elastic behavior of networks in cells[2].

The presence of disparate time, length, and energy scales poses particular challenges for conventional simulation techniques. Biological systems present additional problems, because they are often far from equilibrium and are driven by strong spatially and temporally varying forces. The modeling of these systems often requires the use of coarse-grained or mesoscopic approaches that mimic the behavior of atomistic systems on the length scales of interest. The goal is to incorporate the essential features of the microscopic physics in models which are computationally efficient and are easily implemented in complex geometries and on parallel computers, and can be used to predict emergent properties, test physical theories, and provide feedback for the design and analysis of experiments and industrial applications[2]. In many situations, a simple continuum description, e.g., based on the Navier-Stokes equation is not sufficient, since molecular-level details play a central role in determining the dynamic behavior. A key issue is to resolve the interplay between thermal fluctuations, hydrodynamic interactions (HI), and spatiotemporally varying forces.

The desire to bridge the length- and time-scale gap has stimulated the development of mesoscale simulation methods such as Dissipative Particle Dynamics (DPD)[4–6], Lattice-Boltzmann (LB)[7–9], Direct Simulation Monte Carlo (DSMC)[10–12], and Multiparticle Collision dynamics (MPC)[13,14]. Common to these approaches is a simplified, coarse-grained description of the solvent degrees of freedom. Embedded solute particles, such as polymers or colloids, are often treated by conventional molecular dynamics simulations. All these approaches are essentially alternative ways of solving the Navier-Stokes equation for the fluid dynamics.

In this contribution, the MPC approach – also denoted as stochastic rotation dynamics (SRD) – is discussed, which has been introduced by Malevanets and Kapral[13,14], and is an extension of the DSMC method to fluids. The fluid is represented by point particles and their dynamics proceeds in two steps: A streaming and a collision step. Collisions occur at fixed discrete time intervals, and although space is discretized into cells to define

the multiparticle collision environment, both the spatial coordinates and the velocities of the particles are continuous variables. The algorithm exhibits unconditional numerical stability and has an H-theorem[13,14]. MPC defines a discrete-time dynamics which has been shown to yield the correct longtime hydrodynamics. It also fully incorporates both thermal fluctuations and hydrodynamic interactions. In addition, HI can be easily switched off in MPC algorithms, making it easy to study the importance of such interactions[15,16].

It must be emphasized that all local algorithms, such as MPC, DPD, and LB, model compressible fluids, so that it takes time for the hydrodynamic interactions to "propagate" over longer distances. As a consequence, these methods become quite inefficient in the Stokes limit, where the Reynolds number approaches zero. MPC is particularly well suited for studying phenomena where both thermal fluctuations and hydrodynamics are important, for systems with Reynolds and Peclet numbers of order $0.1 - 10$, if exact analytical expressions for the transport coefficients and consistent thermodynamics are needed, and for modeling complex phenomena for which the constitutive relations are not known. Examples include chemically reacting flows, self-propelled objects, or solutions with embedded macromolecules and aggregates. If thermal fluctuations are not essential or undesirable, a more traditional method such as a finite-element solver or a LB approach is recommended. If, on the other hand, inertia and fully resolved hydrodynamics are not crucial, but fluctuations are, one might be better served using Langevin or BD simulations.

2 Multiparticle Collision Dynamics

In MPC, the solvent is represented by N point-like particles of mass m. The algorithm consists of individual streaming and collision steps. In the streaming step, the particles move independent of each other and experience only possibly present external forces (cf. Sec. 7). Without such forces, they move ballistically and their positions $\boldsymbol{r}_i$ are updated according to

$$\boldsymbol{r}_i(t+h) = \boldsymbol{r}_i(t) + h\boldsymbol{v}_i(t), \tag{1}$$

where $i = 1, \ldots, N$, $\boldsymbol{v}_i$ is the velocity of particle i, and h is the time interval between collisions, which will be denoted as collision time. In the collision step, a coarse-grained interaction between the fluid particles is imposed by a stochastic process. For this purpose, the system is divided in cubic collision cells of side length a. An elementary requirement is that the stochastic process conserves momentum on the collision-cell level, only then HI are present in the system. There are various possibilities for such a process. Originally, the rotation of the relative velocities, with respect to the center-of-mass velocity of the cell, around a randomly orientated axis by a fixed angle α has been suggested[13,14], i.e,

$$\boldsymbol{v}_i(t+h) = \boldsymbol{v}_i(t) + (\mathcal{D}(\alpha) - \mathcal{E})\left(\boldsymbol{v}_i(t) - \boldsymbol{v}_{cm}(t)\right), \tag{2}$$

where $\mathcal{D}(\alpha)$ is the rotation matrix, $\mathcal{E}$ is the unit matrix, and

$$\boldsymbol{v}_{cm} = \frac{1}{N_c}\sum_{i=1}^{N_c}\boldsymbol{v}_i \tag{3}$$

is the center-of-mass velocity of the N_c particles contained in the cell of particle i. The orientation of the rotation axis is chosen randomly for every collision cell and time step. As

is easily shown, the algorithm conserves mass, momentum, and energy in every collision cell, which leads to long-range correlations between particles.

The rotations can be realized in different ways. On the one hand, the rotation matrix

$$\mathcal{D}(\alpha) = \begin{pmatrix} \mathcal{R}_x^2 + (1-\mathcal{R}_x^2)c & \mathcal{R}_x\mathcal{R}_y(1-c) - \mathcal{R}_z s & \mathcal{R}_x\mathcal{R}_z(1-c) + \mathcal{R}_y s \\ \mathcal{R}_x\mathcal{R}_y(1-c) + \mathcal{R}_z s & \mathcal{R}_y^2 + (1-\mathcal{R}_y^2)c & \mathcal{R}_y\mathcal{R}_z(1-c) - \mathcal{R}_x s \\ \mathcal{R}_x\mathcal{R}_z(1-c) - \mathcal{R}_y s & \mathcal{R}_y\mathcal{R}_z(1-c) + \mathcal{R}_x s & \mathcal{R}_z^2 + (1-\mathcal{R}_z^2)c \end{pmatrix} \quad (4)$$

can be used, with the unit vector $\boldsymbol{\mathcal{R}} = (\mathcal{R}_x, \mathcal{R}_y, \mathcal{R}_z)^T$, $c = \cos\alpha$, and $s = \sin\alpha$. The Cartesian components of $\boldsymbol{\mathcal{R}}$ are defined as

$$\mathcal{R}_x = \sqrt{1-\theta^2}\cos\varphi\ ,\ \mathcal{R}_y\sqrt{1-\theta^2}\sin\varphi\ ,\ \mathcal{R}_z = \theta, \quad (5)$$

where φ and θ are uncorrelated random numbers, which are taken from uniform distributions in the intervals $[0, 2\pi]$ and $[-1, 1]$, respectively. On the other hand, a vector rotation can be performed[17]. The vector $\Delta\boldsymbol{v}_i = \boldsymbol{v}_i - \boldsymbol{v}_{cm} = \Delta\boldsymbol{v}_{i,\parallel} + \Delta\boldsymbol{v}_{i,\perp}$ is given by the component $\Delta\boldsymbol{v}_{i,\parallel} = (\Delta\boldsymbol{v}_i\boldsymbol{\mathcal{R}})\boldsymbol{\mathcal{R}}$ parallel to $\boldsymbol{\mathcal{R}}$ and $\Delta\boldsymbol{v}_{i,\perp} = \Delta\boldsymbol{v}_i - \Delta\boldsymbol{v}_{i,\parallel}$ perpendicular to $\boldsymbol{\mathcal{R}}$. Rotation by an angle α transforms $\Delta\boldsymbol{v}_i$ into $\Delta\boldsymbol{v}_i' = \Delta\boldsymbol{v}_{i,\parallel} + \Delta\boldsymbol{v}_{i,\perp}'$. $\Delta\boldsymbol{v}_{i,\perp}'$ can be expressed by the vector $\Delta\boldsymbol{v}_{i,\perp}$ and the vector $\boldsymbol{\mathcal{R}} \times \Delta\boldsymbol{v}_{i,\perp}$, which yields

$$\begin{aligned} \boldsymbol{v}_i(t+h) &= \boldsymbol{v}_{cm}(t) + \cos\alpha\Delta\boldsymbol{v}_{i,\perp} + \sin\alpha\,(\boldsymbol{\mathcal{R}} \times \Delta\boldsymbol{v}_{i,\perp}) + \Delta\boldsymbol{v}_{i,\parallel} \\ &= \boldsymbol{v}_{cm}(t) + \cos\alpha\,[\Delta\boldsymbol{v}_i - (\Delta\boldsymbol{v}_i\boldsymbol{\mathcal{R}})\boldsymbol{\mathcal{R}}] \\ &\quad + \sin\alpha\,\boldsymbol{\mathcal{R}} \times [\Delta\boldsymbol{v}_i - (\Delta\boldsymbol{v}_i\boldsymbol{\mathcal{R}})\boldsymbol{\mathcal{R}}] + (\Delta\boldsymbol{v}_i\boldsymbol{\mathcal{R}})\boldsymbol{\mathcal{R}}\ , \end{aligned} \quad (6)$$

since the vector $\boldsymbol{\mathcal{R}} \times \Delta\boldsymbol{v}_{i,\perp}$ is perpendicular to $\boldsymbol{\mathcal{R}}$ and $\Delta\boldsymbol{v}_{i,\perp}$.

In its original form[2,13,14,18], the MPC algorithm violates Galilean invariance. This is most pronounced at low temperatures or small time steps, where the mean free path $\lambda = h\sqrt{k_BT/m}$ of a particle is smaller than the cell size a. Then, the same particles repeatedly interact with each other in the same cell and thereby build up correlations. In a collision lattice moving with a constant velocity, other particles interact with each other, creating less correlations, which implies breakdown of Galilean invariance. In Refs. 19,20, a random shift of the entire computational grid is introduced to restore Galilean invariance. In practice, all particles are shifted by the same random vector with components uniformly distributed in the interval $[-a/2, a/2]$ before the collision step. After the collision, particles are shifted back to their original positions. As a consequence, no reference frame is preferred.

The velocity distribution is given by the Maxwell-Boltzmann distribution in the limit $N \to \infty$, and the probability to find N_c particles in a cell is given by the Poisson distribution

$$P(N_c) = e^{-\langle N_c\rangle}\langle N_c\rangle^{N_c}/N_c!\ , \quad (7)$$

where $\langle N_c\rangle$ is the average number of the particles in a cell.

As an alternative collision rule, Maxwell-Boltzmann, i.e., Gaussian distributed relative velocities $\boldsymbol{v}_i^{\mathrm{ran}}$ of variance $\sqrt{k_BT/m}$ can be used to create new velocities according to[17,21,22]

$$\boldsymbol{v}_i(t+h) = \boldsymbol{v}_{cm}(t) + \boldsymbol{v}_i^{\mathrm{ran}} - \frac{1}{N_c}\sum_{j=1}^{N_c}\boldsymbol{v}_i^{\mathrm{ran}}\ . \quad (8)$$

Here, a canonical ensemble is simulated and no further thermalization is needed in non-equilibrium simulations, where there is viscose heating. From a numerical point of view, however, the calculation of the Gaussian random numbers is somewhat more time consuming, hence the performance is slower compared to SRD[2]. In Refs. 21–23 algorithms are presented, which additionally preserve angular momentum.

3 Embedded Objects and Boundary Conditions

A very simple procedure for coupling embedded objects such as colloids or polymers to a MPC solvent has been proposed in Refs. 24–26. In this approach, every colloid particle or monomer in a polymer is taken to be a point-particle which participates in the MPC collision. If monomer k has mass M and velocity $\boldsymbol{V}_k$ the center-of-mass velocity of all particles (MPC and monomers) in a collision cell is

$$\boldsymbol{v}_{cm} = \frac{\sum_{i=1}^{N_c} m\boldsymbol{v}_i + \sum_{k=1}^{N_m^c} M\boldsymbol{V}_k}{mN_c + MN_m^c}, \tag{9}$$

where N_m^c is the number of monomers in the collision cell. A stochastic collision of the relative velocities of both the solvent particles and embedded monomers is then performed in the collision step, which leads to an exchange of momentum between them. The dynamics of the monomers is typically treated by molecular dynamics simulations (MD), applying the velocity Verlet integration scheme[27,28]. Hence, the new monomer momenta are used as initial conditions for the the subsequent streaming step (MD) of duration h. In this approach, the average mass of solvent particles per cell $m\langle N_c\rangle$, should be of the order of the monomer or colloid mass M (assuming one embedded particle per cell). This corresponds to a neutrally buoyant object which responds quickly to the fluid flow but is not kicked around too violently. It is also important to note that the average number of monomers per cell $\langle N_m\rangle$ should be smaller than unity in order to properly resolve HI between them. On the other hand, the average bond length in a semiflexible polymer or rodlike colloid should also not be much larger than the cell size a, in order to capture the anisotropic friction of rodlike molecules due to HI (which leads to a twice as large perpendicular than parallel friction coefficient for long stiff rods[29,30]), and to avoid an unnecessarily large ratio of the number of solvent to solute particles. Hence, the average bond length should be of order a.

To accurately resolve the local flow field around a colloid, methods have been proposed which exclude fluid-particles from the interior of the colloid and mimic slip[14,31] or no-slip[2,32–34] boundary conditions. No-slip boundary conditions are modeled by the bounce-back rule. Here, the velocity of a particle is inverted from $\boldsymbol{v}_i$ to $-\boldsymbol{v}_i$ when it intersects the surface of an impenetrable particle, e.g., colloid or blood cell, or wall. Since walls or surfaces will generally not coincide with the collision cell boundaries, in particular due to random shifts, the simple bounce-back rule fails to guarantee no-slip boundary conditions. The following generalization of the bounce-back rule has therefore been suggested[32]: For all cells that are intersected by walls, fill the wall part of the cell with a sufficient number of virtual particles in order to make the total number of particles equal to $\langle N_c\rangle$. The velocities of the wall particles are taken from a Maxwell-Boltzmann distribution with zero mean and variance k_BT/m. Since the sum of Gaussian random numbers

is also Gaussian distributed, the velocities of the individual virtual particles need not be determined explicitly, it suffices to determine a momentum $\boldsymbol{p}$ from a Maxwell-Boltzmann distribution with zero mean and variance mN_pk_BT, where $N_p = \langle N_c \rangle - N_c$ is the number of virtual particles corresponding to the partially filled cell of N_c particles. The center-of-mass velocity of the cell is then

$$\boldsymbol{v}_{cm} = \frac{1}{m\langle N_c \rangle} \left(\sum_{i=1}^{N_c} m\boldsymbol{v}_i + \boldsymbol{p} \right). \tag{10}$$

Results for a Poiseuille flow obtained by this procedure, both with and without cell shifting, are in good agreement with the correct parabolic flow profile[32] (see Sec. 7.2).

However, this does not completely prevent slip, because the average center-of-mass position of all particles in a collision cell – including the virtual particle – does not coincide with the wall. In order to further reduce slip, the following modification of the original approach has been proposed[35]. To treat a surface cell on the same basis as a cell in the bulk, i.e., the number of particles satisfies the Poisson distribution with the average $\langle N_c \rangle$, we take fluctuations in the particle number into account by adding N_p virtual particles to every cell intersected by a wall such that $\langle N_p + N_c \rangle = \langle N_c \rangle$. There are various ways to determine the number N_p. For a system with parallel walls, we suggest to use the number of fluid particles in the opposite surface cell, i.e., the opposing surface cell cut by the other wall. The average of the two numbers is equal to $\langle N_c \rangle$. Alternatively, N_p can be taken from a Poisson distribution with average $\langle N_c \rangle$ accounting for the fact that there are already N_c particles in the cell. Now, the center-of-mass velocity of the particles in a boundary cell is

$$\boldsymbol{v}_{cm} = \frac{1}{m(N_c + N_p)} \left(\sum_{i=1}^{N_c} m\boldsymbol{v}_i + \boldsymbol{p} \right). \tag{11}$$

The momentum $\boldsymbol{p}$ of the effective virtual particle is obtained as described above.

4 Cell-Level Canonical Thermostat

In any nonequilibrium situation, the presence of external fields destroys energy conservation and a control mechanism has to be implemented to maintain temperature (a brief review on existing thermostats is presented in Ref. 36). A basic requirement of any thermostat is that it does not violate local momentum conservation, smear out local flow profiles, or distort the velocity distribution too much. A simple and efficient way to maintain a constant temperature is velocity scaling. For a homogeneous system, a single global scaling factor is sufficient. For an inhomogeneous system, such as shear flow or Poiseuille flow, a local, profile-unbiased thermostat is required. Here, the relative velocities $\Delta\boldsymbol{v}_i = \boldsymbol{v}_i - \boldsymbol{v}_{cm}$ (2) are scaled, before or after the rotation (velocity scaling exchanges with the rotation), i.e., $\Delta\boldsymbol{v}_i' = \kappa\Delta\boldsymbol{v}_i$, where κ is the scale factor.

In its simplest form, velocity scaling keeps the kinetic energy at the desired value. For a profile-unbiased *global scaling* scheme, the scale factor is give by

$$\kappa = \left(\frac{3(N - N_{cl})k_BT}{2E_k} \right)^{1/2} \tag{12}$$

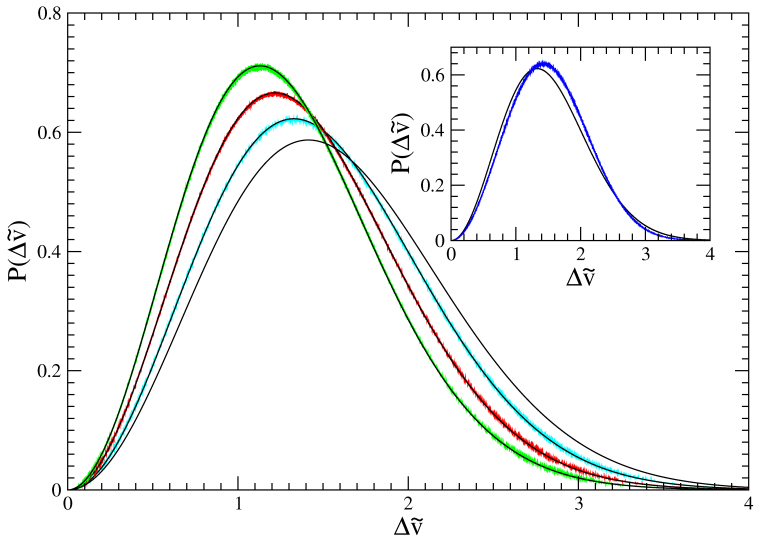


Figure 1. Distribution functions of the particle velocities $|\Delta \boldsymbol{v}|$ in a collision cell under shear flow for the average particle numbers $\langle N_c \rangle = 3$ (green), 5 (red), 10 (blue), and ∞ (black). The solid lines are determined using Eq. 17. $\Delta\tilde{v}$ is an abbreviation for $\Delta\tilde{v} = \Delta v/\sqrt{k_B T/m}$. The inset shows the distribution function for velocity scaling with the thermal energy $E_k = 3(N_c - 1)k_B T/2$ for $\langle N_c \rangle = 10$ in comparison to the correct Maxwell-Boltzmann result (black)[36].

in three-dimensional space, where N_{cl} is the number of collision cells and $E_k = \sum_{i=1}^{N} m\Delta \boldsymbol{v}_i^2/2$ the kinetic energy of all particles with respect to their cells' center-of-mass velocities. The corresponding expression for *cell-level* scaling is

$$\kappa = \left(\frac{3(N_c - 1)k_B T}{2E_k} \right)^{1/2}, \tag{13}$$

where now $E_k = \sum_{i=1}^{N_c} m\Delta \boldsymbol{v}_i^2/2$ is the kinetic energy of the particles within the particular cell. Note that the scale factor is different for every cell.

This kind of temperature control corresponds to an isokinetic rather than isothermal, i.e., canonical ensemble. As shown in Sec. 7.2, this may have sever consequences on certain properties such as local temperature or particle number[36]. Such artifacts are avoided by a cell-level canonical thermostat. Instead of using the thermal energy as reference, an kinetic energy is determined from its distribution function in a canonical ensemble[36]

$$P(E_k) = \frac{1}{E_k \Gamma(f/2)} \left(\frac{E_k}{k_B T} \right)^{f/2} \exp\left(-\frac{E_k}{k_B T} \right). \tag{14}$$

Here, $f = 3(N_c - 1)$ denotes the degrees of freedom of the considered system and $\Gamma(x)$ is the gamma function. The distribution function $P(E_k)$ itself is denoted as gamma distribution. In the limit $f \to \infty$, the gamma distribution turns into a Gaussian function with the mean $\langle E_k \rangle = f k_B T/2$ and variance $f(k_B T)^2/2$.

To thermalize the velocities of the MPC fluid on the cell level, a different energy E_k is taken from the distribution function (14) for every cell and time step and the velocities are

scaled by the factor

$$\kappa = \left(\frac{2E_k}{\sum_{i=1}^{N_c} m\Delta \boldsymbol{v}_i^2} \right)^{1/2}. \tag{15}$$

For a fixed N_c, we then obtain the following distribution function for the relative velocity of a particle in a cell in the limit of a large number of MPC steps

$$P(\Delta \boldsymbol{v}, N_c) = \left(\frac{m}{2\pi k_B T(1-1/N_c)} \right)^{3/2} \exp\left(-\frac{m}{2k_B T(1-1/N_c)} \Delta \boldsymbol{v}^2 \right). \tag{16}$$

However, the number of fluid particles in a cell is fluctuating in time. Thus, the actual distribution function is obtained by averaging Eq. 16 over the Poisson distribution (7)

$$P(\Delta \boldsymbol{v}) = \sum_{N_c=2}^{\infty} e^{-\langle N_c \rangle} \frac{\langle N_c \rangle^{N_c}}{N_c!} P(\Delta \boldsymbol{v}, N_c) / \left(1 - (\langle N_c \rangle + 1) e^{-\langle N_c \rangle} \right). \tag{17}$$

Fig. 1 provides an example of velocity distributions of a MPC fluid under shear flow. Evidently excellent agreement is obtained between the simulation result and the theoretical expression.

5 Transport Coefficients

A major advantage of the MPC dynamics is that the transport properties that characterize the macroscopic laws may be computed and analytical expressions be derived[18]. In the following, the self-diffusion coefficient and the viscosity of the MPC solvent will be discussed. Other aspects are presented in Refs. 2, 18, 37.

5.1 Diffusion Coefficient

The diffusion coefficient D of a particle i can be obtained from the Green-Kubo relation[2,18,20,38]

$$D = \frac{h}{6} \left\langle \boldsymbol{v}_i(0)^2 \right\rangle + \frac{h}{3} \sum_{n=1}^{\infty} \left\langle \boldsymbol{v}_i(nh) \boldsymbol{v}_i(0) \right\rangle \tag{18}$$

for a discrete-time random system in three-dimensional space. $t_n = nh$ denotes the discrete time of the nth collision. The average $\langle \ldots \rangle$ comprises both, averaging over the orientation of the rotation axis ($\boldsymbol{\mathcal{R}}$) and the distribution of velocities. The two are independent. To evaluate the expression, the velocity auto-correlation function is required. An exact evaluation of the correlation function is difficult or even impossible, because it would imply that the full correlated dynamics of the particles can analytically be calculated. However, an approximate expression can be derived.

In a first step, the average over the random orientation of the rotation axis is performed. Since the orientation is isotropic in space, all odd moments of the Cartesian components of $\boldsymbol{\mathcal{R}}$ vanish and the second moments are given by $\langle \mathcal{R}_\beta \mathcal{R}_{\beta'} \rangle = \delta_{\beta\beta'}/3$. Thus,

$$\langle \boldsymbol{\mathcal{R}} \Delta \boldsymbol{v}_i \rangle = \frac{1}{3} \left(1 + 2\cos\alpha \right) \Delta \boldsymbol{v}_i, \tag{19}$$

which yields

$$\langle \boldsymbol{v}_i(t+h)\boldsymbol{v}_i(t)\rangle = \langle \boldsymbol{v}_{cm}(t)\boldsymbol{v}_i(t)\rangle + \frac{1}{3}\left(1+2\cos\alpha\right)\langle \Delta\boldsymbol{v}_i(t)\boldsymbol{v}_i(t)\rangle\,. \tag{20}$$

To evaluate the correlation function with the center-of-mass velocity, we apply the molecular chaos assumption, which assumes that different particles are independent, i.e., $\langle \boldsymbol{v}_j(t)\boldsymbol{v}_i(t')\rangle = \delta_{ij}\langle \boldsymbol{v}_i(t)\boldsymbol{v}_i(t')\rangle$ and $\langle \boldsymbol{v}_{cm}(t)\boldsymbol{v}_i(t)\rangle = \sum_{j=1}^{N_c}\langle \boldsymbol{v}_j(t)\boldsymbol{v}_i(t)\rangle/N_c = \left\langle \boldsymbol{v}_i(t)^2\right\rangle/N_c$. Hence,

$$\langle \boldsymbol{v}_i(t+h)\boldsymbol{v}_i(t)\rangle = \left[1-\frac{2}{3}(1-\cos\alpha)\left(1-\frac{1}{N_c}\right)\right]\left\langle \boldsymbol{v}_i(t)^2\right\rangle. \tag{21}$$

To account for particle number fluctuations, this expression has to be averaged applying the Poisson distribution (7). Since we consider a particular particle in a cell, the probability distribution of finding N_c-1 other particles in that cell is $N_cP(N_c)/\langle N_c\rangle$. Averaging over this distribution gives[2,18,39]

$$\sum_{N_c=1}^{\infty} e^{-\langle N_c\rangle}\frac{\langle N_c\rangle^{(N_c-1)}}{(N_c-1)!}\left(1-\frac{1}{N_c}\right) = \frac{1}{\langle N_c\rangle}\left(e^{-\langle N_c\rangle}+\langle N_c\rangle-1\right). \tag{22}$$

Thus,

$$\langle \boldsymbol{v}_i(t+h)\boldsymbol{v}_i(t)\rangle = (1-\gamma)\left\langle \boldsymbol{v}_i(t)^2\right\rangle,\ \text{with}\ \gamma = \frac{2(1-\cos\alpha)}{3\langle N_c\rangle}\left(e^{-\langle N_c\rangle}+\langle N_c\rangle-1\right) \tag{23}$$

This expression reduces to Eq. 21 for $\langle N_c\rangle \gg 1$. In fact, we can replace N_c by $\langle N_c\rangle$ already for $N_c \gtrsim 5$.

More generally, iteration yields

$$\langle \boldsymbol{v}_i(nh)\boldsymbol{v}_i(0)\rangle = (1-\gamma)^n\left\langle \boldsymbol{v}_i(0)^2\right\rangle. \tag{24}$$

This relation suggests that the velocity correlation function decays exponentially, which is not the case and is a result of the applied approximation, which neglects all correlations. In contrast, HI lead to a long-time tail of the velocity correlation function

$$\langle \boldsymbol{v}_i(t)\boldsymbol{v}_i(0)\rangle = \frac{k_BT}{4m\langle N_c\rangle\pi^{3/2}}\frac{1}{\left([\nu+D]t\right)^{3/2}} \tag{25}$$

with an algebraic decay[40–42] in the limit $t\to\infty$. This relation can be derived from the Navier-Stokes equation[30,40,42–44].

With Eq. 24, the diffusion coefficient follows as[37,38]

$$D = \frac{h\left\langle \boldsymbol{v}_i(0)^2\right\rangle}{3}\left(\frac{1}{\gamma}-\frac{1}{2}\right) = \frac{hk_BT}{m}\left(\frac{1}{\gamma}-\frac{1}{2}\right) \tag{26}$$

within the molecular chaos assumption.

Fig. 2 shows simulation results for various collision time steps[41]. As expected, the molecular chaos assumption works well for large collision time steps $h/\sqrt{ma^2/(k_BT)} = \lambda/a > 1$, and hence large mean-free paths λ, where the particles are exposed to a nearly random collision environment at every step. This is reflected in Fig. 2, where the velocity correlation function decays exponentially for large collision time steps

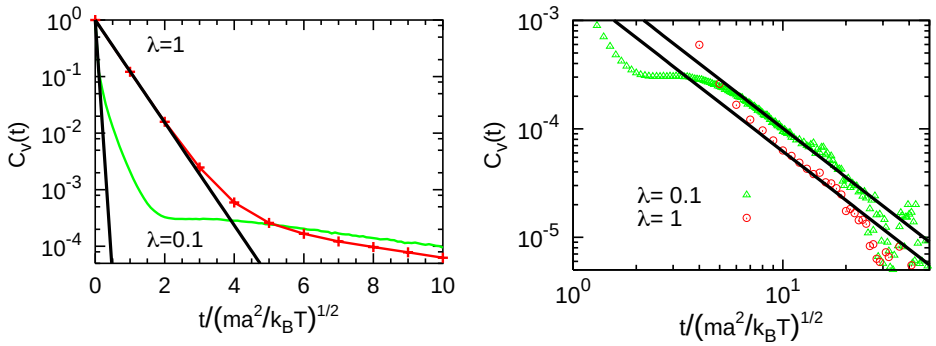


Figure 2. Fluid velocity auto-correlation functions $C_V(t) = \langle \boldsymbol{v}(t)\boldsymbol{v}(0)\rangle / \langle \boldsymbol{v}(0)^2\rangle$ for the mean free paths $\lambda/a = h/\sqrt{ma^2/(k_BT)} = 0.1$ and 1. Left: The semi-logarithmic representation shows an exponential decay (solid lines) of the correlation function at short times and large mean-free paths. Right: The (black) solid lines are calculated according to Eq. 25 and show the power-law decay $\sim t^{-3/2}$.

over a certain time window. For small collision times, the same particles collide several times with each other, which builds up correlations. Here, sound propagation plays an important role and contributes to the decay[45] at short times. For longer times, vorticity determines the time dependence of the correlation function, which then decays by the power-law (25)[45]. The simulations results are in close quantitative agreement with the theoretical prediction (25).

Calculations of the diffusion coefficient reflect the same behavior[18,41]. For $\lambda/a \gtrsim 0.5$, the numerical results for D agree very well with the analytical expression, whereas for smaller λ values, a somewhat large D is obtained[41].

5.2 Viscosity

The shear viscosity is one of the most important properties of complex fluids. In particular, it characterizes their non-equilibrium behavior, e.g., in rheology. Various ways have been suggested to obtain an analytical expression for the viscosity of a MPC fluid. In Refs. 2, 20, 38, 46, 47, linear hydrodynamic equations (Navier-Stokes equation) and Green-Kubo relations are exploited. Alternatively, non-equilibrium simulations can be performed and transport coefficients are obtained from the linear response to an imposed gradient. The two approaches are related by the fluctuation-dissipation theorem.

In simple shear flow, with the velocity field $v_x = \dot{\gamma}y$, where v_x is the fluid flow field along the x-direction (flow direction), y the gradient direction, and $\dot{\gamma}$ the shear rate, the viscosity η is related to the stress tensor via

$$\sigma_{xy} = \eta\dot{\gamma}. \tag{27}$$

Hence, an expression is required for the stress tensor to either derive η analytically and/or to determine it in simulations. In Refs. 39,48, the kinetic theory moment method has been applied to derive an analytical expression.

5.2.1 Stress Tensor

In this lecture note, an expression for the stress tensor is obtained by the virial theorem[35,49] starting from the equation of motion of particle i

$$\ddot{r}_{i\beta} = F_{i\beta} \ , \tag{28}$$

where the force $\boldsymbol{F}_i$ will be specified later. For a system with periodic boundary conditions, $\boldsymbol{r}_i$ refers to the position of the particle in the infinite system, i.e., we do not jump to an image, which is located in the primary box (see Fig. 6), when a particle crosses a boundary of the periodic lattice. Hence, $\boldsymbol{r}_i$ is a continuous function of time. Multiplication of Eq. 28 by $r_{i\beta'}$ and summation over all N particles yields

$$\frac{d}{dt}\sum_{i=1}^{N} m_i v_{i\beta} r_{i\beta'} = \sum_{i=1}^{N} m_i v_{i\beta} v_{i\beta'} + \sum_{i=1}^{N} F_{i\beta} r_{i\beta'}. \tag{29}$$

The average over time (or an ensemble) yields

$$\left\langle \sum_{i=1}^{N} m_i v_{i\beta} v_{i\beta'} \right\rangle + \left\langle \sum_{i=1}^{N} F_{i\beta} r_{i\beta'} \right\rangle = 0, \tag{30}$$

because the term on the left hand side of Eq. 29 vanishes for a diffusive or confined system[50,51]. Eq. 30 is a generalization of the virial theorem[49,50].

In the presence of shear flow, the time average of the left-hand side of Eq. 29 does not vanish anymore[35]. In order to arrive at a vanishing term, we subtract the derivative of the velocity profile $d(\dot{\gamma} r_{iy})/dt = \dot{\gamma} v_{iy}$ from both sides of Eq. 28. This leads to the modified equation

$$\frac{d}{dt}\sum_{i=1}^{N} m(v_{ix} - \dot{\gamma} r_{iy}) r_{iy} = \sum_{i=1}^{N} m(v_{ix} - \dot{\gamma} r_{iy}) v_{iy} + \sum_{i=1}^{N} F_{ix} r_{iy} - \dot{\gamma} \sum_{i=1}^{N} m v_{iy} r_{iy} \tag{31}$$

and

$$\sum_{i=1}^{N} \langle m(v_{ix} - \dot{\gamma} r_{iy}) v_{iy} \rangle + \sum_{i=1}^{N} \langle F_{ix} r_{iy} \rangle - \dot{\gamma} \sum_{i=1}^{N} \langle m v_{iy} r_{iy} \rangle = 0 \tag{32}$$

in the flow-gradient plane.

For the MPC method, the force on a particle can be expressed as

$$\boldsymbol{F}_i(t) \equiv \boldsymbol{F}_i^c(t) = \sum_{q=0}^{\infty} \Delta \boldsymbol{p}_i(t) \delta(t - t_q), \tag{33}$$

where $\Delta \boldsymbol{p}_i(t) = m(\boldsymbol{v}_i(t) - \hat{\boldsymbol{v}}_i(t))$ is the change in momentum during collision [Eq. 2]; $\hat{\boldsymbol{v}}_i(t)$ denotes the velocity after streaming and before the collision. Note that the velocities are not necessarily constant during the streaming step due to an external field. The required time average is defined as follows

$$\begin{aligned} \left\langle F_{i\beta}^c r_{i\beta'} \right\rangle &= \lim_{t\to\infty} \frac{1}{t} \int_0^t F_{i\beta}^c(t') r_{i\beta'}(t') \, dt' = \lim_{N_s\to\infty} \frac{1}{N_s h} \sum_{q=1}^{N_s} \Delta p_{i\beta}(t_q) r_{i\beta'}(t_q) \\ &= \frac{1}{h} \left\langle \Delta p_{i\beta} r_{i\beta'} \right\rangle_{N_s} . \end{aligned} \tag{34}$$

The last line defines an average over collision steps N_s. Denoting the position (image or real) of a particle in the primary box by $\boldsymbol{r}'_i(t)$, the particle position itself is given by $\boldsymbol{r}_i(t) = \boldsymbol{r}'_i(t) + \boldsymbol{R}_i(t)$, where $\boldsymbol{R}_i(t) = (n_{ix}L_x, n_{iy}L_y, n_{iz}L_z)^T$ is the lattice vector at time t. The $n_{i\beta}$ are integer numbers and L_β denotes the box length along the Cartesian axis β. Applying these definitions, the velocity terms of Eq. 32 become

$$\begin{aligned} \langle (v_{ix} - \dot{\gamma} r_{iy}) v_{iy} \rangle &= \langle \hat{v}_{iy} \hat{v}'_{ix} \rangle_{N_s} + \frac{\dot{\gamma} h}{2} \left\langle \hat{v}_{iy}^2 \right\rangle_{N_s}, \\ \langle v_{iy} r_{iy} \rangle &= \frac{1}{2} \left\langle (v_{iy} + \hat{v}_{iy}) r_{iy} \right\rangle_{N_s} \end{aligned} \tag{35}$$

in the stationary state. v'_{ix} denotes the velocity in the primary simulation box, i.e., $v_{ix} = v'_{ix} + \dot{\gamma} R_{iy}$. Note that the expression $\langle (\hat{v}_{ix} - \dot{\gamma} r_{iy}) \hat{v}_{iy} \rangle_{N_s}$ reduces to $\langle \hat{v}'_{ix} \hat{v}_{iy} \rangle_{N_s}$, because the average $\left\langle \hat{v}_{iy} r'_{iy} \right\rangle_{N_s}$ vanishes. The particle velocities along the other spatial directions are identical for each periodic image.

We now define instantaneous external σ^e_{xy} and internal σ^i_{xy} stress tensors according to

$$\sigma^e_{xy} = \frac{1}{Vh} \sum_{i=1}^{N} \Delta p_{ix} R_{iy} - \frac{\dot{\gamma}}{2V} \sum_{i=1}^{N} m (v_{iy} + \hat{v}_{iy}) R_{iy}, \tag{36}$$

$$\sigma^i_{xy} = -\frac{1}{V} \sum_{i=1}^{N} m \hat{v}'_{ix} \hat{v}_{iy} - \frac{\dot{\gamma} h}{2V} \sum_{i=1}^{N} m v_{iy}^2 - \frac{1}{Vh} \sum_{i=1}^{N} \Delta p_{ix} r'_{iy}, \tag{37}$$

which obey the relation $\langle \sigma^i_{xy} \rangle_{N_s} = \langle \sigma^e_{xy} \rangle_{N_s}$. Eq. 36 corresponds to the mechanical definition of the stress tensor as force/area, since $R_{iy} \sim L_y$, and Eq. 37 corresponds to the momentum flux across a surface[52]. Correspondingly, the external stress tensor includes only force terms, i.e., collisional contributions, whereas the internal stress tensor comprises kinetic and collisional contributions. The term $\sim \dot{\gamma}$ in σ^i_{xy} results from the streaming dynamics and vanish in the limit $h \to 0$. Since a discrete time dynamics is fundamental for the MPC method, the collision time will always be finite. Expressions for the stress tensors in the presence of walls are presented in Ref. 35.

An example of the time dependence of the internal and external stress tensors, i.e., $\langle \sigma^i_{xy} \rangle_{N_s}$, $\langle \sigma^e_{xy} \rangle_{N_s}$, under shear is shown in Fig. 3. Both expressions approach the same limiting value in the asymptotic limit. Thereby, the fluctuations of the external stress tensor component are larger.

5.2.2 Viscosity of MPC Fluid: Analytical Expressions

The derived expressions for the stress tensors are independent of any particular collision rule. The viscosity of a system, however, depends on the applied collision procedure. Analytical expressions for the viscosity of an MPC fluid have been derived by various approaches[2, 14, 18, 19, 22, 39, 35, 47, 48].

In simple shear flow, the viscosity η is given by Eq. 27, where the (macroscopic) stress tensor follows from $\sigma_{xy} = \langle \sigma^i_{xy} \rangle_{N_s} = \langle \sigma^e_{xy} \rangle_{N_s}$. For a MPC fluid, the stress tensor is composed of a kinetic and collisional contribution[2, 14, 18, 19, 22, 39, 35], i.e, $\sigma_{xy} = \sigma^{\mathrm{kin}}_{xy} + \sigma^{\mathrm{col}}_{xy}$, which implies that the viscosity $\eta = \eta_{\mathrm{kin}} + \eta_{\mathrm{col}}$ consists of a kinetic η_{kin} and collisional η_{col} part too. For a system with periodic boundary conditions, the two contributions are conveniently obtained from the internal stress tensor (37).

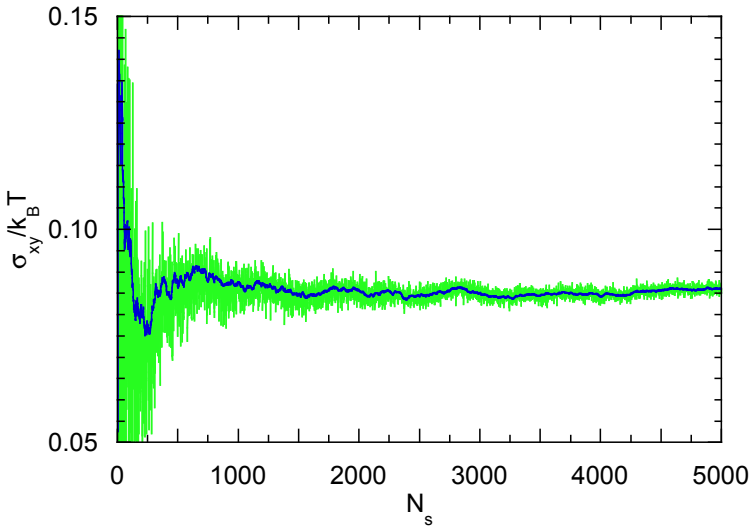


Figure 3. Internal $\langle \sigma^i_{xy} \rangle_{N_s}$ (blue) and external $\langle \sigma^e_{xy} \rangle_{N_s}$ (green, large fluctuations) stress tensor components as function of the number of collision steps N_s. The collision time is $h/\sqrt{ma^2/(k_BT)} = 0.1$. At $t = 0$, the system is in a stationary state[35].

The kinetic contribution $\eta_{\rm kin}$ is determined by streaming, i.e., the velocity dependent terms in Eq. 37. To find the mean $\langle \hat{v}'_{ix} \hat{v}_{iy} \rangle_{N_s}$, we consider a complete MPC dynamics step. The correlation $\langle v'_{ix}(t) v_{iy}(t) \rangle_{N_s}$ before streaming is related to that after streaming $\langle \hat{v}'_{ix}(t+h) \hat{v}_{iy}(t+h) \rangle_{N_s}$ via

$$\begin{aligned} &\langle \hat{v}'_{ix}(t+h) \hat{v}_{iy}(t+h) \rangle_{N_s} = \langle [\hat{v}_{ix}(t+h) - \dot{\gamma} r_{iy}(t+h)] \hat{v}_{iy}(t+h) \rangle_{N_s} \\ &= \langle [v_{ix}(t) - \dot{\gamma} r_{iy}(t)] v_{iy}(t) \rangle_{N_s} - \dot{\gamma} h \left\langle v_{iy}(t)^2 \right\rangle_{N_s} = \langle v'_{ix}(t) v_{iy}(t) \rangle_{N_s} - \dot{\gamma} h \left\langle v^2_{iy} \right\rangle_{N_s} . \end{aligned} \tag{38}$$

Note that the average comprises both, a time average and an ensemble average over the orientation of the rotation axis. The velocities after streaming are changed by the subsequent collisions, which yields, with the corresponding momenta of the rotation operator $\mathcal{D}(\alpha)$, $\langle v'_{ix}(t) v_{iy}(t) \rangle_{N_s} = f \langle \hat{v}'_{ix}(t) \hat{v}_{iy}(t) \rangle_{N_s}$ and $f = 1 + (1 - 1/N_c)(2\cos(2\alpha) + 2\cos\alpha - 4)/5$[39,22,35]. Note, velocity correlations between different particles are neglected, i.e., molecular chaos is assumed. Thus, in the steady stead $[\langle \hat{v}'_{ix}(t) v'_{iy}(t) \rangle_{N_s} = \langle \hat{v}'_{ix}(t+h) v'_{iy}(t+h) \rangle_{N_s}]$, we find

$$\langle \hat{v}'_{ix} v_{iy} \rangle_{N_s} = -\frac{\dot{\gamma} h}{1-f} \left\langle v^2_{iy} \right\rangle_{N_s} \tag{39}$$

by using Eq. 38. Hence, with the equipartition of energy $\left\langle v^2_{iy} \right\rangle_{N_s} = k_BT/m$, the kinetic viscosity is given by

$$\eta_{\rm kin} = \frac{N k_B T h}{V} \left[\frac{5 N_c}{(N_c - 1)(4 - 2\cos\alpha - 2\cos(2\alpha))} - \frac{1}{2} \right] . \tag{40}$$

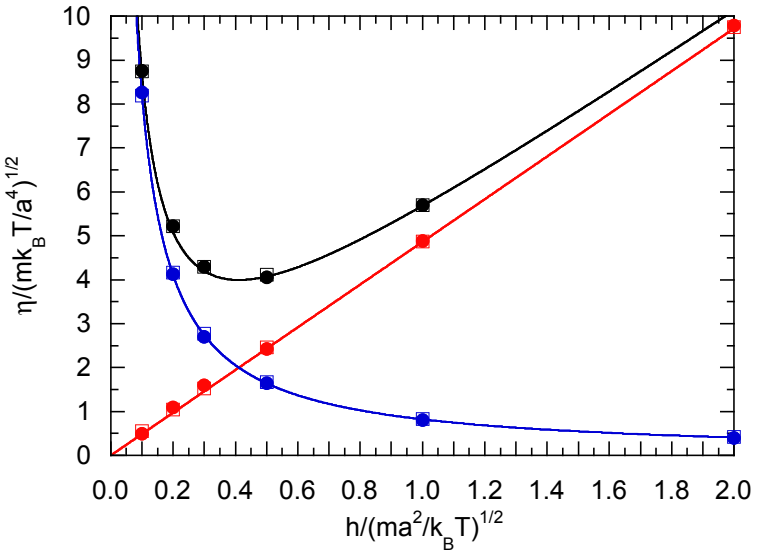


Figure 4. Viscosities determined via the internal (bullets) and external (open squares) stress tensors for a system confined between walls as function of the collision time. The analytical results for the total (black), the kinetic (red, $\sim h$), and collisional (blue, $\sim 1/h$) contributions are presented by solid lines.

The collisional viscosity $\eta_{\rm col}$ is determine by the momentum change of the particles during the collision step. Since the collisions in the various cells are independent, it is sufficient to consider one cell only. The positions of the particles within a cell can be expressed as $\boldsymbol{r}_i' = \boldsymbol{r}_c + \Delta\boldsymbol{r}_i$, where $\boldsymbol{r}_c$ is chosen as the center of the cell. Because of momentum conservation, the term $\sum_{i=1}^{N_c} \Delta p_{ix} r_{iy}'$ becomes $\sum_{i=1}^{N_c} \Delta p_{ix} \Delta r_{iy}$. The averages over thermal fluctuations and random orientations of the rotation axis yield

$$\langle \Delta p_{ix} \Delta r_{iy} \rangle_{N_s} = \frac{2m\dot{\gamma}}{3} (\cos\alpha - 1) \left[\left(1 - \frac{1}{N_c} \right) \langle \Delta r_{iy}^2 \rangle_{N_s} - \frac{1}{N_c} \sum_{j \neq i = 1}^{N_c} \langle \Delta r_{iy} \Delta r_{jy} \rangle_{N_s} \right]. \tag{41}$$

The average over the uniform distribution of the positions within an cell yields $\langle \Delta r_{iy} \Delta r_{jy} \rangle_{N_s} = 0$ for $i \neq j$ and

$$\frac{1}{a} \int_{-a/2}^{a/2} \Delta r_{iy}^2 dr_{iy} = \frac{a^2}{12}. \tag{42}$$

Hence, the collisional viscosity is given by

$$\eta_{\rm col} = \frac{Nma^2}{18Vh} (1 - \cos\alpha) \left(1 - \frac{1}{N_c} \right). \tag{43}$$

Here, we assume that the number of particles in a collision cell N_c is sufficiently large ($N_c > 3$) to neglect fluctuations. For a small number of particles, density fluctuations have to be taken into account as explained in Sec. 5.1.

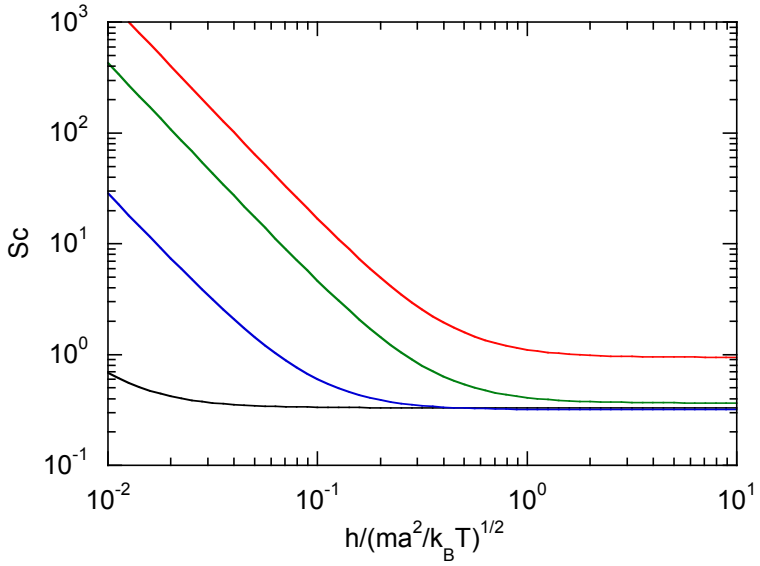


Figure 5. Theoretical Schmidt numbers as function of the collision time step h for the rotation angles $\alpha = 15°$ (black), $45°$ (blue), $90°$ (green), and $130°$ (red). The mean particle number is $\langle N_c \rangle = 10$.

Simulations for various MPC dynamics parameters exhibit very good agreement between the viscosities determined via Eqs. 36, 37 and the analytical expressions Eqs. 40 and 43[39,41]. Fig. 4 displays results for the viscosity determined for an MPC fluid confined between two walls[35]. As shown in Ref. 35, the same analytical expressions are obtained for such a system. For small h, the viscosity is determined by the collisional contribution, whereas for $h \gg 1$, the kinetic contribution dominates. Note that the analytical expression for η_{kin} has been derived assuming molecular chaos, which does not apply for small collision time steps. Hence, there are small deviations between the simulation and analytical results for $h/\sqrt{ma^2/(k_B T)} \lesssim 1$.

5.3 Schmidt Number

A convenient measure of the importance of hydrodynamics is the Schmidt number $Sc = \nu/D$, where $\nu = \eta/(m\langle N_c \rangle)$ is the kinematic viscosity[41]. Thus, Sc is the ratio between momentum transport and mass transport. As is known, this number is smaller than but on the order of unity for gases, while in fluids, like water, it is on the order of 10^2 to 10^3. A prediction for the Schmidt number of a MPC fluid can be obtained from the theoretical expressions (40) and (43) for the viscosity, and the diffusion coefficient (26). In Fig. 5, the theoretical prediction for Sc is displayed for different values of the rotation angle. This shows that Sc becomes considerably larger than unity for $h \to 0$. In fact, Sc increases like $1/h^2$ as soon as the collisional viscosity dominates over the kinetic viscosity.

6 MPC without Hydrodynamics

The importance of HI in complex fluids is generally accepted. A standard procedure for determining the influence of HI is to investigate the same system with and without HI. In order to compare results, however, the two simulations must differ as little as possible – apart from the inclusion of HI. A well-known example of this approach is Stokesian dynamics simulations (SD), where the original Brownian dynamics (BD) method can be extended by including HI in the mobility matrix by employing the Oseen tensor[29,30].

A method for switching off HI in MPC has been proposed in Refs. 15, 39. The basic idea is to randomly interchange velocities of all solvent particles after each collision step, so that momentum (and energy) are *not* conserved *locally*. Hydrodynamic correlations are therefore destroyed, while leaving friction coefficients and fluid self-diffusion coefficients largely unaffected. Since this approach requires the same numerical effort as the original MPC algorithm, a more efficient method has been suggested recently in Refs. 2, 16. If the velocities of the solvent particles are uncorrelated, it is no longer necessary to trace their trajectories. In a random solvent, the solvent-solute interaction in the collision step can thus be replaced by the interaction with a heat bath. This strategy is related to the way no-slip boundary conditions are modeled of solvent particles at a planar wall[32] (see Sec. 3). Since the particle positions within a cell are irrelevant in the collision step, no explicit particles have to be considered. Instead, each monomer of mass $M = m\langle N_c\rangle$ is coupled to an effective solvent momentum $\mathbf{P}$ which is directly chosen from a Maxwell-Boltzmann distribution of variance Mk_BT and a mean given by the average momentum of the fluid field – which is zero at rest, or $(M\dot{\gamma}r_{iy}, 0, 0)$ in the case of an imposed shear flow. The total center-of-mass velocity, which is used in the collision step, is then given by[16]

$$\mathbf{v}_{cm,i} = \frac{M\mathbf{v}_i + \mathbf{P}}{2M} \,. \tag{44}$$

The solute trajectory is determined by MD simulations, and the interaction with the solvent is performed every collision time h.

The relevant parameters of MPC and random MPC are the average number of particles per cell, $\langle N_c\rangle$, the rotation angle α, and the collision time h which can be chosen to be the same. For small values of the density ($\langle N_c\rangle < 5$), fluctuation effects have been noticed[39] and could also be included in the random MPC solvent by a Poisson-distributed density. The velocity autocorrelation functions[41] of a random MPC solvent show a simple exponentially decay, which implies some differences in the solvent diffusion coefficients. Other transport coefficients such as the viscosity depend on HI only weakly[37] and consequently are expected to be essentially identical in both solvents.

7 External Fields

7.1 Shear Flow

To impose shear flow on a periodic MPC solvent system, Lees-Edwards boundary conditions are applied[53,27]. As indicated in Fig. 6, the infinite periodic system is subject to a uniform shear in the xy-plane[29]. The layer of boxes with the primary box is stationary, whereas the layer above moves with the velocity $u = \dot{\gamma}L_y$ to the right and the layer below

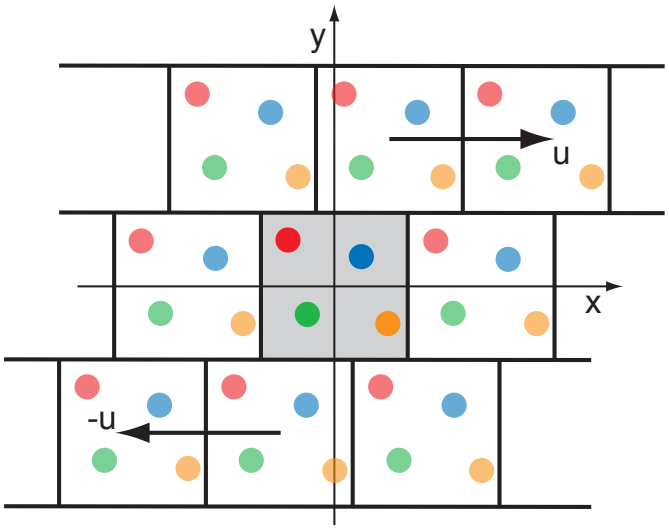


Figure 6. Lees-Edwards homogeneous shear boundary conditions. The primary box is highlighted in gray. The opaque particles are periodic images of the particles of the primary box. The upper layer is moving with the velocity $u = \dot{\gamma} L_y$ to the right, and the bottom layer to the left. Note that the shear velocity is zero in the center of the primary box. See also Ref. 29.

with $-u$ to the left. The corresponding further layers move with the respective integral multiple of u. However, these further layers are not required in practice. Whenever a MPC particle leaves the primary box, it is replaced by its periodic image. This avoids build-up of a substantial difference in the x-coordinates[29]. In the simulation program, the boundary crossing is efficiently handled as follows[29]

```
cory  = anint(ry(i)/ly)
rx(i) = rx(i) - cory*delrx
rx(i) = rx(i) - anint(rx(i)/lx)*lx
ry(i) = ry(i) - cory*ly
rz(i) = rz(i) - anint(rz(i)/lz)*lz
vx(i) = vx(i) - cory*delvx
```

Here, `delvx` $= u$ and `delrx` stores the displacement of the upper box layer. `anint` provides the nearest whole number, i.e., it rounds the argument. Note the change in velocity. The results shown in Fig. 3 for the stress tensor are obtained by applying these boundary conditions.

When walls are present, shear flow can be imposed by the opposite movement of the confining walls with the velocities $u = \pm\dot{\gamma} L/2$ (the reference frame is fixed in the center of the simulation box). Here, shear is imposed in two ways, by applying bounce-back boundary conditions, i.e., the momentum of a particle changes as $\Delta \boldsymbol{p}_i = -2m\boldsymbol{v}_i + 2m\boldsymbol{u}$ ($\boldsymbol{u} = (u, 0, 0)^T$), and by the virtual wall particles[35]. There momenta are determined from a Boltzmann distribution as described in Sec. 3, only along the flow direction the extra

momentum is added

$$p_u = mN_p \left(u + \frac{\dot{\gamma}}{2} \Delta y \right) \tag{45}$$

for a surface at $+L/2$. For the surface at $-L/2$, $u \to -u$ and $\Delta y \to -(a - \Delta y)$ for a given random shift. Δy is the fraction of the wall-truncated collision cell insight the wall and N_p denotes the number of virtual particles[35]. The viscosities of Fig. 4 have been determined applying this scheme.

7.2 Poiseuille Flow

A parabolic flow profile of a fluid confined between walls is obtained by a constant pressure gradient or a uniform body force, e.g., gravitational force, combined with non-slip boundary conditions. For two planer walls parallel to the xz-plane at $y = 0$ and $y = L_y$, the Stokes equation yields the velocity profile

$$v_x(y) = \frac{4 v_{\max} y (L_y - y)}{L_y^2}, \qquad \text{with} \qquad v_{\max} = \frac{m \langle N_c \rangle g L_y^2}{8\eta}. \tag{46}$$

$m \langle N_c \rangle g$ is the gravitational (volume) force density[32].

In MPC simulations a parabolic flow profile is obtained in a similar manner. Here, the same geometry is considered as in Sec. 7.1, with periodic boundary conditions parallel to the walls, and every fluid particle is exposed to the gravitational force $F_x = mg$ along the x-direction. Naturally, other channel geometries, such as square channels[54] or capillaries[55–57] can be considered. Then, the particle velocities and positions are updated according to

$$\begin{aligned} v_{i\beta}(t+h) &= v_{i\beta}(t) + gh\delta_{\beta x}, \\ r_{i\beta}(t+h) &= r_{i\beta}(t) + v_{i\beta}(t)h + \frac{1}{2} g h^2 \delta_{\beta x} \end{aligned} \tag{47}$$

in the streaming step. The bounce-back rule has to be adjusted too. This is simply done after the streaming step (47) is complete. The velocities and positions of the particles who penetrated into a wall are corrected according to

$$\begin{aligned} \hat{v}_{i\beta}(t+h) &= -v_{i\beta}(t+h) + 2g\Delta h_i \delta_{\beta x}, \\ \hat{r}_{i\beta}(t+h) &= r_{i\beta}(t+h) - 2v_{i\beta}(t+h)\Delta h_i \delta_{\beta x} + 2g\Delta h_i^2 \delta_{\beta x}. \end{aligned} \tag{48}$$

The time Δh_i, during which the particle moves insight the wall, follows from the dynamics along the y-direction: $\Delta h_i = [r_{iy}(t+h) - L_y \Theta(r_{iy}(t+h) - L_y)]/v_{iy}(t+h)$, where $\Theta(x)$ is the Heaviside function.

Fig. 7 show velocity profiles for various thermalization procedures. The results are obtained for the system parameters $L_x = L_y = L_z = 20a$, $h/\sqrt{ma^2/(k_B T)} = 0.1$, $\alpha = 130°$, $\langle N_c \rangle = 10$, and $g/(k_B T/(ma)) = 0.01$[36]. Evidently, a parabolic profile is obtained, which however depends on the way the system is thermalized. Note that, without explicit thermostat, the fluid is thermalized via the virtual particles in the walls. As Fig. 8 shows, an inadequate thermostat leads to inhomogeneous energy and particle density profiles[36,58]. A constant energy and particle density is obtained for the local Maxwellian thermostat presented in Sec. 4.

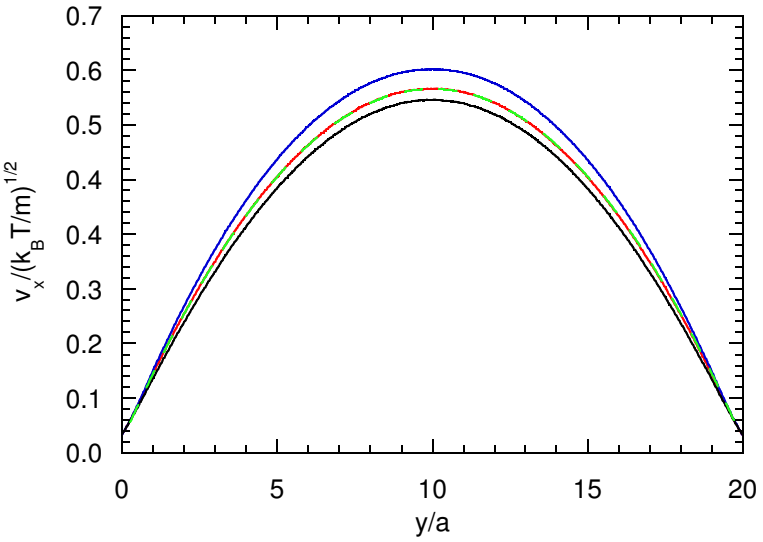


Figure 7. Velocity profiles for a MPC fluid confined between two parallel walls. Bottom (black) line: The fluid is thermalized by the surfaces only. Top (blue) line: The fluid is thermalized by the global thermostat Eq. 12 and the surfaces. Middle (red) line: The fluid is thermalized by the local thermostat (15) and the surfaces. The green dashed lines is a fit of the parabolic profile (46), which yields the viscosity $\eta/\sqrt{mk_BT/a^4} = 8.9$ and a finite slip length $l_s/a = 0.176$.

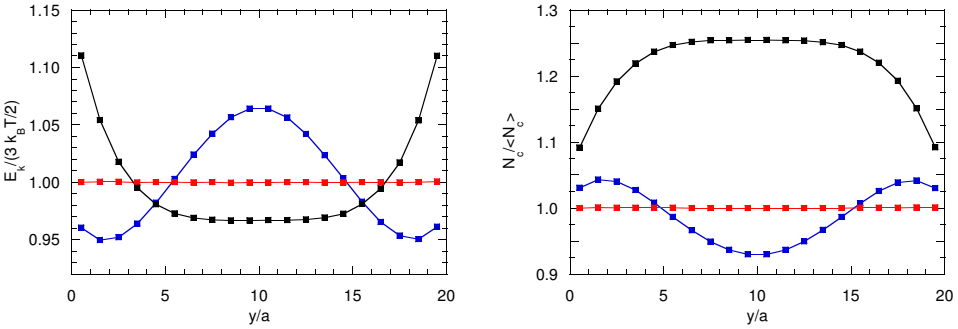


Figure 8. Kinetic energy of fluid particles (left) and mean particle number in a collision cell (right) perpendicular to the confining walls for systems, which are thermalized by the surfaces (black), the global thermostat (12) (blue), and the local thermostat (15) (red).

There is a finite slip at the walls visible in Fig. 7[32,36,58]. The reason is that the average center-of-mass velocities of the cells intersected by walls are not zero but positive (see Sec. 3). A zero velocity can easily be achieved in a linear velocity profile, i.e., in shear flow[35], but would require corrections to the proposed scheme of treating the virtual particles for non-linear flow profiles. A fit of the parabolic velocity profile $v_x \sim (y + l_s)(L_y + l_s - y)$ (46), with the slip length l_s, yields $l_s/a = 0.176$ and the vis-

cosity $\eta/\sqrt{mk_BT/a^4} = 8.9$. This value agrees with the value obtained from shear flow simulations $\eta/\sqrt{mk_BT/a^4} = 8.8$ (see Sec. 5.2), and both are close to the theoretical vale $\eta/\sqrt{mk_BT/a^4} = 8.7$. Note, the theoretical value is somewhat smaller, because the collisional contribution to viscosity is only calculated within the molecular chaos assumption. Looking at the velocity distribution of the system locally thermalized by Maxwellian distributed energies, we find excellent agreement with the Maxwell-Boltzmann distribution[36].

7.3 Gravitational Field

So far, the external field is explicitly interacting with the MPC fluid. In sedimentation or electrophoresis, the field typically interacts with the solute particles only[59–62]. Here, the solute particles, which are dragged by the external field, induce a motion of the MPC fluid. For a system confined between impenetrable walls, this leads to backflow effects[62,63], since fluid in front of the moving solute particles is reflected from the wall and moving in opposite direction to the solute particles. In systems with periodic boundary conditions, there is also a backflow effect, which is obtained as follows.

The equations of motion of (point-like) solute particles exposed to a gravitational field $\boldsymbol{F}_g = M\boldsymbol{g}$ are given by ($k = 1, \ldots, N_m^t$)

$$M\ddot{\boldsymbol{R}}_k(t) = \boldsymbol{F}_k(t) + \boldsymbol{F}_k^c(t) + M\boldsymbol{g}, \tag{49}$$

where $\boldsymbol{F}_k$ denotes the forces between solute particles and $\boldsymbol{F}_k^c$ the forces due to MPC collisions [Eq. 33]. The equations of motion of the MPC particles are given by Eq. 28 with the forces (33). Summation over all solvent and solute particles yields the equation of motion for the center-of-mass velocity $\boldsymbol{v}_{\mathrm{tot}}$ of the total system

$$(MN_m^t + mN)\dot{\boldsymbol{v}}_{\mathrm{tot}} = \sum_{i=1}^{N} m\ddot{\boldsymbol{r}}_i + \sum_{k=1}^{N_m^t} M\ddot{\boldsymbol{R}}_k = MN_m^t\boldsymbol{g}. \tag{50}$$

The sum over the (pairwise) solvent-solvent forces vanishes, as well as the MPC collisional forces due to momentum conservation. Hence,

$$\boldsymbol{v}_{\mathrm{tot}} = \frac{MN_m^t\boldsymbol{g}}{MN_m^t + mN}t, \tag{51}$$

when the total momentum is zero at $t = 0$. I.e., the center-of-mass velocity increases linearly in time. We want to adopt a reference frame, where the center-of-mass velocity is zero[59–61]. Hence, $\boldsymbol{v}_{\mathrm{tot}}$ is subtracted from every velocity: $\boldsymbol{v}_i' = \boldsymbol{v}_i - \boldsymbol{v}_{\mathrm{tot}}$ and $\boldsymbol{V}_k' = \boldsymbol{V}_k - \boldsymbol{v}_{\mathrm{tot}}$. The equations of motion of the primed variables are given by

$$m\ddot{\boldsymbol{r}}_i' = \boldsymbol{F}_i^c - \frac{mMN_m^t}{MN_m^t + mN}\boldsymbol{g}, \tag{52}$$

$$M\ddot{\boldsymbol{R}}_k' = \boldsymbol{F}_k(t) + \boldsymbol{F}_k^c(t) + \frac{mMN}{MN_m^t + mN}\boldsymbol{g}. \tag{53}$$

The total momentum in this reference frame is evidently zero. In the streaming step, the velocities and positions of the fluid particles are then updated according to (omitting the

prime)

$$\boldsymbol{v}_i(t+h) = \boldsymbol{v}_i(t) - \frac{MN_m^t}{MN_m^t + mN}\boldsymbol{g}h, \tag{54}$$

$$\boldsymbol{r}_i(t+h) = \boldsymbol{r}_i(t) + \boldsymbol{v}_i(t)h - \frac{MN_m^t}{2(MN_m^t + mN)}\boldsymbol{g}h^2. \tag{55}$$

The dynamics of the solute particles is treated by MD[27,28]. There is a flux of MPC particles opposite to the flux of solute particles, i.e., backflow is present.

8 Hydrodynamic Simulations of Polymers in Flow Fields

As an example of a complex fluid in a flow field, I will briefly touch the nonequilibrium properties of a linear polymer in shear flow. Shear flow is a paradigmatic case, where the polymer dynamics can be studied in a stationary nonequilibrium state. The MPC dynamics approach has been shown to properly account for HI in polymer systems[25,26,64] and provides thus an excellent way to incorporate fluid properties. We adopt a hybrid simulation approach, combining MPC for the solvent with molecular dynamics simulations for the polymer molecule, where the two are coupled in the collision step according to Eq. 9 (see Sec. 3).

Single molecule experiments reveal a remarkably reach structural and dynamical behavior of individual polymers in flow fields[65,66]. In particular, fluorescence microscopy studies on single DNA molecules in shear flow find large conformational changes due to tumbling motion[65–68]. A polymer chain continuously undergoes stretching and compression cycles and never reaches a steady-state extension. The detailed evolution itself depends upon the shear rate. By the same experimental technique, valuable quantitative information has been obtained for the non-equilibrium properties of DNA molecules, such as their deformation, orientation, and viscosity, both, for free and tethered molecules[65,67–72].

8.1 Model

The polymer is comprised of N_m beads of mass M, which are connected by linear springs[64]. The bond potential is

$$U_l = \frac{\kappa_l}{2}\sum_{k=1}^{N_m-1}\left(|\boldsymbol{R}_{k+1} - \boldsymbol{R}_k| - l\right)^2, \tag{56}$$

where l is the bond length, κ_l the spring constant, and $\boldsymbol{R}_k$ the position of monomer k. Excluded-volume interactions are taken into account by the shifted and truncated Lennard-Jones potential

$$U_{LJ}(r) = 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} + \frac{1}{4}\right] \tag{57}$$

for monomer distances $r < \sqrt[6]{2}\sigma$ and $U_{LJ} = 0$ otherwise. The monomer dynamics is determined by Newton's equations of motion, which are integrated by the velocity Verlet algorithm with time step h_p[28,27].

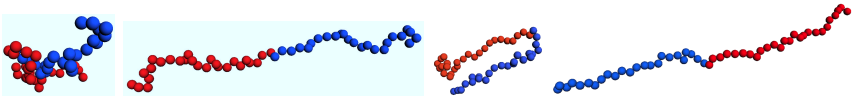


Figure 9. Sequence of snapshots illustrating the conformational changes of a polymer of length $N_m = 50$ in shear flow during a tumbling cycle.

Three-dimensional periodic boundary conditions are considered with Lees-Edwards boundary conditions to impose shear flow[27] (see Sec. 7.1). The local Maxwellian thermostat, as described in Sec. 4, is used to maintain a constant temperature. We employ the parameters $\alpha = 130°$, $h/\sqrt{ma^2/(k_BT)} = 0.1$, $\langle N_c \rangle = 10$, $M = m\,\langle N_c \rangle$, $l = \sigma = a$, $k_BT/\epsilon = 1$, $h/h_p = 50$, the bond spring constant $\kappa_l = 5 \times 10^3 k_BT/a^2$, the mass density $\varrho = m\langle N_c \rangle$, and the polymer length $N_m^t = 50$. In dilute solution, the equilibrium end-to-end vector relaxation time of this polymer is $\tau_0/\sqrt{ma^2/(k_BT)} = 6169$.[64]

8.2 Conformations

Fig. 9 shows a sequence of snapshots illustrating the conformations and the tumbling dynamics. Starting from a coiled state, the flow field stretches the polymer – the angle φ between the end-to-end vector and flow direction is positive (for the definition of φ, see Fig. 11) – and the polymer is aligned. Thermal fluctuations cause the polymer orientation angle to become negative, i.e., $\varphi < 0$, which leads to a polymer collapse. Later the angle becomes positive again, the polymer stretches, and the cycle starts again.

The conformational properties are characterized by the radius of gyration tensor

$$\langle G_{\beta\beta'} \rangle = \frac{1}{N_m} \sum_{k=1}^{N_m} \langle \Delta R_{k,\beta} \Delta R_{k,\beta'} \rangle \,, \tag{58}$$

where $\Delta \boldsymbol{R}_k$ is the monomer position in the center-of-mass reference frame. The ratios of the diagonal components $\langle G_{\beta\beta} \rangle / \langle G^0_{\beta\beta} \rangle$, $\langle G^0_{\beta\beta} \rangle = R_G^2/3$ is the equilibrium value, with R_G^2 the radius of gyration, are displayed in Fig. 10 (left). A significant polymer stretching along the flow direction appears for $\mathrm{Wi} > 1$, where Wi is the Weissenberg number, defined as $\mathrm{Wi} = \dot{\gamma}\tau_0$. At large shear rates, the stretching saturates at a maximum, which is smaller than the value corresponding to a fully stretched chain ($\langle G_{xx} \rangle \approx l^2 N_m^2/12$) and reflects the finite size of a polymer. This is consistent with experiments on DNA[67,70], where the maximum extension is on the order of half of the contour length, and theoretical calculations[73]. It is caused by the large conformational changes of polymers in shear flow, which yields an average extension smaller than the contour length. Nevertheless, molecules assume totally stretched conformations at large Weissenberg numbers during their tumbling cycles. In the gradient and the vorticity directions, the polymers shrink, with a smaller shrinkage in the vorticity direction due to excluded-volume interactions[64].

8.3 Alignment

The deformation is associated with a preferred alignment of a polymer. This is typically characterized by the angle χ between the main axis of the gyration tensor and the flow

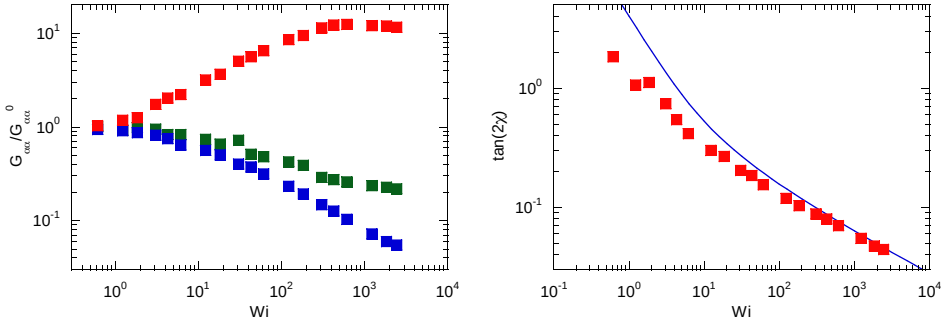


Figure 10. Shear rate dependence of the gyration tensor components along the flow (red), gradient (blue), and vorticity direction (green) (left). Dependence of the alignment angle on the Weissenberg number (right). The solid line is obtained from the theoretical expression of Ref. 74.

direction[64,70,73]. It is obtained from the components of the gyration tensor via[64]

$$\tan(2\chi) = \frac{2\,\langle G_{xy}\rangle}{\langle G_{xx}\rangle - \langle G_{yy}\rangle}. \tag{59}$$

The dependence of $\tan(2\chi)$ on shear rate is shown in Fig. 10 (right). In the limit $\mathrm{Wi} \to 0$, theory[74,73] predicts $\tan(2\chi) \sim \mathrm{Wi}^{-1}$, which seems to be in qualitative agreement with the simulation data. However, there is a quantitative difference, which might be due to excluded-volume interactions not taken into account in the analytical calculations. For larger Weissenberg numbers, excluded-volume interactions seem to be of minor importance. Here, $\tan(2\chi)$ decreases asymptotically as $\mathrm{Wi}^{-1/3}$.

Fig. 11 shows probability distributions of the angle φ. (For the definition of φ, see Fig. 11.) The distribution function $P(\varphi)$ exhibits a maximum at $\tan(2\varphi_m) \approx \tan(2\chi)$. Hence, φ_m is very close to the angle χ of Eq. 59. The width $\Delta\varphi$ of the distribution function depends on the Weissenberg number and decreases with increasing Wi. In the limit $\mathrm{Wi} \to \infty$, the asymptotic dependence $\Delta\varphi \sim \mathrm{Wi}^{-1/3}$ is obtained for the full width at half maximum[68,74,75]. The polymer model of Refs.[74,73] predicts the same dependence on the Weissenberg number, only certain numerical factors are different. Evidently, the theoretical curves are in excellent agreement with the simulation data.

8.4 Tumbling dynamics

The distribution function $P(\varphi)$ is strongly linked to the tumbling dynamics of a polymer. The existence of such a cyclic motion is not a priori evident from the theoretical model. $P(\varphi)$ does not provide any hint on a periodic motion. Only experiments and computer simulations reveal the presence of a cyclic dynamics. $P(\varphi)$ reveals that the polymer is not rigidly oriented in the flow-gradient plane, but the end-to-end vector fluctuates. The fact that also negative φ values are assumed points toward a reorientation of the bond vector. Tumbling is a consequence of the fact that shear flow is a superposition of a rotational and an extensional flow. It is the rotational part that leads to reorientation. A polymer in elongational flow behaves very differently, in particular its orientation is fixed along the

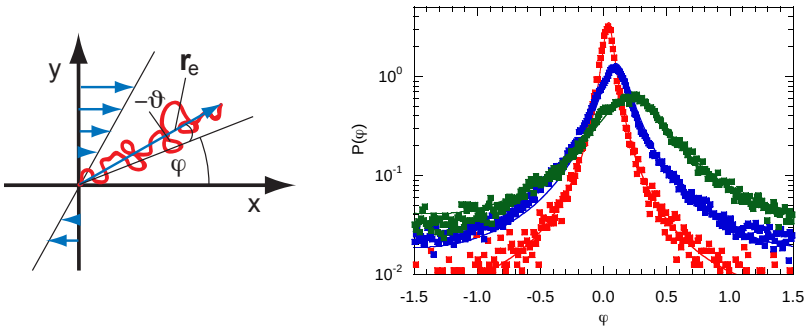


Figure 11. Orientation of the polymer end-to-end vector $\boldsymbol{r}_e$ (left). ϑ is the angle between the bond vector $\boldsymbol{r}_e = |\boldsymbol{r}_e|(\cos\varphi\cos\vartheta, \sin\varphi\cos\vartheta, \sin\vartheta)^T$ and its projection onto the flow-gradient plane and φ is the angle between this projection and the flow direction. Probability distributions of the angle φ for the Weissenberg numbers $\mathrm{Wi} = 617$ (red), $\mathrm{Wi} = 62$ (blue), and $\mathrm{Wi} = 12.3$ (green) (right). The lines are calculated by the theoretical expressions of Ref. 74.

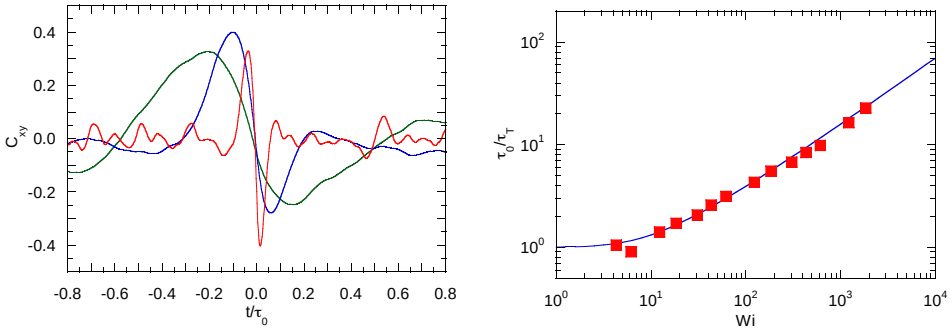


Figure 12. Cross-correlation functions (60) of the gyration tensor components along the flow and gradient direction for the Weissenberg numbers $\mathrm{Wi} = 617$ (red), $\mathrm{Wi} = 62$ (blue), and $\mathrm{Wi} = 12.3$ (green) (left). Normalized tumbling times τ_T as function of Weissenberg number (right)[64,73,74].

flow direction aside from thermal fluctuations[76,77].

The tumbling time can be obtained from the correlation function

$$C_{xy}(t) = \frac{\langle G'_{xx}(t_0) G'_{yy}(t_0+t)\rangle}{\sqrt{\langle G'^2_{xx}(t_0)\rangle \langle G'^2_{yy}(t_0)\rangle}}, \tag{60}$$

where $G'_{\beta\beta}(t) = G_{\beta\beta}(t) - \langle G_{\beta\beta}\rangle$ denotes the deviation from the average stationary value of the gyration tensor. The correlation function captures the time dependent correlations in the deformation along the flow and gradient direction. As shown in Fig. 12, a correlation function exhibits a pronounced maximum at negative lag time (t_-) and a deep minimum at positive lag time (t_+). In the limit $t \to \pm\infty$, the correlation function vanishes. The tumbling time is then defined as $\tau_T = 2(t_+ - t_-)$[69]. These times nicely follow the theoretical prediction, as is evident from Fig. 12, and indicates that the tumbling time is equal to the polymer relaxation time for a given shear rate. Alternative definitions of the tumbling time lead to the same dependence on the Weissenberg number[67,68].

9 Conclusions

In the short time since Malevanets and Kapral[13,14] introduced the MPC dynamics approach as a particle-based mesoscale simulation technique, the method developed into a versatile tool to study hydrodynamic properties of complex fluids. By now, several collision algorithms have been proposed and employed, and the method has been generalized to describe multi-phase flows and viscoelastic fluids[2]. A major advantage of the algorithm is that it is very straightforward to model the dynamics of embedded particles using a hybrid MPC-MD simulations approach. Results of such studies are in excellent quantitative agreement with both theoretical predictions and results obtained using other simulation techniques. In the future, we will see more applications of the method in non-equilibrium and driven soft-matter systems. Specifically, systems where thermal fluctuations play a major role. Here, the full advantage of the method can be exploited, because the interactions of colloids, polymers, and membranes with the mesoscale solvent can be treated on the same basis.

References

1. H. Löwen, *Colloidal soft matter under external control*, J. Phys.: Condens. Matter, **13**, R415, 2001.
2. G. Gompper, T. Ihle, D. M. Kroll, and R. G. Winkler, *Multi-Particle Collision Dynamics: A particle-based mesoscale simulation approach to the hydrodynamics of complex Fluids*, Adv. Polym. Sci., **221**, 1, 2009.
3. M. Doi, *Onsager's variational principle in soft matter*, J. Phys.: Condens. Matter, **23**, 284118, 2011.
4. P. J. Hoogerbrugge and J. M. V. A. Koelman, *Simulating microscopic hydrodynamics phenomena with Dissipative Particle Dynamics*, Europhys. Lett., **19**, 155, 1992.
5. P. Espanol, *Hydrodynamics from Dissipative Particle Dynamics*, Phys. Rev. E, **52**, 1734, 1995.
6. P. Espanol and P. B. Warren, *Statistical mechanics of Dissipative Particle Dynamics*, Europhys. Lett., **30**, 191, 1995.
7. G. McNamara and G. Zanetti, *Use of the Boltzmann equation to simulate lattice-gas automata*, Phys. Rev. Lett., **61**, 2332, 1988.
8. X. Shan and H. Chen, *Lattice Boltzmann model for simulating flows with multiple phases and components*, Phys. Rev. E, **47**, 1815, 1993.
9. X. He and L.-S. Luo, *Theory of the lattice Boltzmann method: From the Boltzmann equation to the lattice Boltzmann equation*, Phys. Rev. E, **56**, 6811, 1997.
10. G. A. Bird, *Molecular Gas Dynamics and the Direct Simulation of Gas Flows*, Oxford University Press, Oxford, 1994.
11. F. J. Alexander and A. L. Garcia, *The Direct Simulation Monte Carlo Method*, Comp. in Phys., **11**, 588, 1997.
12. A. L. Garcia, *Numerical Methods for Physics*, Prentice Hall, 2000.
13. A. Malevanets and R. Kapral, *Mesoscopic model for solvent dynamics*, J. Chem. Phys., **110**, 8605, 1999.
14. A. Malevanets and R. Kapral, *Solute molecular dynamics in a mesoscopic solvent*, J. Chem. Phys., **112**, 7260–7269, 2000.

15. N. Kikuchi, A. Gent, and J. M. Yeomans, *Polymer collapse in the presence of hydrodynamic interactions*, Eur. Phys. J. E, **9**, 63, 2002.
16. M. Ripoll, R. G. Winkler, and G. Gompper, *Hydrodynamic screening of star polymers in shear flow*, Eur. Phys. J. E, **23**, 349, 2007.
17. E. Allahyarov and G. Gompper, *Mesoscopic solvent simulations: Multiparticle-collision dynamics of three-dimensional flows*, Phys. Rev. E, **66**, 036702, 2002.
18. R. Kapral, *Multiparticle Collision Dynamics: Simulations of complex systems on mesoscale*, Adv. Chem. Phys., **140**, 89, 2008.
19. T. Ihle and D. M. Kroll, *Stochastic rotation dynamics: A Galilean-invariant mesoscopic model for fluid flow*, Phys. Rev. E, **63**, 020201(R), 2001.
20. T. Ihle and D. M. Kroll, *Stochastic rotation dynamics I: Formalism, Galilean invariance, Green-Kubo relations*, Phys. Rev. E, **67**, 066705, 2003.
21. Hiroshi Noguchi, N. Kikuchi, and G. Gompper, *Particle-based mesoscale hydrodynamic techniques*, EPL, **78**, 10005, 2007.
22. H. Noguchi and G. Gompper, *Transport coefficients of off-lattice mesoscale-hydrodynamics simulation techniques*, Phys. Rev. E, **78**, 016706, 2008.
23. Ingo O. Götze, Hiroshi Noguchi, and Gerhard Gompper, *Relevance of angular momentum conservation in mesoscale hydrodynamics simulations*, Phys. Rev. E, **76**, 046705, 2007.
24. A. Malevanets and J. M. Yeomans, *Dynamics of short polymer chains in solution*, Europhys. Lett., **52**, 231–237, 2000.
25. M. Ripoll, K. Mussawisade, R. G. Winkler, and G. Gompper, *Low-Reynolds-number hydrodynamics of complex fluids by Multi-Particle-Collision dynamics*, Europhys. Lett., **68**, 106, 2004.
26. K. Mussawisade, M. Ripoll, R. G. Winkler, and G. Gompper, *Dynamics of polymers in a particle-based mesoscopic solvent*, J. Chem. Phys., **123**, 144905, 2005.
27. M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids*, Clarendon Press, Oxford, 1987.
28. W. C. Swope, H. C. Andersen, P. H. Berens, and K. R. Wilson, *A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters*, J. Chem. Phys., **76**, 637, 1982.
29. M. Doi and S. F. Edwards, *The Theory of Polymer Dynamics*, Clarendon Press, Oxford, 1986.
30. J. K. G. Dhont, *An Introduction to Dynamics of Colloids*, Elsevier, Amsterdam, 1996.
31. S. H. Lee and R. Kapral, *Friction and diffusion of a Brownian particle in a mesoscopic solvent*, J. Chem. Phys., **121**, 11163, 2004.
32. A. Lamura, G. Gompper, T. Ihle, and D. M. Kroll, *Multiparticle collision dynamics: Flow around a circular and a square cylinder*, Europhys. Lett., **56**, 319–325, 2001.
33. Y. Inoue, Y. Chen, and H. Ohashi, *Development of a simulation model for solid objects suspended in a fluctuating fluid*, J. Stat. Phys., **107**, 85, 2002.
34. I. O. Götze, H. Noguchi, and G. Gompper, *Relevance of angular momentum conservation in mesoscale hydrodynamics simulations*, Phys. Rev. E, **76**, 046705, 2007.
35. R. G. Winkler and C.-C. Huang, *Stress tensors of multiparticle collision dynamics fluids*, J. Chem. Phys., **130**, 074907, 2009.

36. C.-C. Huang, A. Chatterji, G. Sutmann, G. Gompper, and R. G. Winkler, *Cell-level canonical sampling by velocity scaling for multiparticle collision dynamics simulations*, J. Comput. Phys., **229**, 168, 2010.
37. E. Tüzel, T. Ihle, and D. M. Kroll, *Dynamic correlations in stochastic rotation dynamics*, Phys. Rev. E, **74**, 056702, 2006.
38. T. Ihle and D. M. Kroll, *Stochastic rotation dynamics II: Transport coefficients, numerics, long time tails*, Phys. Rev. E, **67**, 066706, 2003.
39. N. Kikuchi, C. M. Pooley, J. F. Ryder, and J. M. Yeomans, *Transport coefficients of a mesoscopic fluid dynamics model*, J. Chem. Phys., **119**, 6388–6395, 2003.
40. M. H. Ernst, E. H. Hauge, and J. M. J. van Leeuwen, *Asymptotic time behavior of correlation functions. I. Kinetic terms*, Phys. Rev. A, **4**, 2055, 1971.
41. M. Ripoll, K. Mussawisade, R. G. Winkler, and G. Gompper, *Dynamic regimes of fluids simulated by Multi-Particle-Collision dynamics*, Phys. Rev. E, **72**, 016701, 2005.
42. R. F. A. Dib, F. Ould-Kaddour, and D. Levesque, *Long-time behavior of the velocity autocorrelation function at low densities and near the critical point of simple fluids*, Phys. Rev. E, **74**, 011202, 2006.
43. J. P. Boon and S. Yip, *Molecular Hydrodynamics*, Dover, New York, 1980.
44. J.-P. Hansen and I. R. McDonald, *Theory of Simple Liquids*, Academic Press, London, 1986.
45. M. Belushkin R. G. Winkler and G. Foffi, *Backtracking of colloids: A multiparticle collision dynamics simulation study*, J. Chem. Phys. B., **115**, 14263, 2011.
46. T. Ihle, E. Tüzel, and D. M. Kroll, *Resummed Green-Kubo relations for a fluctuating fluid-particle model*, Phys. Rev. E, **70**, 035701, 2004.
47. T. Ihle, E. Tüzel, and D. M. Kroll, *Equilibrium calculation of transport coefficients for a fluid-particle model*, Phys. Rev. E, **72**, 046707, 2005.
48. C. M. Pooley and J. M. Yeomans, *Kinetic theory derivation of the transport coefficients of stochastic rotation dynamics*, J. Phys. Chem. B, **109**, 6505, 2005.
49. R. Becker, *Theory of Heat*, Springer Verlag, Berlin, 1967.
50. R. G. Winkler, H. Morawitz, and D. Y. Yoon, *Novel molecular dynamics simulations at constant pressure*, Molec. Phys., **75**, 669, 1992.
51. R. G. Winkler and R. Hentschke, *Liquid benzene confined between graphite surfaces. A constant pressure molecular dynamics study*, J. Chem. Phys., **99**, 5405, 1993.
52. R. B. Bird, C. F. Curtiss, R. C. Armstrong, and O. Hassager, *Dynamics of Polymer Liquids*, vol. 2, John Wiley & Sons, New York, 1987.
53. A. W. Lees and S. F. Edwards, *The computer study of transport processes under extreme conditions*, J. Phys. C, **5**, 1921, 1972.
54. L. Cannavacciuolo, R. G. Winkler, and G. Gompper, *Mesoscale simulation of polymer dynamics in microchannel flows*, EPL, **83**, 34007, 2008.
55. H. Noguchi and G. Gompper, *Shape transitions of fluid vesicles and red blood cells in capillary flow*, Proc. Natl. Acad. Sci. USA, **102**, 14159–14164, 2005.
56. J. L. McWhirter, H. Noguchi, and G. Gompper, *Flow-induced clustering and alignment of vesicles and red blood cells in microcapillaries*, Proc. Natl. Acad. Sci. USA, **106**, 6039, 2009.
57. R. Chelakkot, R. G. Winkler, and G. Gompper, *Migration of semiflexible polymers in microcapillary flow*, EPL, **91**, 14001, 2010.

58. J. K. Whitmer and E. Luijten, *Fluid-solid boundary conditions for multiparticle collision dynamics*, J. Phys.: Condens. Matter, **22**, 104106, 2010.
59. J. T. Padding and A. A. Louis, *Hydrodynamic and Brownian fluctuations in sedimenting suspensions*, Phys. Rev. Lett., **93**, 220601, 2004.
60. M. Hecht, J. Harting, T. Ihle, and H. J. Herrmann, *Simulation of claylike colloids*, Phys. Rev. E, **72**, 011408, 2005.
61. S. Frank and R. G. Winkler, *Polyelectrolyte electrophoresis: Field effects and hydrodynamic interactions*, EPL, **83**, 38004, 2008.
62. A. Wysocki, P. Royall, R. G. Winkler, G. Gompper, H. Tanaka, A. van Blaaderene, and Hartmut Löwen, *Direct observation of hydrodynamic instabilities in a driven non-uniform colloidal dispersion*, Soft Matter, **5**, 1340, 2009.
63. A. Wysocki, C. P. Royall, R. G. Winkler, G. Gompper, H. Tanaka, A. van Blaaderen, and H. Löwen, *Multi-particle collision dynamics simulations of sedimenting colloidal dispersions in confinement*, Faraday Discuss., **144**, 245–252, 2010.
64. C.-C. Huang, R. G. Winkler, G. Sutmann, and G. Gompper, *Semidilute polymer solutions at equilibrium and under shear flow*, Macromolecules, **43**, 10107, 2010.
65. D. E. Smith, H. P. Babcock, and S. Chu, *Single polymer dynamics in steady shear flow*, Science, **283**, 1724, 1999.
66. P. LeDuc, C. Haber, G. Boa, and D. Wirtz, *Dynamics of individual flexible polymers in a shear flow*, Nature, **399**, 564, 1999.
67. C. M. Schroeder, R. E. Teixeira, E. S. G. Shaqfeh, and S. Chu, *Characteristic periodic motion of polymers in shear flow*, Phys. Rev. Lett., **95**, 018301, 2005.
68. S. Gerashchenko and V. Steinberg, *Statistics of tumbling of a single polymer molecule in shear flow*, Phys. Rev. Lett., **96**, 038304, 2006.
69. R. E. Teixeira, H. P. Babcock, E. S. G. Shaqfeh, and S. Chu, *Shear thinning and tumbling dynamics of single polymers in the flow-gradient plane*, Macromolecules, **38**, 581, 2005.
70. C. M. Schroeder, R. E. Teixeira, E. S. G. Shaqfeh, and S. Chu, *Dynamics of DNA in the flow-gradient plane of steady shear flow: Observations and simulations*, Macromolecules, **38**, 1967, 2005.
71. P. S. Doyle, B. Ladoux, and J.-L. Viovy, *Dynamics of a tethered polymer in shear flow*, Phys. Rev. Lett., **84**, 4769, 2000.
72. B. Ladoux and P. S. Doyle, *Stretching tethered DNA chains in shear flow*, Europhys. Lett., **52**, 511, 2000.
73. R. G. Winkler, *Conformational and rheological properties of semiflexible polymers in shear flow*, J. Chem. Phys., **133**, 164905, 2010.
74. R. G. Winkler, *Semiflexible polymers in shear flow*, Phys. Rev. Lett., **97**, 128301, 2006.
75. A. Puliafito and K. Turitsyn, *Numerical study of polymer tumbling in linear shear flow*, Physica D, **211**, 9, 2005.
76. T. T. Perkins, D. E. Smith, and S. Chu, *Single polymer dynamics in an elongational flow*, Science, **276**, 2016, 1997.
77. T. Hofmann, R. G. Winkler, and P. Reineker, *Dynamics of a polymer chain in an elongational flow*, Phys. Rev. E, **61**, 2840, 2000.

Dissipative Particle Dynamics

Pep Español

Departamento de Física Fundamental, UNED, Madrid, 28040, Spain *E-mail: pep@fisfun.uned.es*

Dissipative Particle Dynamics is a particle model that allows one to simulate complex fluids and soft matter at mesoscopic scales. Since its introduction twenty years ago it has been applied to an enormous variety of different systems. The conceptual underpinning of the model and its connection with the underlying molecular dynamics is now rather clear. We present a review of the method, some of its extensions, and discuss the theoretical basis for the model. We also present a necessarily brief account of applications.

1 Introduction

Molecular dynamics allows us to simulate realistic dynamics of millions of atoms during nanoseconds with present day computer resources. The largest systems considered up to now contain a trillion (10^{12}) of particles for 40 time steps[1] and the largest time scales explored is a microsecond for a protein system of about 3600 particles[2]. Although these record breaking research suggests the possibilities of MD, it is clear that this is still insufficient when trying to study the behaviour of structured soft matter as occurs in the interior of living cells, for example. Complex fluids provide another instance of the inapplicability of MD due to the large disparity of time scales of the mesostructure dynamics and the atomic dynamics. When the system displays multiple characteristic time scales a brute force MD simulation is completely unfeasible because the time step is limited by the short range repulsive interactions and then one needs an enormous number of steps to explore the larger time scales.

MD is CPU consuming because it provides all the microscopic detail at the shortest time scales. There are situations, though, in which having all the molecular detail is not necessary in order to answer relevant scientific questions. In those cases, what is required is a model composed of a reduced number of computation units, smaller than the total number of atoms in the system but that still capture the phenomena that we are interested in. Those models are named coarse-grain (CG) models. There are two broad families of coarse-grain mesoscopic models, those that use a *lattice* to support the CG variables (like the finite elements/volumes/differences methods for solving elasticity or hydrodynamic field theories or the Lattice Boltzmann Equation for hydrodynamics simulations) and those that use off lattice *particles* that carry the CG information. The number of CG particles or lattice nodes is much smaller than the number of atoms and, besides, the time scale of evolution of the CG variables is much slower, permitting much larger time steps than in MD. The computational gain when using CG models is manifest.

A very popular particle model for the simulation of CG dynamics is Dissipative Particle Dynamics (DPD), which was introduced by Hoogerbrugge and Koelman[3] and was formulated as a proper statistical mechanics model later[4]. The formulation of the model was done by resorting to very simple general principles, like having translational, rotational, and Galilean invariance, and requesting momentum conservation. The DPD model consists on a set of point particles that move off-lattice interacting with each other through a

set of prescribed forces. The forces are of three types: a conservative force deriving from a potential, a dissipative force that tries to reduce velocity differences between the particles, and a further stochastic force directed along the line joining the center of the particles. The stochastic differential equations of motion for the dissipative particles are[4]

$$\begin{aligned} \dot{\mathbf{r}}_i &= \mathbf{v}_i \\ m_i \dot{\mathbf{v}}_i &= -\frac{\partial V}{\partial \mathbf{r}_i} - \sum_j \gamma \omega^D(r_{ij})(\mathbf{v}_{ij} \cdot \mathbf{e}_{ij})\mathbf{e}_{ij} + \sum_j \sigma \omega^R(r_{ij}) \frac{dW_{ij}}{dt} \mathbf{e}_{ij} \end{aligned} \tag{1}$$

where $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$ is the relative distance between particles i, j, $\mathbf{v}_{ij} = \mathbf{v}_i - \mathbf{v}_j$ is the relative velocity and $\mathbf{e}_{ij} = \mathbf{r}_{ij}/r_{ij}$ is the unit vector joining particles i and j. dW_{ij} is an independent increment of the Wiener process. In Eq. 1, γ is a friction coefficient and $\omega^D(r_{ij}), \omega^R(r_{ij})$ are bell-shaped functions with a finite support that render the dissipative interactions local. Validity of the fluctuation-dissipation theorem requires[4] σ and γ to be linked by the relation $\sigma^2 = 2\gamma k_B T$, where k_B is the Boltzmann factor and T is the system temperature, and also $\omega^D(r_{ij}) = [\omega^R(r_{ij})]^2$. As a result, the stationary probability distribution of the DPD model is given by the Gibbs canonical ensemble

$$\rho(\{\mathbf{r}, \mathbf{p}\}) = \frac{1}{Z} \exp\left\{ -\beta \sum_i^N \frac{\mathbf{p}_i^2}{2m_i} - \beta V(\{\mathbf{r}\}) \right\} \tag{2}$$

The potential energy V is a suitable function of the positions of the dissipative particles that is transitionally invariant in order to ensure momentum conservation. No matter what is the specific form of the potential function, a model with local momentum conservation will exhibit a hydrodynamic behaviour at sufficiently large scales[5,6]. The dissipative particles are point particles that "represent" or "capture" the behaviour of many underlying atoms. The versatility of the model relies on the fact that very different potential functions may be employed to model different systems. A solvent is described with soft repulsive particles, we may join particles with springs to model polymers, set repulsion forces with different amplitudes to model phase separation between different particles, perform rigid body movements of groups of particles to model solid objects floating in liquids. The main theme in a CG model, though, is that the particles are not atoms and, therefore, the potential function is not a potential function between atoms, but rather, it is a coarse-grained potential (a free energy to be precise, see later) whose functional form is much softer than the singular impenetrable potentials between atoms. Of course, we may still use Eqs. 1 when the particles are actually atoms if we are interested in sampling the distribution function (2).

2 The Meaning of a Dissipative Particle

A recurrent question when discussing about the DPD model is: What is in fact a dissipative particle? The literature is plagued with suggestive images as "lumps of atoms", "clusters", "groups of atoms moving coherently", but it is only recently that there seems to emerge a clear picture of the meaning of a dissipative particle. The discussion requires the distinction of two very different situations, those in which the dissipative particle represents *bounded* groups of atoms, and those which represents *unbounded* groups of atoms as those

constituting a simple liquid. In the former case, one can resort to the theory of coarse-graining[7,8] and formulate the equations of motion for the *center of mass* of the bounded group of atoms. The resulting equations are similar to those of the DPD model, with explicit expressions in terms of molecular averages for the CG potential and friction forces. We may say, therefore, that we have a clear definition and meaning of what a dissipative particle is when it represent a set of *bounded* atoms.

For the case of dissipative particles modelling simple fluids made of *unbounded* atoms or molecules, the situation is much more difficult because there is no notion of group of atoms that retain its entity as times proceeds. Due to the diffusing nature of the unbounded atoms in a liquid the notion of "the center of mass of a group of atoms" is a rather fuzzy one[9]. In this case, it proofs more useful to connect the DPD model directly with the *continuum* Navier-Stokes equations, which are eventually linked to the microscopic dynamics through a well-know procedure[7]. Let us discuss these two cases separately in the following sections.

3 DPD for Unbounded Atoms: The Simulation of Simple Fluids

Originally, the DPD model was introduced to model complex fluids made of structures floating in a simple liquid solvent. The structures are implemented through springs and/or repulsive potentials between certain bounded particles. The main problem is, then, how to specify the particular form of the potential function and friction forces between the dissipative particles that model the *fluid solvent*.

In the original model, the conservative potential V and force $\mathbf{F}_i$ were assumed to be pair-wise and of the form

$$V = \frac{1}{2}\sum_{ij} a_{ij}(1 - r_{ij}/r_c)^2$$
$$\mathbf{F}_i = -\frac{\partial V}{\partial \mathbf{r}_i} = \sum_j F^C(r_{ij})\mathbf{e}_{ij} \tag{3}$$

with $F^C(r_{ij}) = a_{ij}(1-r_{ij}/r_c)$ where a_{ij} is a particle interaction constant and r_c is a cutoff radius. This force produces a repulsion that could be thought of as representing the "pressure forces" exerted between portions of a fluid. Without any other guidance, the weight function $\omega^D(r)$ in the dissipative and random forces is given the same linear functional form. The resulting set of dissipative particles display hydrodynamic behaviour, because momentum is conserved. In fact, DPD has been used for the study of the hydrodynamics of simple fluids in several works[10–13].

However, the fact that DPD conserves momentum does not mean that it makes a good model for solving hydrodynamics. Indeed, molecular dynamics itself is also momentum conserving and displays hydrodynamic behaviour, but it is not the method of choice for solving hydrodynamic problems (although it can be used and has been used for that[14]). We will argue that for the efficient simulation of hydrodynamics the original DPD model is not well suited. Instead, the model can be improved in several directions in order to overcome its limitations when modelling simple fluids. Let us review now these limitations.

3.1 Limitations of DPD for Modelling Simple Fluids

- **Equation of state:** The pressure equation of state is an outcome of the simulation, not an input. The linear conservative forces of the original DPD model produce an equation of state that is quadratic in the density[15]. The thermodynamics of the model cannot be changed at will. One would like to be able to specify, through the functional form of the conservative force, the desired thermodynamic behaviour.

- **Simplistic friction forces.** According to the DPD friction force, if a dissipative particle is orbiting in a circumference around a reference particle, it will not exert any force on this particle. Nevertheless, on simple physical grounds one expect that the motion of the dissipative particle must drag in some way the reference particle. Of course, if many DPD particles are involved simultaneously in between the two particles, this will result in an effective drag. The same is true for a purely conservative molecular dynamics simulation. It would be nice, though, to have this effect captured directly in terms of modified friction forces in a way that a smaller number of particles need to be used to reproduce large scale hydrodynamics.

- **Viscosity as output, not input.** Even though the macroscopic behaviour of the model is hydrodynamic[5], it is not possible to relate in a simple direct way the viscosity of the fluid with the model parameters. Only after a recourse to the methods of kinetic theory can one estimate what input values for the friction coefficient should be imposed to obtain a given viscosity[6,16,17].

- **Scale of DPD.** It is difficult in advance to specify the scale at which a DPD simulation is operating. In particular, there is no parameter in the model that sets the physical scale of the particle. The cutoff radius r_c simply sets the number of neighbours, and the distance between particles (or the total number density in the container) could be in principle changed at will and there is no prescription about what is an appropriate number. There are many attempts to restore a *scale free* property for DPD[18–20]. This property refers to the ability of the simulation method to get *convergent* results as the number of particles increases, this is, up to a certain number of particles, having more particles should not change the results. This is obviously connected to the idea that the DPD model needs to incorporate the notion of *resolution* and that finer resolutions lead to converged results. To get this property, the parameters in the model need to depend on the level of coarse-graining, but this is not specified in the original model.

- **Importance of thermal fluctuations.** The problem of the scale of a dissipative particle is closely related to the fact that DPD cannot switch off thermal fluctuations according to the size of a DPD particle. On general statistical mechanics grounds, thermal fluctuations should scale as $1/\sqrt{N}$ where N is the number of coarse-grained degrees of freedom. "Larger" dissipative particles should display smaller fluctuations. But there is no size associated to a dissipative particle. This problem is crucial, for instance, in the case of suspended colloidal particles or in microfluidics applications where the physical dimensions of the suspended objects or physical dimensions of the operating device determine whether and, more importantly, *to which extent* thermal fluctuations come into play.

- **Isothermal model.** The DPD model is isothermal and cannot sustain realistic energy transport.

During the years, several DPD-like models have been introduced in order to deal with these limitation. In the following subsections we briefly review these DPD models.

3.2 Many-Body Dissipative Particle Dynamics: MDPD

Pagonabarraga and Frenkel presented a model in which the conservative forces where derived from a many-body potential that derives from a free energy density[21]. The model has been studied in detail by Trofimov et al.[22] presenting a multicomponent version, whereas liquid-vapor coexistence and drop dynamics has been considered by Warren[23] and surface tension of the model has been studied in Refs. 24, 25. The essential feature in MDPD is the incorporation of a density variable associated to every particle

$$d_i = \sum_j^N W(r_{ij}) \tag{4}$$

where $W(r)$ is a suitable weight function normalized to unity $\int d\mathbf{r} W(\mathbf{r}) = 1$ and with a bell shape. If around particle i there are many particles j, the density d_i defined above will be a large number.

The total potential energy of interaction between dissipative particles is made to depend on this density.

$$V(\mathbf{r}_1, \cdots, \mathbf{r}_N) = \sum_i \psi(d_i) \tag{5}$$

The resulting potential is, therefore, many-body although the forces are still pair-wise in form and easy to implement in a code, this is

$$\mathbf{F}_i = -\frac{\partial V}{\partial \mathbf{r}_i} = -\sum_i [\psi'(d_i) + \psi'(d_j)] W'(r_{ij}) \mathbf{e}_{ij} \tag{6}$$

The connection between the thermodynamics of the system and a density dependent potential has been worked out in detail in Ref. 26. Roughly speaking the pressure of the homogeneous system is given by $P = k_B T d + d^2 \psi'(d)$, which allows one to interpret the function $\psi(d)$ in Eq. 5 as the excess free energy of the system. In this way, it is possible to *introduce* the global thermodynamics of the system through the particular functional form of the many-body potential.

3.3 Fluid Particle Model

The realization that a circularly orbiting dissipative particle in the original DPD model does not produce any friction force on a particle located at the origin lead us to introduce the Fluid Particle Model with a general form of the friction forces that have a shear component in addition to the usual central friction forces of DPD[27]. Of course the friction forces are no longer central and angular momentum is not conserved. This is remedied in the FPM model by introducing a spin variable that accounts for the missing angular momentum. The spin is regarded as the angular momentum of the atoms within a fluid particle with

respect to the center of mass of the fluid particle. The kinetic theory of the model was presented in Ref. 18 and it was shown that on macroscopic scales the spin variable becomes slaved by the vorticity of the fluid. Shearing forces between dissipative particles have been introduced for the simulation of colloids[28,29]. The shear friction was introduced in FPM as a theoretically appealing feature, but it has been given further support from a top-down approach in which the viscous term of the Navier-Stokes equation is discretized according to the Smoothed Particle Hydrodynamics methodology[30]. The resulting friction forces display shear components in a natural way. We will review this approach below.

3.4 EDPD

The original DPD model is isothermal and cannot sustain temperature gradients. For this reason it does not give the correct transport of energy across the system. This can be remedied by introducing an internal energy variable associated to every particle, along with a temperature variable. The kinetic energy that is lost due to friction forces is invested into increasing the internal energy of the particles. In addition, a thermal conduction term ensures thermal equilibration between dissipative particles in that it drives the temperatures of the particles towards a common value. This energy conserving model was named EDPD and was introduced independently in Refs. 31, 32 and further studied in Refs. 33, 34, with several recent applications in Refs. 35, 36.

3.5 A Comprehensive Model: SDPD

The three models above (MDPD, FPM, EDPD) try to solve some of the quoted problems of DPD, but none of them solves all the problems simultaneously. In 2003 we introduced the Smoothed Dissipative Particle Dynamics (SDPD) model, which has many features in common with the above three models, but does not suffer from the problems and limitations of DPD quoted above[30]. In fact, the model is just a version of the well-known method of Smoothed Particle Hydrodynamics (SPH)[37], with thermal fluctuations included in a thermodynamically consistent way[38]. SPH is a Lagrangian mesh-less discretization of the Navier-Stokes equations which was introduced by Lucy[39] and Monaghan[40] in the 70's in order to solve hydrodynamic problems in astrophysical contexts. Generalizations of SPH that include viscosity and thermal conduction and address laboratory scale situations like viscous flow and thermal convection were presented much later[41–43]. SPH is now used in a number of applications, particularly because of the easy treatment of free boundaries[44]. The SDPD method of Ref. 30 has a structure very similar to both DPD and SPH and extracts the best of both models (fluctuation from DPD, connection to Navier-Stokes from SPH). The computational simplicity of SDPD is comparable to that of DPD.

The essential idea of SDPD is that the dissipative particles ought to be regarded as truly thermodynamic subsystems of the whole system moving with the flow. In those cases, we prefer to name the particle as a *fluid particle*. In addition to the position and velocity a fluid particle has a volume (given as the inverse of the density (4)), an internal energy, and entropy. The volume is a function of the positions and the internal energy is a function of the entropy and volume of the particles. Therefore, the independent variables characterizing the state of the fluid particles is $x = \{\mathbf{r}_i, \mathbf{p}_i, S_i\}$ where $\mathbf{r}_i$ is the position of the fluid particle, $\mathbf{p}_i$ its momentum, and S_i its entropy. We could equally select as

independent variable the internal energy, or the temperature, instead of the entropy. The total energy and entropy of the system are

$$E(x) = \sum_i \left[\frac{p_i^2}{2m_i} + U(S_i, \mathcal{V}_i) \right]$$
$$S(x) = \sum_i S_i \tag{7}$$

Here $\mathcal{V}_i$ is the volume of the fluid particle which is defined as the inverse of the density, $\mathcal{V}_i = d_i^{-1}$, where the density is defined as a function of the position of the neighbouring particles in Eq. 4. Usually the bell-shaped weight function $W(r)$ has compact support, a sphere of radius h. Finally, $U(S_i, \mathcal{V}_i)$ is the equilibrium relation giving the internal energy of particle i as a function of the mass, entropy and volume of the fluid particle. In this way, we are assuming the principle of local equilibrium, that states that the local thermodynamic behaviour is identical to the global thermodynamic one. The energy function $E(x)$ may be interpreted as a coarse-grained Hamiltonian, with a many-body potential U that depends not-only on the position of the particles but also on the entropy (or the local temperature) variable. Through the functional form of the internal energy of a fluid particle, the global thermodynamic behaviour of the system is fixed, and in this way the equation of state is an input of the model.

The idea underlying the formulation of the SDPD model is that any formulation of a fluid particle should produce a set of interactions between fluid particles that are reminiscent of hydrodynamics. One can construct a model of fluid particles by discretizing the equations of hydrodynamics on a set of nodes that follows the flow field. These nodes can be interpreted as fluid particles with definite amounts of mass, momentum, energy, volume, and entropy. The discretization procedure establishes how the extensive quantities between fluid particles are exchanged and how the fluid particles should eventually move. The discretization of the second derivatives terms appearing in the Navier-Stokes equations is done with the help of the weight function $W(r)$. It can be shown that an approximation to order h^2 for a second space derivative is given in terms of the values of the function in neighbour points by[30]

$$\nabla^\alpha \nabla^\beta A(\mathbf{r}_i) = \sum_j \frac{1}{d_j} F(r_{ij})(A_i - A_j) \left[5\mathbf{e}_{ij}^\alpha \mathbf{e}_{ij}^\beta - \delta^{\alpha\beta} \right] + \mathcal{O}(\nabla^4 A h^2) \tag{8}$$

where $\mathbf{e}_{ij} = \frac{(\mathbf{r}_j - \mathbf{r}_i)}{r_{ij}}$ and $A(\mathbf{r})$ is an arbitrary hydrodynamic field, and $A_i = A(\mathbf{r}_i)$. Expression (8) allows one to estimate the value of the second derivatives at a given point in terms of the value of the function at neighbouring points. The function $F(r)$ is defined through

$$\nabla W(r) = -\mathbf{r} F(r) \tag{9}$$

A common selection for $W(r)$ is the Lucy function,

$$W(r) = \frac{105}{16\pi h^3} \left(1 + 3\frac{r}{h} \right) \left(1 - \frac{r}{h} \right)^3 \tag{10}$$

from which the function $F(r)$ follows

$$F(r) = \frac{315}{4\pi h^5} \left(1 - \frac{r}{h} \right)^2, \qquad F(r) \geq 0 \tag{11}$$

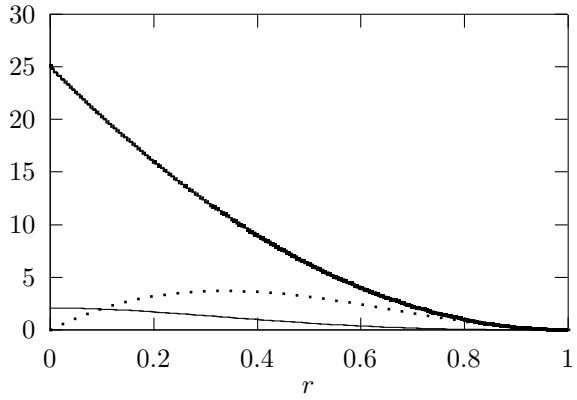


Figure 1. The functions $W(r)$ (solid line), $F(r)$ (bold line), and $rF(r)$ (dotted line).

In Fig. 1 we plot the functions $W(r), F(r)$, and $rF(r)$. However, one should be aware of other kernels $W(r)$ that may perform better[37]. With the use of Eq. 8 we can discretize the second order derivatives of the hydrodynamic equations, which correspond to the irreversible part of the dynamics. The equations also contain first order derivatives (∇P and $\nabla\cdot\mathbf{v}$), which correspond to the reversible part of the dynamics. These first order derivatives are approximated by resorting to the definition of the density, Eq. 4, and to the conservation of energy. The final discrete equations of hydrodynamics are (for simplicity we assume zero bulk viscosity)[30]

$$\begin{aligned}
\dot{\mathbf{r}}_i &= \mathbf{v}_i \\
m\dot{\mathbf{v}}_i &= \sum_j \left[\frac{P_i}{d_i^2} + \frac{P_j}{d_j^2}\right] F_{ij}\mathbf{r}_{ij} - \frac{5\eta}{3}\sum_j \frac{F_{ij}}{d_i d_j}\left(\mathbf{v}_{ij} + \mathbf{e}_{ij}\mathbf{e}_{ij}\cdot\mathbf{v}_{ij}\right) \\
T_i\dot{S}_i &= -2\kappa\sum_j \frac{F_{ij}}{d_i d_j}T_{ij} + \frac{5\eta}{6}\sum_j \frac{F_{ij}}{d_i d_j}\left(\mathbf{v}_{ij}^2 + (\mathbf{e}_{ij}\cdot\mathbf{v}_{ij})^2\right)
\end{aligned} \tag{12}$$

Here, P_i, T_i are the pressure and temperature of the fluid particle i, which are functions of d_i, S_i through the equilibrium equations of state (easily derived from $U(S_i, \mathcal{V}_i)$). In addition, $\mathbf{v}_{ij} = \mathbf{v}_i - \mathbf{v}_j$, and $T_{ij} = T_i - T_j$. It is easily shown that the above model conserves mass, momentum and energy and that the total entropy is a non-decreasing function of time. As the number of particles increases, the resulting flow converges towards the solution of the Navier-Stokes equations, by construction.

The physical picture that emerges from these equations is very appealing and closely resembles the interpretation of dissipative particles in DPD. Particles of constant mass m move according to their velocities and exert forces of a range h to each other of different nature. First, a repulsive force directed along the line joining the particles that has a magnitude given by the pressure and densities of the particles. Roughly speaking, the larger is the pressure in a given region, the higher the repulsion between them. The fluid particles are also subject to friction forces that depend on the relative velocities of the particles. As

opposed to the friction force of the DPD model, there is a component of this forces, directly proportional to $\mathbf{v}_{ij}$ that breaks the conservation of total angular momentum. If one wishes to respect this conservation law, then it is necessary to introduce in the model a spin variable associated to every particle[18]. For a sufficiently large number of particles, the violation of angular momentum is negligible[18]. The terms in the entropy equation represent heat conduction and viscous heating. The heat conduction term tries to reduce temperature differences between particles by suitable energy exchange[43], whereas the viscous heating term ensures that the kinetic energy dissipated by the friction forces is transformed into internal energy of the fluid particles.

The above model can be regarded as one more version of SPH, of which there are many[41–43,45]. However, to our knowledge this is the first model that strictly respects the Second Law. In fact, it is possible to cast the above model into the the GENERIC framework[30,38], which is a thermodynamically consistent and a rather universal framework for non-equilibrium dynamics. As it is apparent from the GENERIC framework, the Second Law is expressed through a dissipative matrix which is positive definite. The dissipative matrix actually governs the amplitude of thermal fluctuations through the Fluctuation-Dissipation theorem. In this sense, it is not possible to have properly defined thermal fluctuations in a model if the Second Law is not respected exactly. Because the set of deterministic equations (12) can be cast in the GENERIC form[30], the introduction of thermal noise into Eqs. 12 is reasonably simple. The stochastic version of Eqs. 12 is presented below. One introduces a stochastic term $md\tilde{\mathbf{v}}_i$ in the momentum equation and a stochastic term $T_i d\tilde{S}_i$ in the entropy equation. They are given by

$$\begin{aligned} md\tilde{\mathbf{v}}_i &= \sum_j A_{ij} d\mathbf{W}_{ij} \cdot \mathbf{e}_{ij} \\ T_i d\tilde{S}_i &= -\frac{1}{2} \sum_j A_{ij} d\mathbf{W}_{ij} : \mathbf{e}_{ij}\mathbf{v}_{ij} + \sum_j C_{ij} dV_{ij} \end{aligned} \tag{13}$$

We have introduced, for each pair i, j of particles, a matrix of independent increments of the Wiener process $d\mathbf{W}_{ij}$. In Eq. 13 we have also introduced an independent increment of the Wiener process for each pair of particles, dV_{ij}. This term will give rise to the heat conduction terms. Finally, the functions A_{ij}, C_{ij} might depend on the state of the system through the positions and entropy of the particles. We postulate the following symmetry properties, that will ensure momentum and energy conservation

$$\begin{aligned} d\mathbf{W}_{ij} &= d\mathbf{W}_{ji} \\ dV_{ij} &= -dV_{ji} \\ A_{ij} &= A_{ji} \\ C_{ij} &= C_{ji} \end{aligned} \tag{14}$$

The independent increments of the Wiener processes satisfy the following Itô mnemotechnical rules

$$\begin{aligned} d\mathbf{W}_{ii'}^{\alpha\alpha'} d\mathbf{W}_{jj'}^{\beta\beta'} &= [\delta_{ij}\delta_{i'j'} + \delta_{ij'}\delta_{i'j}]\delta^{\alpha\beta}\delta^{\alpha'\beta'} dt \\ dV_{ii'} dV_{jj'} &= [\delta_{ij}\delta_{i'j'} - \delta_{ij'}\delta_{i'j}] dt \\ d\mathbf{W}_{ii'}^{\alpha\alpha'} dV_{ii'} &= 0 \end{aligned} \tag{15}$$

which respect the symmetries (14) under particle interchange. As a convention, superindices refer to tensorial components while subindices label different particles. The noise amplitudes A_{ij}, C_{ij} are fixed by the Fluctuation-Dissipation theorem[30]

$$A_{ij} = \left[8k_B \frac{T_i T_j}{T_i + T_j} \frac{5\eta}{3} \frac{F_{ij}}{d_i d_j} \right]^{1/2}$$
$$C_{ij} = \left[4\kappa k_B T_i T_j \frac{F_{ij}}{d_i d_j} \right]^{1/2} \tag{16}$$

The final stochastic equations for the fluid particle model are given by[30]

$$\begin{aligned} d\mathbf{r}_i &= \mathbf{v}_i dt \\ m d\mathbf{v}_i &= \sum_j \left[\frac{P_i}{d_i^2} + \frac{P_j}{d_j^2} \right] F_{ij} \mathbf{r}_{ij} dt - \frac{5\eta}{3} \sum_j \frac{F_{ij}}{d_i d_j} \left(\mathbf{v}_{ij} + \mathbf{e}_{ij} \mathbf{e}_{ij} \cdot \mathbf{v}_{ij} \right) dt + m d\tilde{\mathbf{v}}_i \\ T_i dS_i &= \frac{5\eta}{6} \sum_j \frac{F_{ij}}{d_i d_j} \left(\mathbf{v}_{ij}^2 + (\mathbf{e}_{ij} \cdot \mathbf{v}_{ij})^2 \right) dt - 2\kappa \sum_j \frac{F_{ij}}{d_i d_j} T_{ij} dt + T_i d\tilde{S}_i \end{aligned} \tag{17}$$

where we have neglected, for the sake of the presentation, small terms of the order of k_B/C_i, where C_i is the heat capacity at constant volume of particle i. These terms are important if one wishes to obtain exact energy conservation[30].

We understand the SDPD model in Eqs. 17 as the "proper" DPD model valid for the simulation of Newtonian fluids at mesoscopic scales when thermal fluctuations are important. Microfluidics, colloidal suspensions and dilute polymeric solutions, for which a clear *Newtonian solvent* exists, benefit from the SDPD formulation. The model in Eqs. 17 has a similar simplicity as the original DPD model but with a sounded physical meaning. It should be regarded as an SPH model with thermal fluctuations included in a consistent way. The model solves *all* the conceptual problems of DPD mentioned in Sec. 3.1. In particular, the pressure and any other thermodynamic information is introduced as an input as in the MDPD model. The conservative forces of the original model become physically sounded pressure forces. Energy is conserved and we can study transport of energy in the system as in EDPD. The Second Law is satisfied. The transport coefficients are input of the model. The range functions of DPD have now very specific forms (see Fig. 1), and one can use the large body of knowledge generated in the SPH community to improve on the more adequate shape for the weight function $W(r)$[37]. The particles have a physical size given by its physical volume and it is possible to specify the physical scale being simulated. In addition, the *deterministic* model is *scale free*, in the sense that by increasing the number of particles above certain number does not change the results appreciably. One should understand the density number of particles as a way of controlling the *resolution* of the simulation. The amplitude of thermal fluctuations, however, scales with the size of the fluid particles: *large* fluid particles display smaller thermal fluctuations, in accordance with the usual notions of equilibrium statistical mechanics. While the fluctuations scale with the size of the fluid particles, the resultant stochastic forces on *suspended* bodies are independent of the size of the fluid particles and only depend on the overall size of the object[46]. In this way, the same stochastic model allows to simulate a colloidal particle (that will display Brownian diffusive behaviour due to the fluctuations of the solvent) or a ball in

a quiescent swimming pool (that does not practically diffuse). This property of automatic switching off thermal fluctuations with the scale of the problem is completely absent in the original DPD model.

3.6 Internal Variables

We have seen that the SDPD model is obtained from the discretization of the continuum Navier-Stokes equations, recast in the thermodynamically consistent framework of GENERIC, that provides for a straightforward introduction of thermal fluctuations in a respectful way. Of course, nothing refrains to use more complex continuum equations that are traditionally used for the description of complex fluids. Usually, this requires to introduce additional structural or order parameter variables associated to each fluid particle. In general, the continuum models of the GENERIC type for complex fluids typically involve *additional structural or internal variables* that are coupled with the conventional hydrodynamic variables. The coupling renders the behaviour of the fluid non-Newtonian and complex. For example, polymer melts are characterized by additional conformation tensors, colloidal suspensions can be described by further concentration fields, mixtures are characterized by several density fields (one for each chemical specie), emulsions are described with the amount and orientation of interface, etc. All these continuum models rely on the hypothesis of local equilibrium and, therefore, the fluid particles are regarded as thermodynamic subsystems. The physical picture that emerges from these fluid particles is that they represent "large" portions of the fluid and therefore, the scale of these fluid particles is *supramolecular*. This allows one to study larger time scales than the less coarse models where the mesostructures are represented explicitly through additional interactions between particles (i.e. necklaces for representing polymers, spherical solid particles to represent colloid, different types of particles to represent mixtures) . The price, of course, is the need for a deep understanding of the physics at this more coarse-grained level, which appears in the form of entropy and energy functionals depending on internal variables and kinematic and dissipative matrices describing the complex coupling of the internal microstructure and flow.

For example, in order to describe polymer solutions, we may take a level of CG in which every fluid particle contains already many polymer molecules. This is a more coarse-grained model than describing viscoelasticity by joining dissipative particles with springs[47]. The state of the polymer molecules within a fluid particle may be described either with the average end-to-end vector of the molecules[48,49], or with a conformation tensor[50]. In this latter case, the continuum limit of the model leads to the Olroyd-B model of polymer rheology. Another example where the strategy of internal variables is successful is in the simulation of mixtures. Instead of modelling a mixture with two types of dissipative particles as it is usually done in DPD, one may take a thermodynamically consistent view in which each fluid particle contains the concentration of one of the species, for examples see Refs. 51,52. These two examples show how one can address viscoelastic flow problems and mixtures with a simple methodology that involves fluid particles with internal variables. The idea can, of course, be applied to other complex fluids where the continuum equations are known.

4 Two Technical Points: Integrators and Boundary Conditions

4.1 Integrators

The DPD equations are a set of stochastic differential equations that, as such, require careful consideration[53]. Because the conservative part of the dynamics has a Hamiltonian structure, it is natural to look for generalizations of the usual MD integrators, which are symplectic and, therefore, have good energy conserving properties at large simulation times. However, the fact that the dissipative friction is velocity dependent implies that the usual Verlet algorithm needs to be reconsidered. Groot and Warren introduced a modification of the Verlet algorithm for DPD[15] that uses an intermediate velocity predictor in order to deal with the velocity dependent dissipative forces, and Pagonabarraga and Frenkel introduced the so called self-consistent integrator which is an implicit method that needs to be solved iteratively, but has the advantage of being time reversible[21]. These integrators and some variants were compared in Ref. 54 with the conclusion that the self-consistent integrator produces much reduced numerical artifacts on the observables considered (temperature, radial distribution function, and velocity correlations). Later, Shardlow introduced a splitting method that treats the dissipative forces in an implicit way, but in a pair-wise fashion. This makes the method more efficient than the self-consistent method of Ref. 55. Further comparisons between methods[56] showed that the Sharlow integrator stands among the best integrators for DPD. The major advantage of implicit methods relies on the fact that the method is stable even in the large friction regime were the equations become stiff. Another method to deal with the time-step problems of naive integrators for DPD is the Lowe-Andersen thermostat[57]. Any dynamics having (2) as its equilibrium distribution is named a *thermostat*, and we may speak of Eqs. 1 as the DPD thermostat when we use DPD just to sample the equilibrium ensemble (2). The Lowe-Andersen thermostat has been slightly modified by Peters[58] in order to show that these thermostats are, essentially, *implicit* integrators of the original DPD equations, explaining why these methods may deal with large integration steps and still recover faithfully the equilibrium ensemble (2). A number of momentum conserving thermostats exists now[57–62]. Methods similar to the splitting method of Sharlow have been considered in Ref. 63 with an splitting with an iterative procedure, while the Trotter expansion used in the design of symplectic integrators is pursued in Refs. 64–66. The value of these latter integrators is that they may naturally generalize towards the formulation of the SDE emerging in models that, in addition to position and momenta, include extra variables (as is the case for SDPD that includes a thermal variable, or FPM that includes a spin variable). These quality integrators for the extended DPD models need yet to be fully investigated.

4.2 Boundary Conditions

The full statement of a flow problem requires the specification of boundary conditions. When DPD is used to model fluid flow, one needs to pay attention to this issue. In DPD the boundary conditions are expressed in terms of external forces to the DPD that try to mimic the effect of a wall in a liquid. Usually, solid walls are represented by "frozen" dissipative particles, an approach already used in the first application of DPD[67]. The consideration of the no-slip boundary condition at a wall was considered for the first time in Ref. 68, where an effective force was analytically computed by taking the continuum limit of a particulate

solid wall made of frozen dissipative particles. The method is not easily generalized to non-planar walls, though, and for this reason keeping the solid walls made of frozen particles still is the method of choice. The particulate nature of DPD usually leads to the creation of inhomogeneous density profiles near walls, in a similar way as a molecular fluid structures itself near hard walls. This is regarded as an artifact because macroscopic measurements of the fluid viscosity may be affected by this layering. Consequently, remedies have been devised. In Ref. 69 the authors propose an iterative method for specifying the density near the wall, by adjusting a normal force on the particles near the walls. Adhesive walls for the study of wetting have been constructed in the MDPD model[70] by freezing a liquid structured region to form the solid wall. The inhomogeneous structure of the wall together with the interaction forces proposed reduce the amount of layering near the wall.

The boundary conditions on the surfaces of colloidal particles has been treated with both methods, a continuum friction force[29] and through the frozen particle method[46]. In this latter case, a convenient way to impose no-slip boundary conditions is through the assignment of fictitious velocities inside the frozen wall particles that ensure the correct interpolation at the surface of the colloid, a method introduced in Ref. 71 in the SPH context.

5 Microscopic Foundation of DPD

5.1 DPD for Unbounded Atoms

The SDPD mode, which in our view is the appropriate DPD model for modelling Newtonian fluid flow, has been derived from a top-down approach by discretizing the Navier-Stokes equations and ensuring thermodynamic consistency of the resulting discrete equations. The Navier-Stokes equations are already a CG model in which the atoms are eliminated in favor of mass, momentum, and energy density fields. The derivation of the equations of hydrodynamics from the underlying Hamiltonian dynamics of the atoms is a well studied problem that dates back to Boltzmann and the origins of kinetic theory. It still deserves attention in that *discrete* versions of hydrodynamics, which is what we need in order to simulate hydrodynamics, have been only recently obtained from molecular considerations[72–74]. These latter works show how an *Eulerian* description of hydrodynamics can be derived from the Hamiltonian dynamics of the underlying atoms, by defining mass, momentum, and energy of cells which surround certain points fixed in space. However, *Lagrangian* descriptions in which the cells move, are much more tricky to deal with. From a continuum point of view, the rate of change of an infinitesimal volume moving with the flow field satisfies the following equation $\dot{\mathcal{V}} = \mathcal{V}\nabla \cdot \mathbf{v}$. If we define the density field as $\rho = m/\mathcal{V}$, the continuity equation simply tells us that the mass of a fluid particle is constant[75] – a strong argument to use in favor of a fixed mass of the fluid particles in the discrete SDPD model. However, any definition of a fluid particle in terms of a moving region of space should have molecules entering or leaving the moving cells, something that seems to be against the idea of constant mass fluid particles. Work remains to be done to define the CG variables of a model for lagrangian fluid particles that is fully satisfactory.

Perhaps the earliest attempt to derive effective forces between CG particles representing a fluid from MD simulations was given in Ref. 76 where fluid particles were constructed from a Voronoi tessellation whose centers were moving according to the underlying MD.

An effective excluded volume potential was obtained from the radial distribution function of the Voronoi centers. Using a similar idea the BLOBS method has been introduced in Ref. 77 where an initial single blob moves according to the underlying dynamics and the information about their dynamical correlations is compiled. Subsequently a system of N blobs is constructed in order to reproduce the above correlations. Recently, another attempt to obtain DPD from the underlying MD has been undertaken in Ref. 78 by using the rigorous approach of the theory of coarse-graining. However, in order to construct the "fluid particles" these authors constraint a collection of Lennard-Jones to move bounded maintaining a specified radius of gyration. The fluid no longer is a simple atomic fluid but rather a fluid made of complex "molecules" (the atomic clusters constrained to have a radius of gyration) whose rheology is necessarily complex.

Our impression is that we still have not solved satisfactorily the problem of deriving from the microscopic dynamics the dynamics of CG particles that capture the behaviour of a simple fluid made of *unbounded* atoms. The best model from a conceptual point of view, up to now, is the SDPD model that discretize the Navier-Stokes equations (which do have been derived from MD) as its starting point.

5.2 DPD for Bounded Atoms

The situation is much more satisfactory when considering the CG dynamics of clusters of atoms that are bounded together. In that case, the theory of coarse-graining[7] (this is, the Mori-Zwanzig formalism) does allows one to derive the equations of DPD from the underlying molecular dynamics, by just simply considering a DPD particle as the center of mass of the bounded atoms.

The first attempt to derive the DPD equations of motion from the underlying Hamiltonian dynamics was given in Ref. 79 for a very simple model of harmonic 1D lattice. While the general idea was reflected there, the equations were derived for a model of a solid, for which a number of issues concerning the non-Markovian nature of the description arise[80–82]. Mori-Zwanzig theory for deriving the effective dynamics of clusters of bonded atoms has been used recently in order to derive the equations of DPD[83,84]. The idea is to group several atoms of parts of a molecule (or a whole molecule itself) into M clusters, labeled with Greek indices. The μ-th cluster is made of N_μ atoms whose positions and momenta are $\mathbf{r}_{i_\mu}$, $\mathbf{p}_{i_\mu}$ where the index i_μ runs from $1, \cdots, N_\mu$, while the index μ runs from $1, \cdots, M$. The Hamiltonian of the system is

$$H(z) = \sum_{\mu=1}^{M} \sum_{i_\mu=1}^{N_\mu} \frac{\mathbf{p}_{i_\mu}^2}{2m_{i_\mu}} + \phi \tag{18}$$

where m_{i_μ} is the mass of the atom i_μ and $\phi(q)$ is the potential energy.

At a coarse-grained level, we represent each cluster with just the position $\mathbf{R}_\mu$ and momentum $\mathbf{P}_\mu$ of its center of mass. These relevant variables are the following functions

of the atomic variables

$$\mathbf{R}_\mu(z) = \frac{1}{M_\mu}\sum_{i_\mu=1}^{N_\mu} m_{i_\mu}\mathbf{r}_{i_\mu}$$
$$\mathbf{P}_\mu(z) = \sum_{i_\mu=1}^{N_\mu} \mathbf{p}_{i_\mu} \tag{19}$$

where $M_\mu = \sum_{i_\mu=1}^{N_\mu} m_{i_\mu}$ is the total mass of the molecule μ. Once the CG variables are selected, the Mori-Zwanzig formalism allows to obtain the dynamics of the CG variables. The resulting closed set of stochastic differential equations for the evolution of the position and momentum of the center of mass of cluster μ is given by[84]

$$\frac{d\mathbf{R}_\mu}{dt} = \mathbf{V}_\mu$$
$$\frac{d\mathbf{P}_\mu}{dt} = \langle \mathbf{F}_\mu \rangle^{\mathbf{R}} + \sum_\nu \boldsymbol{\gamma}_{\mu\nu}(\mathbf{R},\mathbf{P})\mathbf{V}_{\mu\nu} + k_B T \sum_\nu \frac{\partial \boldsymbol{\gamma}_{\mu\nu}}{\partial \mathbf{P}_\nu}(\mathbf{R},\mathbf{P}) + \tilde{\mathbf{F}}_\mu \tag{20}$$

Here we use the shorthand notations $\mathbf{R} = (\mathbf{R}_1,\ldots,\mathbf{R}_M)$, $\mathbf{P} = (\mathbf{P}_1,\ldots,\mathbf{P}_M)$ and we denote by $\langle\cdot\rangle^{\mathbf{R},\mathbf{P}}$ the equilibrium expectation conditional to fixed $(\mathbf{R},\mathbf{P})$. The conditional expectation force on cluster μ can be written as

$$\langle \mathbf{F}_\mu \rangle^{\mathbf{R}} = -\frac{\partial V}{\partial \mathbf{R}_\mu}(\mathbf{R}) \tag{21}$$

where $V(\mathbf{R})$ is the so called potential of mean force which is defined by

$$V(\mathbf{R}) \equiv -k_B T \ln \int dz \frac{1}{\mathcal{Q}} \exp\{\beta\phi(z)\} \prod_\mu \delta(\mathbf{R}_\mu(z) - \mathbf{R}_\mu) \tag{22}$$

By definition, the potential of mean force is actually a coarse-grained free energy.

The friction tensor is given by a Green-Kubo expression as

$$\boldsymbol{\gamma}_{\mu\nu}(\mathbf{R},\mathbf{P}) = \frac{1}{k_B T}\int_0^\infty dt \langle \delta\mathbf{F}_\mu \exp\{t\mathcal{R}\}\delta\mathbf{F}_\nu\rangle^{\mathbf{R},\mathbf{P}} \tag{23}$$

where $\delta\mathbf{F}_\mu = \mathbf{F}_\mu - \langle\mathbf{F}_\mu\rangle$ and $\mathbf{F}_\mu$ is the total force acting on the molecule μ:

$$\mathbf{F}_\mu = \sum_\nu \mathbf{F}_{\mu\nu} \equiv \sum_\nu \sum_{i_\mu j_\nu} \mathbf{F}_{i_\mu j_\nu} \tag{24}$$

Here $\mathbf{F}_{i_\mu j_\nu}$ is the force that atom j_ν exerts on atom i_ν, and $\mathbf{F}_{\mu\nu}$ is the total force that molecule ν exerts on molecule μ. The evolution operator $\exp\{t\mathcal{R}\}$ describes a Hamiltonian dynamics constrained to produce the specified values of $\mathbf{R},\mathbf{P}$[84].

Finally, the stochastic force $\tilde{\mathbf{F}}_\mu$ is given in terms of a linear combination of Wiener processes as, for example, $\mathbf{F}_\mu = \sum_\alpha B_{\mu\alpha} dW_\alpha(t)/dt$ with

$$\sum_\alpha B_{\mu\alpha}B_{\nu\alpha} = 2k_B T \boldsymbol{\gamma}_{\mu\nu} \tag{25}$$

This is the Fluctuation-Dissipation theorem for this problem. For a general form of the friction tensor, in order to obtain of $B_{\mu\alpha}$ one would need to perform a Cholesky decomposition of the friction tensor that may be time consuming (although the matrix problem, with a linked list for the particles, is sparse).

The only non-trivial assumption that is required for the validity of Eqs. 20 is that the time scale of evolution of the CG variables is much larger than the time scale of evolution of its derivatives. This allows for a Markovian assumption and allows to interpret the above equations as bona fide Ito stochastic differential equations. The structure of Eq. 20 is very similar to the structure of DPD, in that conservative, friction forces depending on relative velocities, and stochastic forces appear. However, several differences should be noted. The potential of mean force $V(\mathbf{R})$ is given by the explicit microscopic expression (22), which is the usual definition for this quantity as the CG free energy. The scale and shape of the CG potential is dictated by the level of coarse-graining selected (how many atoms constitute the CG particle). The parabolic profile typical of DPD should be regarded as a very crude approximation for the actual CG potential. Another big difference between the usual DPD equations and Eqs. 20 is that the friction coefficient $\boldsymbol{\gamma}_{\mu\nu}(\mathbf{R},\mathbf{P})$ is, in fact, a friction tensor that depends on the positions and momenta of *all* the molecules in the system and not only on the distance $|\mathbf{R}_\mu - \mathbf{R}_\nu|$ of the pair as in DPD. Obviously, several approximations are required in order to find tractable expressions for the friction tensor. A simple one is to assume that the correlation between the forces on molecule μ and ν will depend on the positions of these two molecules but will not depend much on the positions and momenta of the rest of molecules. We thus introduce the following functional ansatz that was first used by Akkermans and Briels[85,84]

$$\boldsymbol{\gamma}_{\mu\nu}(\mathbf{R},\mathbf{P}) \approx -\gamma_\perp(R_{\mu\nu})(\mathbf{1} - \mathbf{e}_{\mu\nu}\mathbf{e}_{\mu\nu}^T) - \gamma_{||}(R_{\mu\nu})\mathbf{e}_{\mu\nu}\mathbf{e}_{\mu\nu}^T. \tag{26}$$

The right-hand side of this equation only depends on $\mathbf{R}_\mu$ and $\mathbf{R}_\nu$ and it is a general form for a tensor that is invariant by rotations along the axis joining the particles μ, ν. Compatibility of (26) with Eq. 23 then requires that

$$\begin{aligned}
\gamma_{||}(R_{\mu\nu}) &= -\frac{1}{k_B T}\int_0^\infty dt \langle (\delta\mathbf{F}_\mu(t)\cdot\mathbf{e}_{\mu\nu})(\delta\mathbf{F}_\nu(0)\cdot\mathbf{e}_{\mu\nu})\rangle^{\mathbf{R}_{\mu\nu}} \\
\gamma_{\perp}(R_{\mu\nu}) &= -\frac{1}{k_B T}\int_0^\infty dt \langle (\delta\mathbf{F}_\mu(t)\cdot\mathbf{e}^\perp_{\mu\nu})(\delta\mathbf{F}_\nu(0)\cdot\mathbf{e}^\perp_{\mu\nu})\rangle^{\mathbf{R}_{\mu\nu}}
\end{aligned} \tag{27}$$

With this approximate model for the friction tensor, the resulting friction forces are identical to the general friction forces of the FPM and SDPD models. For the case of bounded atoms, the explicit Green-Kubo expressions provide one route to the explicit calculation of the friction tensor. For this simple form of the friction tensor, the stochastic forces which satisfy the Fluctuation-Dissipation theorem are given simply by those of the FPM[18].

5.3 Methods to Obtain the Potential of Mean Force

Of course, the formal expression (22) for the potential of mean force does not allow to compute explicitly this potential. Note that the potential of mean force is defined in the high dimensional space of $\mathbf{R}$ and, therefore, a brute force sampling of the probability distribution (the logarithm of the potential) is unfeasible, a phenomenon known as the curse of dimensionality.

There are many different methods for the calculation of approximate versions for the potential of mean force. Most of them try to get a pair-wise potential of mean force, because they are much simpler to simulate at the CG level. Also, most of them formulate a parametrized model for potential of mean force and aim at obtaining the parameters according to several criteria. For example, one can use the Ornstein-Zernike equation that relates the direct correlation function with the radial distribution function and use a closure (Percus-Yevick or Hipernetted Chain) to obtain the potential[86]. Refs. 87, 88 consider an iterative adjustment of potential parameters by running CG simulations with the target potential and updating the parameters in a way to reduce the difference between the target and model radial distribution functions. Also iterative methods like the inverse Monte Carlo technique[89,90] or the Newton inversion method[91], by matching thermodynamic properties[92] or by using directly the underlying all-atom interactions through a force matching procedure have been considered. In the latter case, the potential of mean force is obtained by requiring that the actual forces between CG particles and the forces which are modelled with a parametrized model are as similar as possible[93–96]. Recently, a very elegant procedure has been introduced by Shell to obtain parametrized models of the effective potential by using the concept of relative entropy[97]. While all the above methods may be termed as variational in that a difference between a model and a target is minimized, a non-variational approach is given in Refs. 78, 84, 98 where the idea is to compute the conditional expectation that defines the mean force on the CG particles through a constrained MD where the CG particles remain fixed. Another approach is to consider a many-body potential of the embedded atom form usually considered in MDPD[99].

5.4 Methods to Obtain the Friction Tensor

As compared with the vast literature on the calculation of the potential of mean force between CG particles, the calculation of the friction between CG particles has received much less attention. While the potential of mean force gives all the static properties at the CG level, friction is crucial to obtain the dynamic aspects of the CG procedure[84]. To our knowledge, the first calculation of the friction coefficient between to CG particles was given by Akkermans and Briels in their consideration of a dimer CG version of a linear polymer[98]. The same method of running constrained dynamics simulations that served to obtain the potential of mean force allows to obtain the correlations of the fluctuating forces between the centers of mass of the two beads of the dimer. This provides, from the Green-Kubo expression (23) the corresponding friction tensor. A similar methodology has been followed in constructing the position dependent friction tensor between star polymer molecules[84] and by Karniadakis' group when dealing with the Lennard-Jones blobs bounded through the radius of gyration[78].

One should note that the Green-Kubo expression (23) provides a route to calculate the friction tensor through a constrained dynamics. We have recently shown that the alternative routes through the Einstein-Helfand relation, or the Onsager regression hypothesis, are also feasible[100] to compute the transport coefficient. In fact, trying to adjust the friction coefficient in order to recover the short time dynamics of the momentum is as good (and "microscopic") as computing the Green-Kubo expression[100]. In this way, Refs. 101, 102 construct effective potentials from the radial distribution function and adjust the friction parameters in order to fit the CG dynamics. While this may seem at first sight as "just fitting" and less fundamental than the calculation of the Green-Kubo expression it is, in

fact, not so[100]. Very recently, we have introduced a generalization of Shell method in order to obtain the best diffusion process that fits a CG signal generated from MD[103]. This should allow, in principle, to recover both the potential of mean force and the friction tensors simultaneously.

A general cautionary remark needs to be formulated on the validity of a CG description for dynamics when the grouping of atoms into a CG particle gives few atoms per CG particle. In this case, we do not expect that the time scales of the CG velocity and the CG forces are clearly separated and, therefore, the Markovian assumption implicit in the Mori-Zwanzig CG method fails. In the other limit of large groupings, the velocity time scale is large because the center of mass of a big object moves much slower than its constituent atoms, while the force, which is basically determined by collisions, vary in a fast time scale. In this case the Markovian approximation is valid. When dealing with small objects, one should be aware of non-Markovian effects. Gao and Fang present[104] a proper CG of water molecules following the lines of Ref. 84 and accounting for non-Markovian behaviour by a simple rescaling of the overall friction in order to match diffusivity. The value of this empiricism relies on the fact that *viscosity* turns out to be correct with this method. However, the treatment of non-Markovian effects is an open area for more fundamental research.

6 Conclusion

DPD was a very appealing model for the simulation of complex fluids and soft matter in general, because of its simplicity and versatility. However, there has been always an uneasy feeling about what is a dissipative particle really, despite its colorful descriptions as representing many underlying atoms. The situation seems to be settled now and can be summarized as follows. If you want to model unbounded atoms with moving CG particles, then the best thing to do is to understand the dissipative particles as fluid particles (i.e. thermodynamic subsystems flowing with the flow) for which a thermodynamically consistent model based on the SPH methodology of discretizing Navier-Stokes equations exists. The resulting DPD+SPH= SDPD model does not suffer from the limitations of the original DPD model when modelling simple fluids. Of course, this requires that the fluid particles "contain many atoms", not just three or four. If one insists on working with a model of dissipative particles where each dissipative particle represents "three water molecules", there is no theory that support the picture and, therefore, one should do whatever is necessary to make sense of the simulation results.

On the other hand, if you want to model groups of bounded atoms with a dissipative particle, then there exists a solid theory for constructing the effective potentials and frictions from the underlying molecular dynamics. Despite of this solid basis, however, challenging technical problems remain in order to deal with the curse of dimensionality imposed by the dependence of the effective potential and friction with respect to all the CG variables of the system. Therefore, we still need to recourse to our best modelling skills in order to tackle the construction of CG models of the DPD type for bounded atoms.

Acknowledgements

This work has been supported by the MICINN of Spain under project FIS2010-22047-C05-03.

References

1. T.C. Germann and K. Kadau, *Trillion-atom molecular dynamics becomes a reality*, International Journal of Modern Physics C, **19**, no. 9, 1315–1319, Sept. 2008.
2. Y. Duan and P.A. Kollman, *Pathways to a Protein Folding Intermediate Observed in a 1-Microsecond Simulation in Aqueous Solution*, Science, **282**, no. 5389, 740–744, Oct. 1998.
3. P.J. Hoogerbrugge and J.M.V.A. Koelman, *Simulating Microscopic Hydrodynamic Phenomena with Dissipative Particle Dynamics*, EPL (Europhysics Letters), **19**, no. 3, 155, 1992.
4. P.Español and P. Warren, *Statistical mechanics of dissipative particle dynamics*, EPL (Europhysics Letters), **30**, no. May, 191, 1995.
5. P. Español, *Hydrodynamics from dissipative particle dynamics*, Physical Review E, **52**, no. 2, 1734–1742, 1995.
6. C.A. Marsh, G. Backx, and M.H. Ernst, *Static and dynamic properties of dissipative particle dynamics*, Physical Review E, **56**, no. 2, 1676, 1997.
7. H. Grabert, *Projection Operator Techniques in Nonequilibrium Statistical Mechanics*, Springer Verlag, Berlin, 1982.
8. P. Español, *Statistical mechanics of coarse-graining*, Novel Methods in Soft Matter Simulations, pp. 2256–2256, 2004.
9. H. Bock, K. Gubbins, and S. Klapp, *Coarse Graining of Nonbonded Degrees of Freedom*, Physical Review Letters, **98**, no. 26, 1–4, June 2007.
10. E.E. Keaveny, I.V. Pivkin, M. Maxey, and G.E. Karniadakis, *A comparative study between dissipative particle dynamics and molecular dynamics for simple- and complex-geometry flows.*, The Journal of chemical physics, **123**, no. 10, 104107, Sept. 2005.
11. J.W. van de Meent, A. Morozov, E. Somfai, E. Sultan, and W. van Saarloos, *Coherent structures in dissipative particle dynamics simulations of the transition to turbulence in compressible shear flows*, Physical Review E, **78**, no. 1, 1–4, July 2008.
12. T. Steiner, C. Cupelli, R. Zengerle, and M. Santer, *Simulation of advanced microfluidic systems with dissipative particle dynamics*, Microfluidics and Nanofluidics, **7**, no. 3, 307–323, Jan. 2009.
13. N. Filipovic, M. Kojic, and M. Ferrari, *Dissipative particle dynamics simulation of circular and elliptical particles motion in 2D laminar shear flow*, Microfluidics and nanofluidics, **10**, no. 5, 1–8, Dec. 2011.
14. K. Kadau, J.L. Barber, T.C. Germann, B.L. Holian, and B.J. Alder, *Atomistic methods in fluid simulation.*, Philosophical transactions. Series A, Mathematical, physical, and engineering sciences, **368**, no. 1916, 1547–60, Apr. 2010.
15. R.D. Groot and P.B. Warren, *Dissipative particle dynamics: Bridging the gap between atomistic and mesoscopic simulation*, The Journal of Chemical Physics, **107**, no. 11, 4423, 1997.
16. A.J. Masters and P.B. Warren, *Kinetic theory for dissipative particle dynamics: The importance of collisions*, EPL (Europhysics Letters), **48**, 1, 1999.
17. G.T. Evans, *Dissipative particle dynamics: Transport coefficients*, The Journal of chemical physics, **110**, no. 3, 1338, 1999.
18. P. Español, *Fluid particle model*, Physical Review E, **57**, no. 3, 2930, 1998.

19. R.M. Füchslin, H. Fellermann, A. Eriksson, and H.J. Ziock, *Coarse graining and scaling in dissipative particle dynamics.*, The Journal of chemical physics, **130**, no. 21, 214102, June 2009.
20. M. Arienti, W. Pan, X. Li, and G.E. Karniadakis, *Many-body dissipative particle dynamics simulation of liquid/vapor and liquid/solid interactions.*, The Journal of chemical physics, **134**, no. 20, 204114, May 2011.
21. I. Pagonabarraga and D. Frenkel, *Dissipative particle dynamics for interacting systems*, The Journal of Chemical Physics, **115**, no. 11, 5015, 2001.
22. S. Y. Trofimov, E. L. F. Nies, and M. A. J. Michels, *Thermodynamic consistency in dissipative particle dynamics simulations of strongly nonideal liquids and liquid mixtures*, The Journal of Chemical Physics, **117**, no. 20, 9383, 2002.
23. P. Warren, *Vapor-liquid coexistence in many-body dissipative particle dynamics*, Physical Review E, **68**, no. 6, 1–8, Dec. 2003.
24. Anupam Tiwari, H. Reddy, S. Mukhopadhyay, and J. Abraham, *Simulations of liquid nanocylinder breakup with dissipative particle dynamics*, Physical Review E, **78**, no. 1, 1–11, July 2008.
25. A. Ghoufi and P. Malfreyt, *Calculation of the surface tension from multibody dissipative particle dynamics and Monte Carlo methods*, Physical Review E, **82**, no. 1, 1–11, July 2010.
26. S. Merabia and I. Pagonabarraga, *Density dependent potentials: structure and thermodynamics.*, The Journal of chemical physics, **127**, no. 5, 054903, Aug. 2007.
27. P. Español, *Fluid particle dynamics: A synthesis of dissipative particle dynamics and smoothed particle dynamics*, EPL (Europhysics Letters), **39**, no. May, 605, 1997.
28. V. Pryamitsyn and V. Ganesan, *A coarse-grained explicit solvent simulation of rheology of colloidal suspensions.*, The Journal of chemical physics, **122**, no. 10, 104906, Mar. 2005.
29. W. Pan, I. V. Pivkin, and G. E. Karniadakis, *Single-particle hydrodynamics in DPD: A new formulation*, EPL (Europhysics Letters), **84**, no. 1, 10012, Oct. 2008.
30. P. Español and M. Revenga, *Smoothed dissipative particle dynamics*, Physical Review E, **67**, no. 2, 1–12, Feb. 2003.
31. J.Bonet-Avalós and A.D. Mackie, *Dissipative particle dynamics with energy conservation*, EPL (Europhysics Letters), **40**, no. 2, 141, 1997.
32. P. Español, *Dissipative particle dynamics with energy conservation*, EPL (Europhysics Letters), **40**, no. December, 631, 1997.
33. J. Bonet Avalos and A. D. Mackie, *Dynamic and transport properties of dissipative particle dynamics with energy conservation*, The Journal of Chemical Physics, **111**, no. 11, 5267, 1999.
34. M. Ripoll and P. Español, *Heat conduction modeling with energy conservation dissipative particle dynamics*, Heat Technol, **18**, no. 3, 57–61, 2000.
35. T. Yamada, A. Kumar, Y. Asako, O.J. Gregory, and M. Faghri, *Forced Convection Heat Transfer Simulation Using Dissipative Particle Dynamics*, Numerical Heat Transfer, Part A: Applications, **60**, no. 8, 651–665, 2011.
36. E. Abu-Nada, *Energy Conservative Dissipative Particle Dynamics Simulation of Natural Convection in Liquids*, Journal of Heat Transfer, **133**, no. 11, 112502, 2011.
37. M.B. Liu and G.R. Liu, *Smoothed Particle Hydrodynamics (SPH): an Overview and Recent Developments*, Archives of Computational Methods in Engineering, **17**, no. 1, 25–76, 2010.

38. P. Español, M. Serrano, and H.C. Öttinger, *Thermodynamically Admissible Form for Discrete Hydrodynamics*, Physical Review Letters, **83**, no. 22, 4542–4545, Nov. 1999.
39. L.B. Lucy, *A Numerical Approach to Testing the Fission Hypothesis*, The Astronomical Journal, **82**, no. 12, 1013–1924, 1977.
40. J.J. Monaghan, *Smoothed particle hydrodynamics*, Reports on Progress in Physics, **30**, no. 8, 543–574, 1992.
41. H. Takeda, S.M. Miyama, and M. Sekiya, *Numerical simulation of viscous flow by Smoothed Particle Hydrodynamics*, Prog. Theor. Phys., **92**, 939, 1994.
42. S.J. Watkins, A.S. Bhattal, N. Francis, J.A. Turner, and A.P. Whitworth, *A new prescription for viscosity in Smoothed Particle Hydrodynamics*, Astron. Astrophys. Suppl. Ser., **119**, 177, 1996.
43. P.W. Cleary and J.J. Monaghan, *Conduction modelling using Smoothed Particle Hydrodynamics*, J. Comp. Phys., **148**, 227, 1999.
44. P.W. Randles and L.D. Libersky, *Smoothed Particle Hydrodynamics: Some recent improvements and applications*, Computer Methods in Applied Mechanics and Engineering, **139**, no. 1-4, 375–408, 1996.
45. W.G. Hoover and H.A. Posch, *Numerical heat conductivity smooth particle applied mechanics*, Physical Review E, **54**, 5142–5146, 1996.
46. A. Vázquez-Quesada, M. Ellero, and P. Español, *Consistent scaling of thermal fluctuations in smoothed dissipative particle dynamics.*, The Journal of chemical physics, **130**, no. 3, 034901, Jan. 2009.
47. E. Somfai, A. Morozov, and W. van Saarloos, *Modeling viscoelastic flow with discrete methods*, Physica A: Statistical Mechanics and its Applications, **362**, no. 1, 93–97, Mar. 2006.
48. B.I.M. ten Bosch, *On an extension of Dissipative Particle Dynamics for viscoelastic flow modelling*, Journal of Non-Newtonian Fluid Mechanics, **83**, no. 3, 231–248, July 1999.
49. M. Ellero, P. Español, and E.G. Flekkø y, *Thermodynamically consistent fluid particle model for viscoelastic flows*, Physical Review E, **68**, no. 4, 1–19, Oct. 2003.
50. A. Vázquez-Quesada, M. Ellero, and P. Español, *Smoothed particle hydrodynamic model for viscoelastic fluids with thermal fluctuations*, Physical Review E, **79**, no. 5, 1–17, May 2009.
51. C.A.P. Thieulot, L. Janssen, and P. Español, *Smoothed particle hydrodynamics model for phase separating fluid mixtures. I. General equations*, Physical Review E, **72**, no. 1, July 2005.
52. C.A.P. Thieulot, L. Janssen, and P. Español, *Smoothed particle hydrodynamics model for phase separating fluid mixtures. II. Diffusion in a binary mixture*, Physical Review E, **72**, no. 1, 1–12, July 2005.
53. P.E. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*, Springer, Berlin, 1995.
54. I. Vattulainen, M. Karttunen, G. Besold, and J. M. Polson, *Integration schemes for dissipative particle dynamics simulations: From softly interacting systems towards hybrid models*, The Journal of Chemical Physics, **116**, no. 10, 3967, 2002.
55. I. Pagonabarraga and M.H.J. Hagen, *Self-consistent dissipative particle dynamics algorithm*, EPL (Europhysics, **377**, 1998.

56. P. Nikunen, M. Karttunen, and I. Vattulainen, *How would you integrate the equations of motion in dissipative particle dynamics simulations?*, Computer Physics Communications, **153**, no. 3, 407–423, July 2003.
57. C.P. Lowe, *An alternative approach to dissipative particle dynamics*, EPL (Europhysics Letters), **47**, 145, 1999.
58. E. Peters, *Elimination of time step effects in DPD*, EPL (Europhysics Letters), **66**, no. 1, 311, 2004.
59. S.D. Stoyanov and R.D. Groot, *From molecular dynamics to hydrodynamics: a novel Galilean invariant thermostat.*, The Journal of chemical physics, **122**, no. 11, 114112, Mar. 2005.
60. M.P. Allen and F. Schmid, *A thermostat for molecular dynamics of complex fluids*, Molecular Simulation, **33**, no. 1, 1–14, 2006.
61. E.A. Koopman and C.P. Lowe, *Advantages of a Lowe-Andersen thermostat in molecular dynamics simulations.*, The Journal of chemical physics, **124**, no. 20, 204103, May 2006.
62. L. Pastewka, D. Kauzlarić, A. Greiner, and J. Korvink, *Thermostat with a local heat-bath coupling for exact energy conservation in dissipative particle dynamics*, Physical Review E, **73**, no. 3, 8–11, Mar. 2006.
63. S. Litvinov, M. Ellero, X.Y. Hu, and N.A. Adams, *A splitting scheme for highly dissipative smoothed particle dynamics*, Journal of Computational Physics, **229**, no. 15, 5457–5464, Aug. 2010.
64. M. Serrano, G. Defabritiis, P. Español, and P.V. Coveney, *A stochastic Trotter integration scheme for dissipative particle dynamics*, Mathematics and Computers in Simulation, **72**, no. 2-6, 190–194, Sept. 2006.
65. G. Defabritiis, M. Serrano, P. Espanol, and P. Coveney, *Efficient numerical integrators for stochastic models*, Physica A: Statistical Mechanics and its Applications, **361**, no. 2, 429–440, Mar. 2006.
66. F. Thalmann and J. Farago, *Trotter derivation of algorithms for Brownian and dissipative particle dynamics.*, The Journal of chemical physics, **127**, no. 12, 124109, Sept. 2007.
67. J.M.V.A. Koelman and P.J. Hoogerbrugge, *Dynamic simulations of hard-sphere suspensions under steady state shear*, Europhysics Letters, **21**, 363, 1993.
68. M. Revenga, I. Zúñiga, P. Español, and I. Pagonabarraga, *Boundary model in DPD*, Int. J. Mod. Phys, **9**, 1319–1328, 1998.
69. I. Pivkin and G.E. Karniadakis, *Controlling Density Fluctuations in Wall-Bounded Dissipative Particle Dynamics Systems*, Physical Review Letters, **96**, no. 20, 1–4, May 2006.
70. B. Henrich, C. Cupelli, M. Moseler, and M. Santer, *An adhesive DPD wall model for dynamic wetting*, Europhysics Letters (EPL), **80**, no. 6, 60004, Dec. 2007.
71. J.P. Morris, P.J. Fox, and Y. Zhu, *Modeling Low Reynolds Number Incompressible Flows Using SPH*, Journal of Computational Physics, **136**, 214–226, 1997.
72. J. A. de la Torre and P. Español, *Coarse-graining Brownian motion: From particles to a discrete diffusion equation*, The Journal of Chemical Physics, **135**, no. 11, 114103, 2011.
73. P. Español and I. Zúñiga, *On the definition of discrete hydrodynamic variables.*, The Journal of chemical physics, **131**, no. 16, 164106, Oct. 2009.

74. P. Español, J.G. Anero, and I. Zúñiga, *Microscopic derivation of discrete hydrodynamics.*, The Journal of chemical physics, **131**, no. 24, 244117, Dec. 2009.
75. M. Serrano, P. Español, and I. Zúñiga, *Voronoi Fluid Particle Model for Euler Equations*, Journal of Statistical Physics, **121**, no. 1-2, 133–147, Oct. 2005.
76. P. Español, M. Serrano, and I. Zúñiga, *Coarse-graining of a fluid and its relation with dissipative particle dynamics and smoothed particle dynamics*, International Journal of Modern Physics C-Physics and Computer, **8**, no. 4, 899–908, 1997.
77. G.S. Ayton, H.L. Tepper, D.T. Mirijanian, and G.A. Voth, *A new perspective on the coarse-grained dynamics of fluids.*, The Journal of chemical physics, **120**, no. 9, 4074–88, Mar. 2004.
78. H. Lei, B. Caswell, and G.E. Karniadakis, *Direct construction of mesoscopic models from microscopic simulations*, Physical Review E, **81**, no. 2, 1–10, Feb. 2010.
79. P. Español, *Dissipative particle dynamics for a harmonic chain: A first-principles derivation*, Physical Review E, **53**, no. 2, 1572, 1996.
80. D. Cubero, *Comment on "Markovian approximation in a coarse-grained description of atomic systems" [J. Chem. Phys. 125, 204101 (2006)].*, The Journal of chemical physics, **128**, no. 14, 147101; author reply 147102, Apr. 2008.
81. C. Hijón, M. Serrano, and P. Español *Response to "Comment on Markovian approximation in a coarse-grained description of atomic systems" [J. Chem. Phys. 128, 147101 (2008)]*, The Journal of Chemical Physics, **128**, no. 14, 147102, 2008.
82. D. Kauzlarić, J.T. Meier, P. Español, S. Succi, A. Greiner, and J.G. Korvink, *Bottom-up coarse-graining of a simple graphene model: the blob picture.*, The Journal of chemical physics, **134**, no. 6, 064106, Mar. 2011.
83. T. Kinjo and S.A. Hyodo, *Equation of motion for coarse-grained simulation based on microscopic description*, Physical Review E, **75**, no. 5, 1–9, May 2007.
84. C. Hijón, P. Español, E. Vanden-Eijnden, and R. Delgado-Buscalioni, *Mori-Zwanzig formalism as a practical computational tool.*, Faraday discussions, **144**, no. 1, 301–22; discussion 323–45, 467–81, Jan. 2010.
85. R.L.C. Akkermans and W. J. Briels, *Coarse-grained interactions in polymer melts: A variational approach*, The Journal of Chemical Physics, **115**, no. 13, 6210, 2001.
86. T. Head-Gordon and F.H. Stillinger, *An orientational perturbation theory for pure liquid water*, The Journal of Chemical Physics, **98**, no. 4, 3313, 1993.
87. H. Meyer, O. Biermann, R. Faller, D. Reith, and F. Müller-Plathe, *Coarse graining of nonbonded inter-particle potentials using automatic simplex optimization to fit structural properties*, The Journal of Chemical Physics, **113**, no. 15, 6264, 2000.
88. J.C. Shelley, M.Y. Shelley, R.C. Reeder, S. Bandyopadhyay, and M.L. Klein, *A Coarse Grain Model for Phospholipid Simulations*, The Journal of Physical Chemistry B, **105**, no. 19, 4464–4470, May 2001.
89. S. Garde and H.S. Ashbaugh, *Temperature dependence of hydrophobic hydration and entropy convergence in an isotropic model of water*, The Journal of Chemical Physics, **115**, no. 2, 977, 2001.
90. T. Murtola, E. Falck, M. Karttunen, and I. Vattulainen, *Coarse-grained model for phospholipid/cholesterol bilayer employing inverse Monte Carlo with thermodynamic constraints.*, The Journal of chemical physics, **126**, no. 7, 075101, Mar. 2007.
91. A. Lyubartsev, A. Mirzoev, L. Chen, and A. Laaksonen, *Systematic coarse-graining of molecular models by the Newton inversion method*, Faraday Discuss., **144**, 43–56, 2010.

92. S.J. Marrink, A.H. de Vries, and A.E. Mark, *Coarse grained model for semiquantitative lipid simulations*, The Journal of Physical Chemistry B, **108**, no. 2, 750–760, 2004.
93. F. Ercolessi and J.B. Adams, *Interatomic potentials from first-principles calculations: the force-matching method*, EPL (Europhysics Letters), **26**, 583, 1994.
94. T.D. Hone, S. Izvekov, and G.A. Voth, *Fast centroid molecular dynamics: a force-matching approach for the predetermination of the effective centroid forces.*, The Journal of chemical physics, **122**, no. 5, 54105, Feb. 2005.
95. S. Izvekov and G.A. Voth, *Multiscale coarse graining of liquid-state systems*, The Journal of chemical physics, **123**, 134105, 2005.
96. W.G. Noid, J.-W. Chu, G.S. Ayton, V. Krishna, S. Izvekov, G.A. Voth, A. Das, and H.C. Andersen, *The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models.*, The Journal of chemical physics, **128**, no. 24, 244114, June 2008.
97. M.S. Shell, *The relative entropy is fundamental to multiscale and inverse thermodynamic problems.*, The Journal of chemical physics, **129**, no. 14, 144108, Oct. 2008.
98. R.L.C. Akkermans and W. J. Briels, *Coarse-grained dynamics of one chain in a polymer melt*, The Journal of Chemical Physics, **113**, no. 15, 6409, 2000.
99. R.L.C. Akkermans, *Mesoscale model parameters from molecular cluster calculations.*, The Journal of chemical physics, **128**, no. 24, 244904, June 2008.
100. D. Kauzlarić, P. Español, A. Greiner, and S. Succi, *Three Routes to the Friction Matrix and Their Application to the Coarse-Graining of Atomic Lattices*, Macromolecular Theory and Simulations, **20**, no. 7, 526–540, Aug. 2011.
101. X. Guerrault, B. Rousseau, and J. Farago, *Dissipative particle dynamics simulations of polymer melts. I. Building potential of mean force for polyethylene and cis-polybutadiene.*, The Journal of chemical physics, **121**, no. 13, 6538–46, Oct. 2004.
102. F. Lahmar and B. Rousseau, *Influence of the adjustable parameters of the DPD on the global and local dynamics of a polymer melt*, Polymer, **48**, no. 12, 3584–3592, June 2007.
103. P. Español and I. Zúñiga, *Obtaining fully dynamic coarse-grained models from MD*, Phys. Chem. Chem. Phys., pp. 1–9, 2011.
104. L. Gao and W. Fang, *Semi-bottom-up coarse graining of water based on microscopic simulations*, The Journal of Chemical Physics, **135**, no. 18, 184101, 2011.

Large-Scale Simulations of Blood Flows with Coarse-Grained Cells

Simone Melchionna

CNR-IPCF, Consiglio Nazionale delle Ricerche, P.le A. Moro 2, 00185, Rome, Italy
E-mail: simone.melchionna@roma1.infn.it

When simulating blood and the non trivial rheology arising in arbitrary flow conditions, one needs to account for red blood cells as the majority components of the suspension. I will discuss a methods to include the particulate nature of blood by introducing diffused particles. In case we need to account for the near-field hydrodynamics, the model can be promoted to a solid particle model, retaining the simplicity and robustness of the diffused model.

1 Introduction

Blood is the biological fluid of reference and hemodynamics is an active field of research, both at fundamental level and for understanding the biomechanical causes and remedies to cardiovascular diseases. For example, atherosclerosis is the leading cause of death in western countries and the biomechanical origins of it relate to the disturbed flow patterns[1]. Computer simulation provides a crucial methodology to study flow patterns with blood modeled as a continuum. However, blood circulation entails several physical levels and the usage of complex geometries, spanning from large-scale arteries to microcapillaries. Depending on the scale, blood exhibits different physical behaviors, with visco-elastic and shear-thinning response, at shear rates encountered inside large-scale arteries. The ultimate reason for the non-trivial rheology resides in the corpuscular nature of blood. In fact, more than $99\,\%$ in volume of blood is composed by plasma and red blood cells, where in physiological condition, presents a large volume fraction of red blood cells (RBC), with hematocrit level H ranging between 35 and $50\,\%$.

Red blood cells or erythrocytes are globules with flexible biconcave discs of diameter $6 - 8\,\mu m$, and resemble vesicles, as they are made by a membrane separating an internal fluid composed by hemoglobins from the external plasma. However, the RBC shape is maintained by a cytoskeleton composed of several proteins and thus RBCs are more rigid than vesicles. In presence of a shear field, red blood cells present both a solid-like tumbling and a vesicular motion with the attendant sliding of the membrane, the so-called tank treading.

A detailed representation of RBCs is crucial to study microcirculation, a situation where shape, deformability and near-field hydrodynamic response need to be accurately accounted for. Recent computational models have been put forward to represent red blood cells at this level, where the membrane and the internal fluid are explicitly represented[2–8]. On the other hand, a different class of models target large scale representations of blood usable in situations where the far-field hydrodynamics and the global rheological response of blood are reproduced. In this lecture, I will discuss a model that enables studying blood flows with a computational effort being a trade-off between physical fidelity and computational feasibility. I will discuss one such class, where red blood cells are treated as rigid or quasi-rigid entities[9,10].

I will address the issue of designing a robust physical representation of blood to be used in large-scale conditions from a double perspective. The first one is more focused on far-field hydrodynamic interactions and is provided by the Diffused Particle Model (DPM), the second one is based on a version that ameliorates the near-field hydrodynamics and I will call it the Solid Particle Model (SPM). Both models can be used successfully to modulate the fluid rheology together with accounting for cell crowding in proximity of the vessel walls. The local structuring of RBCs has strong consequences on the endothelial shear stress and other hemodynamical properties near the arterial walls, due to the inhomogeneous distribution of red blood cells and the interplay with the plasma dynamics. The present notes constitute a synthesis of the two papers[9,11], where the two models were discussed in independent ways.

2 Solvent Representation

In modeling the plasma-RBC suspension, I first consider the plasma solvent, a water-like Newtonian fluid that can be treated as a continuum. Among other computational frameworks, I will focus on the Lattice Boltzmann (LB) method, as a robust and well-behaved computational technique that reproduces the Navier-Stokes equations for an incompressible fluid, reading

$$\rho\left(\partial_t \mathbf{u} + \mathbf{u}\cdot\partial\mathbf{u}\right) = -\partial p + \eta\partial^2 u + \mathbf{G} \tag{1}$$

where ρ and $\mathbf{u}$ are the plasma density and velocity, p is the pressure and η is the dynamic viscosity. $\mathbf{G}$ is the body force acting on the fluid and we will use this term to include drag forces arising from the embedded RBCs and acting on the plasma fluid element.

The reasons for choosing LB as the embedding methodology are multiple. At first, LB is a compact and simple scheme to handle plasma both in its theoretical foundations and implementational aspects. Second, LB does not rely on a direct solution of the Navier-Stokes dynamics, but rather circumvents it by solving a minimal and effective microdynamics. Third, LB is rather tolerant in accommodating stiff hydrodynamic forces arising from the suspended particles, an aspect that confers good robustness in a variety of flow conditions (an issue that is much more delicate when dealing with direct Navier-Stokes solvers). Finally, LB reproduces the quasi-incompressible Navier-Stokes dynamics at virtually arbitrary Reynolds numbers and in arbitrary geometries. The physiological conditions of Reynolds $\lesssim 2000$ and shear rates $\lesssim 500\, s^{-1}$ can be accessed in simulation without posing limitations in terms of feasibility. A more in-depth discussion on the method can be found in Ref. 12.

The LB method is based on a microdynamics as prescribed by kinetic theory. The key idea is to evolve the single-particle distribution function $f(\mathbf{x}, \mathbf{v}, t)$ encoding the probability of having a fluid molecule at position $\mathbf{x}$, moving with velocity $\mathbf{v}$ at time t. In discrete terms, plasma is represented over a cartesian mesh having cubic symmetry and the distribution is subdivided in velocity space in elements called populations, representing the probability of moving with discrete speeds $\mathbf{c}_p$ from a mesh point to its mesh neighbors. Therefore, populations associated to $\mathbf{c}_p$ are labelled with the subscript p, as $f(\mathbf{x}, \mathbf{v}, t) \rightarrow f_p(\mathbf{x}, t)$. By using a more precise formulation, the distribution function is expanded as a second-order

Hermite polynomial in velocity space, complemented by Gauss-Hermite quadratures to evaluate the populations moments that correspond to the hydrodynamic fields[13].

The minimal form of LB is based on a relaxational dynamics of the populations towards the local statistical equilbrium, that is, the Maxwell-Boltzmann distribution. This type of dynamics is called the Bhatnagar-Gross-Krook (BGK) equation and reads[14,12]

$$\partial_t f + \mathbf{v} \cdot \partial f + \frac{\mathbf{G}}{m} \cdot \partial_v f = \frac{1}{\mathcal{T}}(f^{eq} - f) \tag{2}$$

where $\mathcal{T}$ is a characteristic relaxation time and m the fluid mass, that we take to be unity from now on (and interchange the name of the body force and acceleration as $\mathbf{G}$). It should be noticed the simple derivative in velocity space to account for the body forces, that is, a term descending from the well-known Liouvillean operator acting on the distribution.

The discrete form of the BGK dynamics over a timestep h reads

$$f_p(\mathbf{x} + h\mathbf{c}_p, t + h) = f_p^*(\mathbf{x}, t) \tag{3}$$

with $f_p^*(\mathbf{x}, t)$ being called the post-collisional population,

$$f_p^* = (1 - \frac{h}{\mathcal{T}})f_p + \frac{h}{\mathcal{T}} f_p^{eq} + h\Delta f_p^{drag} \tag{4}$$

and where the term Δf_p^{drag} accounts for the presence of suspended RBC that act as body forces on the plasma in a hydrokinetic way. The time evolution is given by an upwind Euler propagation, i.e. at first sight this implies a first-order accurate discrete evolution. But this is not completely true, since given the special nature of the streaming operator, the evolution can be made second-order accurate with a slight modification of the basic Euler scheme, as shown later on. Moreover, f_p^{eq} is the Maxwell-Boltzmann equilibrium expressed as a second-order low-Mach expansion in the fluid velocity $\mathbf{u}$,

$$f_p^{eq} = w_p \rho \left[1 + \frac{\mathbf{u} \cdot \mathbf{c}_p}{c^2} + \frac{(\mathbf{u} \cdot \mathbf{c}_p)^2 - c^2 u^2}{2c^4} \right] \tag{5}$$

Eqs. 3-4 encode the effect of streaming, that is, the motion of free particles along straight trajectories, together with the solvent-solvent and the solvent-solute "molecular" collisions. The plasma kinematic viscosity ν relates to the relaxation time $\mathcal{T}$ via $\nu = c^2(\mathcal{T} - h/2)$, where c is the plasma sound speed. A detailed theoretical analysis based on the Chapman-Enskog multiscale analysis indicates that the LB dynamics recovers Newtonian rheology in the macroscopic space/time limit.

For LB a popular choice is to employ the D3Q19 lattice scheme, where one has $c = 1/\sqrt{3}$, and w_p stands for a set of normalized weights with $p = 0, ..., 18$, being equal to $w_p = 1/3$ for the population corresponding to the null discrete speed $\mathbf{c}_0 = (0, 0, 0)$, $w_p = 1/18$ for the ones connecting first mesh neighbors $\mathbf{c}_{1,...,6} = (\pm 1, 0, 0)$, $(0, \pm 1, 0)$, $(0, 0, \pm 1)$, and $w_p = 1/36$ for second mesh neighbors, $\mathbf{c}_{7,...,18} = (\pm 1, \pm 1, 0)$, $(\pm 1, 0, \pm 1)$, $(0, \pm 1, \pm 1)$.

Concerning the palsma-particle interactions, the drag term has the following general expression

$$\Delta f_p^{drag} = h w_p \rho \left[\frac{\mathbf{G} \cdot \mathbf{c}_p}{c^2} + \frac{(\mathbf{G} \cdot \mathbf{c}_p)(\mathbf{u} \cdot \mathbf{c}_p) - c^2 \mathbf{G} \cdot \mathbf{u}}{c^4} \right] \tag{6}$$

being the corresponding second-order Hermite expansion of the body force $\mathbf{G}$[13].

The local plasma density ρ, speed $\mathbf{u}$ and momentum-flux tensor $\mathbf{P}$, are given by the following Gauss-Hermite quadratures of the populations

$$\rho = \sum_p f_p \tag{7}$$

$$\rho \mathbf{u} = \sum_p f_p \mathbf{c}_p \tag{8}$$

$$\mathbf{P} = \sum_p f_p \mathbf{c}_p \mathbf{c}_p \tag{9}$$

These quadratures can be slightly modified in order to ensure second order space/time accuracy of the LB algorithm, being equivalent to a trapezoidal temporal integration[15]. The modification of Eq. 8 reads

$$\rho \mathbf{u} = \sum_p f_p \mathbf{c}_p + \frac{h}{2} \rho \mathbf{G} \tag{10}$$

The kinetic representation of the solvent offers other important advantages. One of them regards the off-diagonal component of the momentum-flux giving the deviatoric shear stress $\boldsymbol{\sigma}$. In the LB scheme, this is related to the non-equilibrium component of the populations and can be computed locally, that is, without using finite difference schemes. The following expression holds

$$\boldsymbol{\sigma} \equiv \nu \rho \left(\boldsymbol{\partial} \mathbf{u} + \boldsymbol{\partial} \mathbf{u}^T \right) = -\frac{3\nu}{c^2 \mathcal{T}} \sum_p \mathbf{c}_p \mathbf{c}_p \left(f_p - f_p^{eq} \right) \tag{11}$$

On the other hand, due to the lack of a local kinetic definition of the antisymmetric component of the displacement tensor $\partial \mathbf{u}$ and the fluid vorticity, these are evaluated via finite-differences.

The local knowledge of the stress tensor is important in hemodynamics because one important indicator of cardiovascular disease is the Endothelial Shear Stress (ESS), a quantity related to the biomechanical disturbances occurring on the vascular endothelium due to the shearing forces induced by the plasma. The ESS is quantified by the second invariant of the stress tensor as $ESS = \sqrt{\frac{1}{2} \boldsymbol{\sigma} : \boldsymbol{\sigma}}$.

3 Diffused Particle Model (DPM)

Let us now turn to designing a workable model for red blood cells. The DPM is a simple and effective way to include the hydrodynamic interactions mediated by the surrounding plasma solvent. As will become apparent in the following, fluid-particle exchange mechanisms can be entirely handled at kinetic level, that is, governed by appropriate collisional terms that avoid to compute hydrodynamic forces and torques via the Green's function method, as employed in Stokesian dynamics[16], a fundamental advantage of hydrokinetic modeling, resulting in an order N computational cost (in sharp constract with Stokesian or Brownian dynamics).

The key idea of DPM is to represent a single cell as an effective diffused body without handling explicitely the globule membrane. Let us first remark that we need to mimick the

inner fluid carried by cell since this contributes significantly to the dissipation of energy by the suspension, with a steep raise in the apparent viscosity with the hematocrit level. One simple way to incorporate the viscosity contrast between the inner and outer fluid is to evolve the solvent with the LB method as a single fluid. In addition, we consider a local enhancement of the LB fluid viscosity within the RBC shape according to the following BGK relaxation time

$$\mathcal{T}(\mathbf{x}) = \mathcal{T}_0 + \Delta h \sum_i \theta(\mathbf{x} - \mathbf{R}_i) \tag{12}$$

where $\mathcal{T}_0$ corresponds to the viscosity of pure plasma and θ_i is the globule characteristic function. The prefactor Δ is a viscosity enhancement factor that can be tuned at will to change the viscosity of the inner fluid.

A RBC is then represented as a diffused ellipsoidal particle, with mass M, position $\mathbf{R}_i$, velocity $\mathbf{V}_i$, angular velocity $\mathbf{\Omega}_i$, and instantaneous orientation given by the matrix

$$\mathbf{Q}_i = \begin{pmatrix} \hat{n}_{x,i} \; \hat{t}_{x,i} \; \hat{g}_{x,i} \\ \hat{n}_{y,i} \; \hat{t}_{y,i} \; \hat{g}_{y,i} \\ \hat{n}_{z,i} \; \hat{t}_{z,i} \; \hat{g}_{z,i} \end{pmatrix} \tag{13}$$

where $\hat{\mathbf{n}}_i$, $\hat{\mathbf{t}}_i$, $\hat{\mathbf{g}}_i$ are orthogonal unit vectors, such that $\mathbf{Q}_i^T \mathbf{Q}_i = \mathbf{1}$. The orthogonal matrix $\mathbf{Q}_i$ transforms between the body and the laboratory frame via $\mathbf{v}' = \mathbf{Q}_i \mathbf{v}$, where the primed and unprimed symbols stand for laboratory and body frames, respectively. The tensor of inertia, $\mathbf{I}_i$, is diagonal in the body frame and transforms to the laboratory frame according to $\mathbf{I}_i' = \mathbf{Q}_i \mathbf{I}_i \mathbf{Q}_i^T$. In the sequel, we shall drop the prime symbol to ease the notation and implicitly mean that the translational motion is handled in the laboratory frame, where the rotational motion is handled in the body frame. We collectively denote the roto-translational state by the symbol $\mathbf{\Gamma}_i \equiv (\mathbf{R}_i, \mathbf{Q}_i, \mathbf{V}_i, \mathbf{\Omega}_i)$.

We introduce an auxiliary function to account for the shape and orientation of the suspended body and choose the following expression, as borrowed from the immersed boundary method[17],

$$\tilde{\delta}(\mathbf{x}, \mathbf{Q}_i) \equiv \prod_{\alpha=x,y,z} \tilde{\delta}_\alpha[(\mathbf{Q}_i \mathbf{x})_\alpha] \tag{14}$$

with

$$\tilde{\delta}_\alpha(y_\alpha) \equiv \begin{cases} \frac{1}{8}\left(5 - 4|y_\alpha/\xi_\alpha| - \sqrt{1 + 8|y_\alpha|/\xi_\alpha - 16 y_\alpha^2/\xi_\alpha^2}\right) & |y_\alpha/\xi_\alpha| \le 0.5 \\ \frac{1}{8}\left(3 - 4|y_\alpha|/\xi_\alpha - \sqrt{-7 + 24|y_\alpha|/\xi_\alpha - 16 y_\alpha^2/\xi_\alpha^2}\right) & 0.5 < |y_\alpha/\xi_\alpha| \le 1 \\ 0 & |y_\alpha|/\xi_\alpha > 1 \end{cases} \tag{15}$$

and ξ_α being a set of three integers, one for each cartesian component $\alpha = x, y, z$, representing the ellipsoidal radii in the three principal directions. The shape function has compact support and for $\xi_x = \xi_y = \xi_z = 2$ generates a spherically symmetric diffused particle with a support extending over 64 mesh points. Given that we handle one RBC via a single shape function, the computational cost of a suspended RBC is proportional to the size of the support in the three cartesian directions.

In this computational model, the particle shape function has two important properties, it is normalized when summed over the cartesian mesh points $\mathbf{x}$, $\sum_{\mathbf{x}} \tilde{\delta}(\mathbf{x} - \mathbf{T}) = 1$ for any continuous displacement $\mathbf{T}$, and obeys the property $\sum_{\mathbf{x}} (x_\alpha - T_\alpha) \partial_\beta \tilde{\delta}(\mathbf{x} - \mathbf{T}) = -\delta_{\alpha\beta}$.

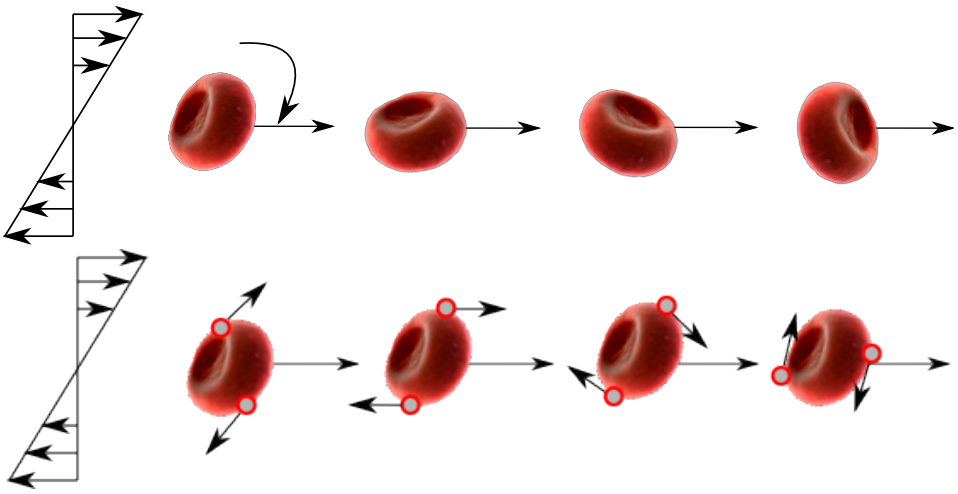


Figure 1. The two types of motion considered in modeling a RBC in a linear shear field, the solid-like, tumbling or flipping coin motion (upper panel), and the vesicular, tank-treading motion (lower panel), where two material points on the RBC membrane move at fixed body orientation.

The translational response of the suspended body is designed according to the RBC-fluid exchange kernel

$$\boldsymbol{\phi}(\mathbf{x}, \boldsymbol{\Gamma}_i) = -\gamma_T \tilde{\delta}(\mathbf{x} - \mathbf{R}_i, \mathbf{Q}_i) \left[\mathbf{V}_i - \mathbf{u}(\mathbf{x})\right] = -\gamma_T \tilde{\delta}_i \left(\mathbf{V}_i - \mathbf{u}\right) \tag{16}$$

where γ_T is a translational coupling coefficient and where the short-hand notation $\tilde{\delta}_i \equiv \tilde{\delta}(\mathbf{x} - \mathbf{R}_i, \mathbf{Q}_i)$ has been introduced.

The body rotational response has different origins and can be analyzed by considering the general decomposition of the deformation tensor in terms of purely elongational and rotational terms

$$\boldsymbol{\partial}\mathbf{u} = \boldsymbol{e} + \boldsymbol{\rho}$$

where $\mathbf{e} = \frac{1}{2}(\boldsymbol{\partial}\mathbf{u} + \boldsymbol{\partial}\mathbf{u}^T)$ is the symmetric rate of strain tensor, related to the dissipative character of the flow, and $\boldsymbol{\rho} = \frac{1}{2}(\boldsymbol{\partial}\mathbf{u} - \boldsymbol{\partial}\mathbf{u}^T)$ is the antisymmetric vorticity tensor, which bears the conservative component of the flow and is related to the vorticity vector $\boldsymbol{\omega} = \boldsymbol{\partial} \times \mathbf{u} = \boldsymbol{\epsilon} : \boldsymbol{\rho}$, where $\boldsymbol{\epsilon}$ is the Levi-Civita tensor[18]. The rotational component of the deformation tensor gives rise to a solid-like tumbling motion, where the rotational and elongational one give rise to the vesicular, tank treading motion, as illustrated in Fig. 1. Consequently, at rotational level, the DPM experiences two distinct components of the torque. The first one arises from the coupling between the body motion and the fluid vorticity, that we represent by the following rotational kernel

$$\boldsymbol{\tau}^A(\mathbf{x}, \boldsymbol{\Gamma}_i) = -\gamma_R \tilde{\delta}(\mathbf{x} - \mathbf{R}_i, \mathbf{Q}_i) \left[\boldsymbol{\Omega}_i - \boldsymbol{\omega}(\mathbf{x})\right] = -\gamma_R \tilde{\delta}_i \left(\boldsymbol{\Omega}_i - \boldsymbol{\omega}\right) \tag{17}$$

where γ_R is a rotational coupling coefficient and the superscript A stands for antisymmetric. This term depends on the body shape and orientation via the shape function and, in a linear shear flow, generates angular motion at constant angular velocity.

The elongational component of the flow contributes to the orientational torque for bodies with ellipsoidal symmetry, being zero for spherical solutes[19]. By defining the stress vector $\mathbf{t}^{\boldsymbol{\sigma}} = \boldsymbol{\sigma} \cdot \hat{\mathbf{n}}$, where $\hat{\mathbf{n}}$ is the outward normal to the surface of the DPM, we replace the surface normal with the vector spanning over the entire volume of the diffused particle, $\hat{\mathbf{n}} = \boldsymbol{\partial}\tilde{\delta}/|\boldsymbol{\partial}\tilde{\delta}|$. The associated torque is represented in analogy with the torque acting on macroscopic bodies[18], by the kernel

$$\boldsymbol{\tau}^S(\mathbf{x}, \boldsymbol{\Gamma}_i) = \alpha\tilde{\delta}_i \mathbf{t}^{\boldsymbol{\sigma}} \times (\mathbf{x} - \mathbf{R}_i) \tag{18}$$

where α is a parameter to be fixed and the superscript S is mnemonic for the symmetric contribution of the flow. As shown in the following, the elongational component of the torque includes an independent contribution arising from tank treading that will allow us to tune the parameter α based on known data on the tumbling to tank treading transition.

The hydrodynamic force and torque acting on the DPM are obtained via integration over the globule spatial extension. Owing to the discrete nature of the mesh, the integrals are written as discrete sums,

$$\mathbf{F}_i = \sum_{\mathbf{x}} \boldsymbol{\phi}(\mathbf{x}, \boldsymbol{\Gamma}_i) = -\gamma_T(\mathbf{V}_i - \tilde{\mathbf{u}}_i) \tag{19}$$

$$\mathbf{T}_i^A = \sum_{\mathbf{x}} \boldsymbol{\tau}^A(\mathbf{x}, \boldsymbol{\Gamma}_i) = -\gamma_R(\boldsymbol{\Omega}_i - \tilde{\boldsymbol{\omega}}_i) \tag{20}$$

$$\mathbf{T}_i^S = \sum_{\mathbf{x}} \boldsymbol{\tau}^S(\mathbf{x}, \boldsymbol{\Gamma}_i) \tag{21}$$

where

$$\tilde{\mathbf{u}}_i \equiv \tilde{\delta}_i \star \mathbf{u} = \sum_{\mathbf{x}} \tilde{\delta}_i \mathbf{u} \tag{22}$$

$$\tilde{\boldsymbol{\omega}}_i \equiv \tilde{\delta}_i \star \boldsymbol{\omega} = \sum_{\mathbf{x}} \tilde{\delta}_i \boldsymbol{\omega} \tag{23}$$

are smeared hydrodynamic fields and the symbol $\star$ denotes convolution over the mesh.

The action of the forces $\mathbf{F}_i$ and torques $\mathbf{T}_i = \mathbf{T}_i^S + \mathbf{T}_i^A$ are counterbalanced by opposite reactions on the fluid side. Conservation of linear and angular momentum in the composite fluid-particle system preserves the basic symmetries of the microdynamics and produces the consistent hydrodynamic response[20]. The action of forces and torques on the fluid populations are expressed according to the following expression

$$\mathbf{G} = -\sum_i \left\{ \mathbf{F}_i \tilde{\delta}_i + \frac{1}{2} \mathbf{T}_i \times \boldsymbol{\partial}\tilde{\delta}_i \right\}$$

The two exchange terms arising from the translational and rotational back-reactions produce distinct modifications of the fluid velocity and vorticity. Some algebra shows that every suspended body preserves mass and linear momentum in the composite fluid-RBC system since $\sum_{\mathbf{x}} \Delta\mathbf{u} = \sum_{\mathbf{x}} \sum_p \Delta f_p \mathbf{c}_p = -\mathbf{F}_i$. Similarly, any suspended body preserves the total angular momentum, since

$$\sum_{\mathbf{x}} \Delta\boldsymbol{\omega} = \sum_{\mathbf{x}} \sum_{\mathbf{p}} \Delta f_p \mathbf{c}_p \times (\mathbf{x} - \mathbf{R}_i) = \frac{1}{2} \sum_{\mathbf{x}} \left(\mathbf{T}_i \times \boldsymbol{\partial}\tilde{\delta}_i \right) \times (\mathbf{x} - \mathbf{R}_i) = -\mathbf{T}_i \tag{24}$$

where we have used the Lagrange rule, $\mathbf{a}\times(\mathbf{b}\times\mathbf{c}) = (\mathbf{a}\cdot\mathbf{c})\mathbf{b} - (\mathbf{a}\cdot\mathbf{b})\mathbf{c}$ and the properties of the shape function.

Tank Treading is the motion that arises from the sliding of the surrounding membrane when subjected to a shearing flow. It can be visualized by considering a material point anchored to the RBC membrane and moving in elliptical orbits while maintaining fixed the orientation of the RBC with respect to the shearing direction (see Fig. 1). The physical parameters controlling tank treading are the ratio of viscosity between plasma and the inner fluid entrained within the RBC, the shape of the RBC and the shear rate $\dot{\gamma}$[21]. It is essential to include tank treading as it acts to orient cells with a privileged angle with respect to the flow direction. In proximity of the vessel walls, the fixed orientation induces a net lift force proportional to $\dot{\gamma}$ that pushes the globule away from the walls[22]. Lift forces, arising from either single-body or many-body effects, are thought to induce the Farhaeus-Lindqvist phenomenon, the drop of blood viscosity in vessels of sub-millimeter diameters, an effect with far-reaching consequences in physiology[23].

In the general case, tank treading takes place in an apparently decoupled fashion from tumbling, as vorticity and elongational contributions have different origins. However, the motion of the rigid-body RBC is partially compensated by the movement of the surrounding membrane and thus effectively couples tumbling and tank treading via the instantaneous orientation of the globule. At small values of the viscosity ratio, orientational torques prevail over the rotational ones and pure tank treading motion with a fixed RBC orientation is observed.

In general, the torque acting on a RBC can be decomposed into three separate components. The first component is due to the angular motion of RBC in a quiescent fluid and produces a frictional torque given by $-\gamma_R\mathbf{\Omega}$, where γ_R is a phenomenological coefficient that can be identified with the one introduced in Eq. 17. The second component is due to the torque from the shearing fluid on the quiescent RBC and for a quiescent membrane, being related to the vorticity component of the flow. The effect of tank treading is determined by considering the local frame moving together with the material point at velocity $\mathbf{V}_{TT}$ and with the infinitesimal membrane element experiencing a force $d\mathbf{F}_{TT} \propto \mathbf{V}_{TT}$ and torque $\boldsymbol{\tau}_{TT} = \oint d\mathbf{F}_{TT}\times(\mathbf{x}-\mathbf{R})$. Consequently, tank treading couples to both the rotational and elongational flow components and results in a net torque with the same angular symmetry of the mechanism associated to the rigid body response. In the DPM, the presence of the cellular membrane is not explicitly considered but tank treading is controlled by tuning the intensity of the elongational torque via the adimensional prefactor α of Eq. 18.

To recap, the DPM is based on the body forces and torques acting on the RBC rather than on surface forces by a proper decomposition of the translational, tumbling and orientational components of the flow. Thus, the suspended cells are active scalars with hydrodynamic shape that is fixed and ellipsoidal. Clearly, the eccentricity of the RBC can be tuned at will. For example, by choosing $\xi_x = 1$ and $\xi_y = \xi_z = 2$, this corresponds to volume $V \simeq 134$, surface $S \simeq 139$ and reduced volume $v \equiv \frac{3V}{4\pi(S/4\pi)^{3/2}} \simeq 0.87$, to be compared with $v = 0.65$ for human RBCs. Finally, in the DPM shape fluctuations can be also introduced by following the analysis of the Maffettone-Minale model[24], by allowing volume-preserving deformations while still maintaining the ellipsoidal symmetry;

In summary, the main assumption of the DPM, i.e. the neglect of shape fluctuations, is largely compensated by its main strength, the possibility to handle large-scale systems of physiological relevance with state-of-the-art computer hardware. This is a major point of

the model that will not be addressed in detail here, but has been discussed in Refs. 25–27.

I will now discuss some numerical results that provide a bird-eye view on the properties featured by the DPM.

The frictional response of a single RBC in plasma is analyzed by computing the Stokes response of an oblate ellipsoidal particle having two possible orientations with respect to the motion direction. For translational displacement, these are the frontal and side-wise motions. For the rotational motion, these correspond to spinning around two principal directions corresponding to the smallest and largest radii. As shown in Ref. 28 for a model of suspended point-like particles, the effective mobility is given by the sum of two components, the mobility associated to the bare frictional parameters, γ_T and γ_R, and the effect of the hydrodynamic field induced on the surrounding solvent that sustains the motion by increasing the particle roto-translational mobilities. In addition, the hydrodynamic components to mobility contains a Stokes-like component that is renormalized by the presence of the numerical finite-spacing mesh[28].

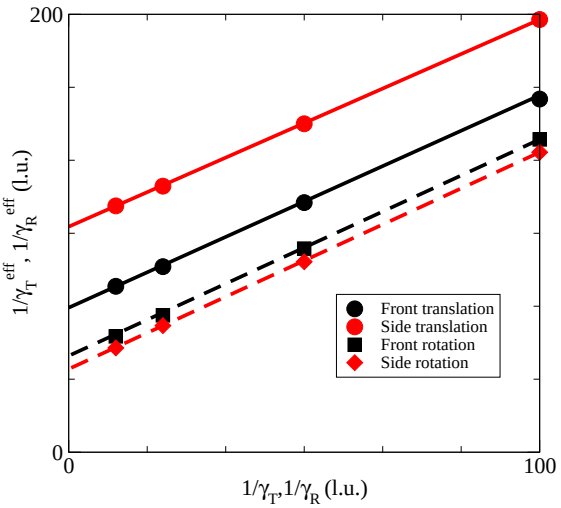


Figure 2. Translational and rotational mobilities as a function of the coupling parameters γ_T and γ_R. Circles correspond to frontal (filled symbols) and lateral translation (open symbols). Squares are for rotations around the smallest principal radius (filled symbols) and around the largest principal radius (open symbols). The lines are guides for the eye.

Fig. 2 shows the computed particle mobilities as a function of the frictional parameters γ_R and γ_T for a RBC of mass $M = 10$ and inertia $I_{x,y,z} = 1000$ (all data are expressed in lattice units unless otherwise stated). At infinite friction, the intercepts correspond to the mesh-induced spurious frictional forces. By associating a residual hydrodynamic radius to frontal and side-wise motion, respectively, one finds that the residual radius is directly

proportional to the size factor ξ_α governing the shape function. In addition, the residual Stokes radii are $a_T^{mesh} = \frac{M\gamma_T^{mesh}}{6\pi\nu\rho} = 0.06$ and 0.03 to frontal and side-wise motion and thus are much smaller than the mesh spacing (being $\Delta x = 1$ in lattice units). Thus modulating the hydrodynamic response to achieve a bulkier suspended body requires increasing the coupling parameter, of the order of $\gamma_T \sim 10$ to have a Stokes radius of order one. A back-of-the-envelop stability analysis shows at most LB can handle a coupling coefficient $\gamma_T < 0.5$ before breaking down, therefore with the DPM we always end up with a tiny hydrodynamic radius. Luckily this is now dramatic, since the far-field hydrodynamics is going to be correctly reproduced.

For angular motion, the intercepts correspond to the residual rotational radii and the mesh-induced friction is larger than the translational counterpart (in fact, comparable to the mesh spacing) and exhibits a weak dependence on the direction of spinning direction, so that it can be considered independent on the latter.

The non-Newtonian behavior of the suspension is further exhibited by the velocity profiles of RBC for different hematocrit levels and vessel diameters, as shown in Fig. 3A. As the hematocrit level increases, the Poiseuille-like parabola modifies into flatter profiles next to the vessel centerline and in a large extension of the channel, whereas in proximity of the walls, the profiles have large slopes and strong dissipation, in particular for the narrower vessels.

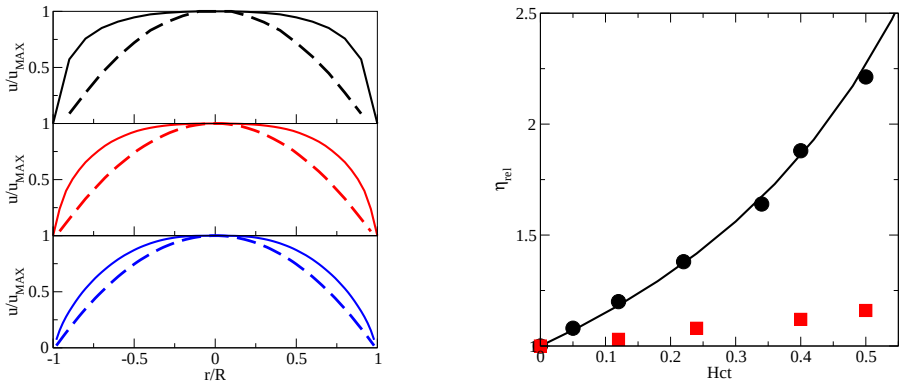


Figure 3. Panel A: velocity profiles for vessel radius of 10 (upper panel), 25 (mid panel) and 50 μm (lower panel). Data correspond to hematocrit levels of 35% (solid lines) and 0% (dashed lines). Panel B: relative viscosity in a channel of radius 50 μm for different hematocrit levels as compared to the experimental data of Pries et al.[29](solid curve). Data are for the enhanced dissipation mechanism of Eq. 12 with $\Delta = 2$ (circles) and without enhancement ($\Delta = 0$) (squares).

The viscosity of the suspension for a cylindrical channel of radius 50 μm is reported in Fig. 3B, where the relative viscosity is $\eta_{rel} = \eta_{app}/\eta_0$, with η_{app} being the apparent viscosity measured in the channel at finite hematocrit and η_0 the viscosity in the same channel

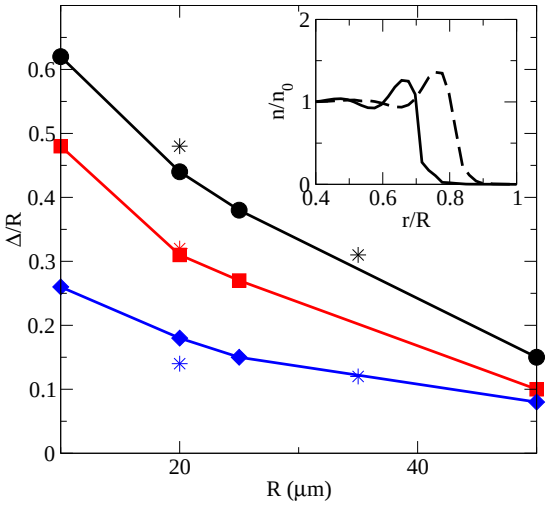


Figure 4. Size dependence of the cell-free layer with the vessel diameter and hematocrit level of $10\,\%$ (diamonds), $20\,\%$ (squares) and $50\,\%$ (circles), as compared to the experimental data of Bugliarello and Sevilla[30] (star symbols). The lines are guides for the eye. Inset: Radial density profiles of RBC, for $R = 10\,\mu m$ (solid line) and $R = 20\,\mu m$ (dashed line), illustrating the cell-free layers in proximity of the vessel wall.

at zero hematocrit. The figure also reports the data on viscosity by setting the enhancement factor of Eq. 12 to $\Delta = 0$. The latter produce a weak modulation of viscosity with hematocrit, while the data with $\Delta = 2$ exhibit an excellent agreement with the experimental results of Ref. 29.

A crucial feature of blood circulation is the decrease of viscosity in a cylindrical vessel, as the vessel radius falls below $100\,\mu m$, namely, the Farhaeus-Lindqvist effect[23]. This effect is ascribed to the formation of a cell-free layer in proximity of the vessel walls. The origin of such depletion is still uncertain but the lateral forces that push the RBCs away from the vessel walls are retained to have different causes, such as tank treading and cell deformation[22], adhesive properties of RBC or shear-induced migration. In the current version of our model, we do not probe the effects of cell deformation. However, the simulation reveals a distinct RBC depletion in proximity of the walls, as shown in Fig. 4. The numerical results reproduce the experimental data quite well, lending good confidence in the numerical model at vessel diameters below the $100\,\mu m$ radius.

We apply the DPM to physiological conditions by considering a realistic bifurcating vessel at 50% hematocrit level, as depicted in Fig. 5A. The bifurcation is extracted as part of a coronary arterial system[31], and is made of a parent vessel of radius $\sim 100\,\mu m$ and one daughter branch having approximately the same size of the parent vessel, and a second daughter branch of radius $\sim 80\,\mu m$. As the snapshot in Fig. 5B reveals, RBCs organize in several different ways throughout the bifurcation and also depending on the value of the

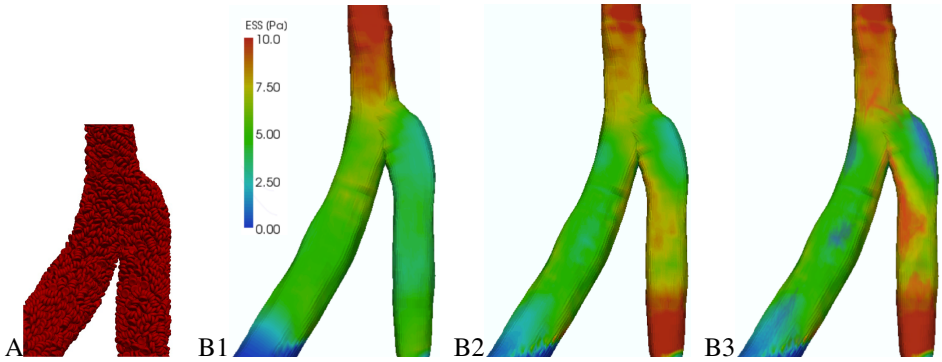


Figure 5. Panel A: detail of the bifurcating vessel showing the organization of RBCs at hematocrit of $50\,\%$ and average shear rate of $80\,s^{-1}$. Panels B1, B2, B3: time-averages of shear stress for pure plasma (B1), hematocrit level of $35\,\%$ (B2) and $50\,\%$ (B3).

shear rate (data not shown). The local organization of RBCs in rouleaux is visible, the typical stack often observed in static or flow conditions, and mostly destroyed as the shear rate increases.

The uneven distribution of RBCs, with the attendant stagnation and persistence of rouleaux in specific regions, can have significant impact on the distribution of shear stress. In particular, low levels of shear stress, as due to disturbed flow patterns and stagnation regions, trigger the growth of plaques. Fig. 5B illustrates the distribution of shear stress for different hematocrit levels. The plot reveals the strong effect of the RBCs throughout the system and in particular the great fluctuations in one daughter vessel. While the overall shear stress distribution is somehow preserved at different hematocrit levels, important local modifications are induced by RBCs. In particular, in proximity of the vessel shoulder, the RBC structuring induces smaller values of the shear stress, followed by larger values next to the inner side of the bifurcation.

4 Solid Particle Model (SPM)

One limitation of the DPM is the fact that the fluid-particle body forces are unable to expell the solvent streamlines from entering the particle extension. This is a clear consequence of the diffused nature of the particle (in fact the DPM is a refined version of a point-like particle) and being in line with the idea of reproducing only the far-field hydrodynamic beviour, akint to the Rotne-Prager level of hydrodynamics[20].

One possible way to imposing no-slip boundary conditions at the particle surface is to explicitly reproduce the particle-plasma interface. A popular scheme is to adapt the model introduced by A.J.C. Ladd used to study colloids[32], where the suspended body is represented as a solid boundary moving within the solvent. An extension of this technique to particle with ellipsoidal shape was presented in Ref. 10. This scheme reproduces hemorheology to high accuracy. One limitation of such method, however, is the fact that the particle surface is a given by a staircased representation, following the underlying cartesian mesh. As RBCs move and rotates in the continuum, the staircased representation changes accord-

ingly, with a variable number of mesh points that enter and exit the particle extension. To avoid numerical artifacts arising from such staircased representation, and in particular any spurious forces and numerical instability, each RBC should contain generally a rather large (~ 100) number of mesh points, resulting in a substantial computational overhead.

We now discuss a model that is somehow a trade-off between the DPM and Ladd's model: i) it circuments the harsh staircased representation, ii) retains the simplicity of DPM and iii) yet serves for taking a closer look at the hydrodynamic field. The idea is to promote the DPM to a solid particle model (SPM). Again, we take a rigid particle for simplicity and, in addition, we consider a spherical particle as a reference, but other shapes can be used with minor modifications of the scheme.

When considering the finite extension of the body, one crucial property of the fluid-body coupling is that it cannot be expressed as a direct force field. In fact, any localized force field acting in the particle region creates a current outside the particle extension with no physical meaning. This is exemplified by a body in a fluid at rest. If a force acts to repel the fluid momentum out of the particle, it creates a persistent current that extends radially away from the particle center. Instead, we actually need a reaction force being active only in flow conditions by creating an exclusion of the streamlines from the body region.

At first, let us consider the body as a rigid body of infinite mass and with local velocity field

$$\vec{W}(\vec{x}) = \Theta_\xi(\vec{x}) \left[\vec{V} + \vec{\omega} \times (\vec{x} - \vec{R}) \right] \tag{25}$$

where $\theta_\xi(\vec{x}) \equiv \theta(|\vec{x} - \vec{R}| - \xi)$ is the characteristic function of a sphere of radius ξ. In the continuum limit, the force acting on the particle is written as the following surface integrals over the particle spherical surface S as $\vec{F}^{drag} = \oint_S \vec{t}(\vec{x}) d\vec{x}$ and for the torque, $\vec{T}^{drag} = \oint_S \vec{t}(\vec{x}) \times \vec{x} d\vec{x}$. Here $\vec{t} = -p\hat{n} + \overleftrightarrow{\sigma} \cdot \hat{n}$ is the vector component of the stress tensor and $\hat{n}$ is the outward normal to the spherical surface.

Let us stipulate that the drag force on the fluid has the form

$$\vec{G}(\vec{x}) = -\theta_\xi(\vec{x}) \vec{\lambda}(\vec{x}) \tag{26}$$

where $\vec{\lambda}$ is a Lagrange multiplier whose value is still undetermined. The streaming step, Eq. 3, can be decomposed as $f_p(\vec{x} + h\vec{c}_p, t + h) = f_p^{*(0)}(\vec{x}, t) + \Delta f_p^{drag}(\vec{x}, t)$ where $f_p^{*(0)}(\vec{x}, t)$ is the uncorrected post-collisional contribution, i.e. the fluid populations in absence of the suspended body and . Similarly, the fluid post-streaming velocity is expressed as $\vec{u}(\vec{x}, t) = \vec{u}^{(0)}(\vec{x}, t) + h\vec{G}(\vec{x}, t)$. In the following, we shall drop the dependence on $\vec{x}$ to ease the notation, unless otherwise expressed.

The reaction force has the form of a contact force whose value is derived by matching the corrected fluid velocity, as obtained after the streaming phase, to the particle one. By exploiting the identity $\theta_\xi^2 = \theta_\xi$, the Lagrange multiplier is found to have the form $\vec{\lambda} = \frac{1}{h} \left(\vec{u}^{(0)} - \vec{W} \right)$ and the fluid corrected velocity reads

$$\vec{u} = \vec{u}_{NoSlip} \equiv (1 - \theta_\xi) \vec{u}^{(0)} + \theta_\xi \vec{W} \tag{27}$$

that matches the body velocity field at the surface and inside the body domain, that is, the no-slip boundary condition.

We now consider a smooth suspended body of infinite mass and approximate the Heaviside function as a smooth shape function

$$\tilde{\theta}_\xi^{(k)}(a) = 1 - \left[1 - \tilde{\delta}_\xi(a)\right]^k \tag{28}$$

with k being an integer parameter that controls the smoothness of the particle, since$\lim_{k\to\infty} \tilde{\theta}_\xi^{(k)}(a) = \theta_\xi(a)$. The body hydrodynamic shape follows the shape function $\tilde{\theta}_\xi^{(k)}$, being 1 inside the body, 0 outside, and decaying smoothly to zero at the body-fluid interface, that is, with a small penetration of the fluid at the interface for finite k.

The reaction force (26) can be approximated by the following iterative correction of the fluid velocity

$$\begin{aligned}
\vec{u}^{(1)} &= (1 - \tilde{\delta}_\xi)(\vec{u}^{(0)} - \vec{W}) + \vec{W} \\
&\cdots \\
\vec{u}^{(k)} &= (1 - \tilde{\delta}_\xi)(\vec{u}^{(k-1)} - \vec{W}) + \vec{W} = (1 - \tilde{\theta}_\xi^{(k)})\vec{u}^{(0)} + \tilde{\theta}_\xi^{(k)}\vec{W}
\end{aligned} \tag{29}$$

that converges to the sought no-slip solution (27) for a sharp body, i.e. $\lim_{k\to\infty} \vec{u}^{(k)} = \vec{u}_{NoSlip}$. In other words, the function $\tilde{\delta}_\xi$ allows to construct incremental and systematic corrections to the fluid velocity as successive sweeping steps and the original problem of handling the boundary contact force at sharp body-fluid interface is rewritten as a volume reaction force within the extension of a smooth body.

On the body side, force balance is such that the body experiences a drag force

$$\vec{F}^{drag} = \sum_{\vec{x}} \Delta x^3 \rho(\vec{x}) \vec{G}(\vec{x}) = \sum_{\vec{x}} \frac{\Delta x^3}{h} \rho(\vec{x}) \tilde{\theta}_\xi^{(k)}(\vec{x}) \left[\vec{u}^o(\vec{x}) - \vec{W}(\vec{x})\right] \tag{30}$$

together with a torque given by

$$\begin{aligned}
\vec{T}^{drag} &= \sum_{\vec{x}} \Delta x^3 \rho(\vec{x}) \vec{G}(\vec{x}) \times (\vec{x} - \vec{R}) \\
&= \sum_{\vec{x}} \frac{\Delta x^3}{h} \rho(\vec{x}) \tilde{\theta}_\xi^{(k)}(\vec{x}) \left[\vec{u}^o(\vec{x}) - \vec{W}(\vec{x})\right] \times (\vec{x} - \vec{R})
\end{aligned} \tag{31}$$

The case of a body of *finite* mass is constructed along similar lines. In this case, however, the fluid and body velocities are corrected simultaneously. This effectively introduces a non-local coupling between fluid elements that are solved by inversion of a linear problem. To see this, let us first consider the forces on the body and add the drag force besides the mechanical force $\vec{F}^{mech}$ as

$$M\frac{d\vec{V}}{dt} = \vec{F}^{tot} \equiv \vec{F}^{mech} + \sum_{\vec{x}} \Delta x^3 \rho(\vec{x}) \tilde{\theta}_\xi^{(k)}(\vec{x} - \vec{R}) \vec{\Lambda}(\vec{x}) \tag{32}$$

and the drag torque enters the rotational dynamics as

$$I\frac{d\vec{\omega}}{dt} = \sum_{\vec{x}} \Delta x^3 \rho(\vec{x}) \tilde{\theta}_\xi^{(k)}(\vec{x} - \vec{R}) \vec{\Lambda}(\vec{x}) \times (\vec{x} - \vec{R}) \tag{33}$$

where the unknown $\vec{\Lambda}$ is yet to be determined.

Let us consider as a reference the Velocity Verlet propagation of the particle position, velocity and angular velocity. The unknown $\vec{\Lambda}$ is obtained at each update of the particle velocity to correct for both the fluid and the body velocity. By focusing on the first half of the velocity update, the equation to correct the fluid velocity locally is

$$\vec{u}(\vec{x}, t + \frac{h}{2}) = \vec{u}^{(0)}(\vec{x}, t) - \frac{h}{2}\tilde{\theta}_{\xi}^{(k)}(\vec{x} - \vec{R}(t))\vec{\Lambda}(\vec{x}) \tag{34}$$

and similarly for the particle velocity

$$\vec{W}(\vec{x}, t + \frac{h}{2}) = \vec{W}^{(0)}(\vec{x}, t + \frac{h}{2}) + \frac{h}{2M}\sum_{\vec{x}'}\Delta x^3 \rho(\vec{x}')\tilde{\theta}_{\xi}^{(k)}(\vec{x}' - \vec{R}(t))\vec{\Lambda}(\vec{x}') \tag{35}$$

so that the following expression holds

$$\tilde{\theta}_{\xi}^{(k)}(\vec{x} - \vec{R})\vec{\Lambda}(\vec{x}) + \frac{1}{M}\sum_{\vec{x}'}\Delta x^3 \rho(\vec{x}')\tilde{\theta}_{\xi}^{(k)}(\vec{x}' - \vec{R})\vec{\Lambda}(\vec{x}') = \frac{2}{hM}(\vec{u}^{(0)}(\vec{x}) - \vec{W}^{(0)}) \tag{36}$$

This is a linear system for the unknown $\vec{\Lambda}$ whose solution is obtained, for example, with few Jacobi iterations. Finally, the obtained value for $\vec{\Lambda}$ is used to correct the fluid velocity and the particle translational and rotational velocities. By employing a simple Euler update for the angular velocity, after each update of the particle position and once the value of $\vec{\Lambda}$ is determined, the angular velocity is updated as

$$\vec{\omega}(t + h) = \vec{\omega}(t) + \frac{h}{I}\sum_{\vec{x}}\rho(\vec{x})\tilde{\theta}_{\xi}^{(k)}(\vec{x} - \vec{R}(t + h))\vec{\Lambda}(\vec{x}) \times (\vec{x} - \vec{R}(t + h)) \tag{37}$$

The SPM can be validated at several levels, by looking at the flow field aroung a single-particle or looking at hydrodynamic forces exerted between pairs of particles. We look here at the first aspect, by showing in Fig. 6 the flow pattern obtained with the SPM and compared with the Stokes solution and demonstrating the high quality of the simulated flow. The match improves at distance $r > 2a$, with a being the effective Stokes radius obtained by an appropriate fitting procedure[11], indicating the overall good quality of the far and intermediate flow, while at short distance the flow is slightly affected by the smooth fluid-particle interface.

The hydrodynamic radius is further used to compare the Stokes frictional force $6\pi\eta a V$ with that drag force directly computed from an independent simulation of a moving particle. In a periodic cubic box of size L^3, the hydrodynamic drag depends on the box size as $\frac{1}{a_L} \equiv \frac{6\pi\eta\phi_z}{F_z^{drag}} = \frac{1}{a_\infty} - \frac{2.84}{L}$, where a_∞ corresponds to the infinite system size ($L \to \infty$)[32] and the fluid volumetric flow rate is $\phi_z = \frac{1}{L^2}\sum_{\vec{x}} u_z$. The frictional force obtained by measuring the frictional resistance and by the Stokes expression differ by less than $4\,\%$. The resulting hydrodynamics radius is now clearly much larger than in the DPM and is $a_\infty = 1.31$.

We next assess the quality of the fluid-particle coupling in transient conditions. In Fig. 7A, we report the temporal decay of the velocity of a particle of mass $M = 100$, prepared with a small initial velocity. After an initial parabolic regime, corresponding to ballistic relaxation for $t \leq t_{ball} \equiv \frac{M}{6\pi\eta a}$, the Stokesian regime sets in for time $t \leq t_{visc} \equiv \frac{a^2}{\nu}$. At much longer times the long-time tail develops the characteristic dependence $\frac{v(t)}{v(0)} = \frac{M}{12}(\pi\nu t)^{-3/2}$. The rotational properties of the single particle is similarly

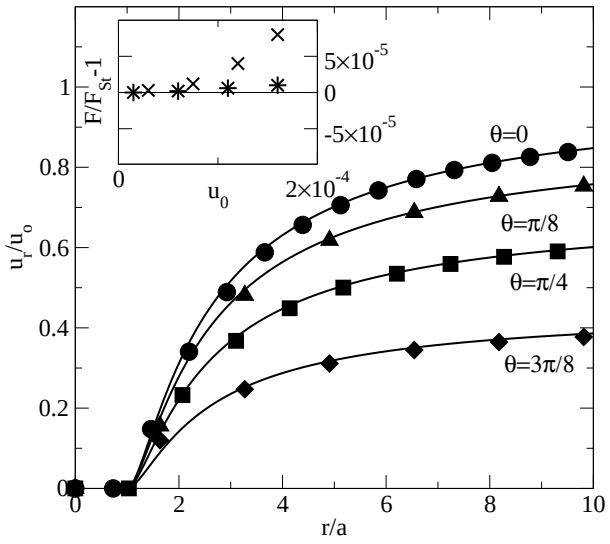
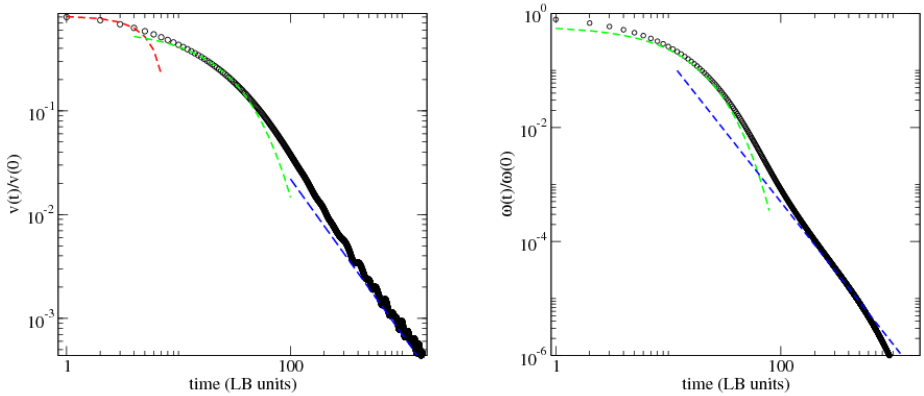


Figure 6. Radial component of the velocity profile for the flow past a spherical particle for $\theta = 0$ (circles), $\pi/8$ (triangles), $\pi/4$ (squares) and $3\pi/8$ rad (diamonds), compared with the Stokes solution at corresponding polar angle θ (solid lines) and for hydrodynamic radius $a = 1.31$. Inset: relative error in the drag force (Eq. 30) computed as $\epsilon \equiv |\vec{F}^{drag}/\vec{F}^{drag}_{St} - 1|$ as a function of the fluid velocity u_0 at $(r, \theta) = (40, 0)$, where $\vec{F}^{drag}_{St}$ is the drag force extrapolated at zero fluid velocity. Forces on fluid are computed according to a first-order (crosses) and second-order (asterisks) accurate LB schemes.

analyzed by looking at the particle with moment of inertia $I = 100$, prepared with a small initial spinning velocity. As shown in Fig. 7B, the ballistic regime rapidly disappears in favor of the Stokes regime that eventually subsides into the long-time tail rotational motion. Both regimes are quantitatively reproduced, showing that the bare moment of inertia correctly describes the rotational dynamics of an isolated body.

5 Excluded Volume Interactions

Being blood a dense suspension of particles (RBCs, white blood cells, platelets, etc.), we need to account for the direct repulsion forces exerted between pairs of globules. By considering the ellipsoidal shape of globules, we can account for body-body excluded volume interactions by soft-core forces and torques represented by the Gay-Berne (GB) potential[33]. The GB potential inhibits interpenetration of pairs of RBCs by introducing an orientation-dependent repulsive interaction derived from the Lennard-Jones potential ($\phi_{LJ}(R_{ij}) = \epsilon[(\sigma/R_{ij})^{12} - (\sigma/R_{ij})^{6}]$, where ϵ is the energy scale and σ the "contact" length scale). The GB potential extends the spherically symmetric Lennard-Jones potential



A B

Figure 7. Panel A: normalized velocity of an impulsively started particle in a quiescent fluid. The red curve illustrates the ballistic regime $\frac{V(t)}{V(0)} \propto 1 - At^2$ while the green curve the Stokes regime $\frac{V(t)}{V(0)} \propto exp(-6\pi\eta a t/M)$. The blue curve is the long-time tail curve for a no-slip sphere $\frac{V(t)}{V(0)} = \frac{M}{12}(\pi\nu t)^{-3/2}$. Panel B: normalized angular velocity of an impulsively spinning particle in a quiescent fluid. The green curve shows the Stokes regime $\frac{\omega(t)}{\omega(0)} \propto exp(-8\pi\eta a^3 t/I)$. The blue curve is the long-time tail curve for a no-slip sphere $\frac{\omega(t)}{\omega(0)} = \frac{\pi I}{\rho}(4\pi\nu t)^{-5/2}$.

to ellipsoidal symmetry, where the potential depends if two RBCs have mutual orientation as face-to-face (maximal repulsion), side-to-side (minimal repulsion) or an arbitrary orientation between the two bodies (intermediate case).

The GB potential is particularly useful to simulation large-scale blood circulation as it can handle interactions between particles of different eccentricity, such as a mixture of ellipsoidal and spherical particles. This flexibility allows to simulate generic biofluids composed of particles with different shapes by employing the same analytical form of the potential. This is the case to study a complete representation of blood, being a mixtures of red and white blood cells, platelets and so on.

The form of the GB potential is easily found in the literature in different versions, for the sake of completeness we provide here an expression as employed for pairs of particles with different shapes, as derived in Refs. 33, 34. Given the principal axes $(a_{i,1}, a_{i,2}, a_{i,3})$ of the i-th globule, the ellipsoidal shape associated to the excluded volume interactions is constructed according to the shape matrix $S_i = \text{diag}(a_{i,1}, a_{i,2}, a_{i,3})$ and the transformed matrix $\mathbf{A}_i = \mathbf{Q}_i \mathbf{S}_i^2 \mathbf{Q}_i^T$ in the laboratory frame. The pair of particles i, j at distance $\mathbf{R}_{ij}$ experiences a characteristic exclusion distance σ_{ij} that depends on the globule-globule distance, shape and mutual orientation, written as

$$\sigma_{ij} = \frac{1}{\sqrt{\phi_{ij}}} \tag{38}$$

$$\phi_{ij} = \frac{1}{2}\hat{\mathbf{R}}_{ij} \cdot \mathbf{H}_{ij}^{-1} \cdot \hat{\mathbf{R}}_{ij} \tag{39}$$

where the matrix $\mathbf{H}_{ij} = \mathbf{A}_i + \mathbf{A}_j$ has been introduced.

A purely repulsive exclusion potential is given by the pairwise form

$$u_{ij} = \begin{cases} 4\epsilon_0(\rho_{ij}^{-12} - \rho_{ij}^{-6}) + \epsilon_0 & \rho_{ij}^6 \leq 2 \\ 0 & \rho_{ij}^6 > 2 \end{cases} \tag{40}$$

with

$$\rho_{ij} = \frac{R_{ij} - \sigma_{ij} + \sigma_{ij}^{min}}{\sigma_{ij}^{min}} \tag{41}$$

where ϵ_0 is the energy scale and σ_{ij}^{min} is a constant, both parameters being independent on the ellipsoidal mutual orientation and distance. For two identical oblate ellipsoids, σ_{ij}^{min} corresponds to a contact distance of the two particles having face-to-face orientation. In general, by considering the minimum particle dimension $a_i^{min} = \min(a_{i,1}, a_{i,2}, a_{i,3})$ then

$$\sigma_{ij}^{min} = \sqrt{2\left[\left(a_i^{min}\right)^2 + \left(a_j^{min}\right)^2\right]}. \tag{42}$$

6 Conclusions

When simulating biological fluids, such as blood, one needs to take into account the corpuscular nature of the biological suspension, and the fact that one deals with suspended particles of multiple species. If one is interested in the transport properties of the suspension as modulated by the structuring of the suspended bodies or by the morphology of the containers, the employed models should be flexible enough to accommodate the basic physical mechanisms, by keeping the numerical simplicity at a bare minimum. The Diffused Particle Model and the Solid Particle Model discussed in this lecture offer ways to emulate the colloidal or vesicular nature of suspended cells. Such simplicity offers major advantages both in terms of numerical robustness, by minimizing the numerical overhead and, last but not least, by implementing efficient parallel softwares to study large-scale biofluidics on leading-edge supercomputers.

References

1. F. Rybicki, S. Melchionna, D. Mitsouras, A. Coskun, A. Whitmore, M. Steigner, L. Nallamshetty, F. Welt, M. Bernaschi, M. Borkin, J. Sircar, E. Kaxiras, S. Succi, P. Stone, and C. Feldman, *Prediction of coronary artery plaque progression and potential rupture from 320-detector row prospectively ECG-gated single heart beat CT angiography: Lattice Boltzmann evaluation of endothelial shear stress*, Intl. J. Cardiovasc. Imaging, Jan. 2009.
2. P. Bagchi, *Mesoscale simulation of blood flow in small vessels*, Biophys.J., **92**, 1858, 2007.
3. H. Noguchi and G. Gompper, *Shape transitions of fluid vesicles and red blood cells in capillary flows*, Proc. Natl. Acad. Sci. USA, **102**, 14159, 2005.
4. R. M. MacMeccan, J. R. Clausen, G. P. Neitzel, and C. K. Aidun, *Simulating Deformable Particle Suspensions Using a Coupled Lattice-Boltzmann and Finite-Element Method*, J. Fluid Mech., **618**, 13, 2009.

5. C. Sun and L.L. Munn, *Particulate nature of blood determines macroscopic rheology: a 2-d Lattice Boltzmann analysis*, Biophys.J., **88**, 1635, 2005.
6. D. A. Fedosov, B. Caswell, and G. E. Karniadakis, *A Multiscale Red Blood Cell Model with Accurate Mechanics, Rheology, and Dynamics*, Biophys. J., **98**, 2215, 2010.
7. J. Zhang, P.C. Johnson, and A.S. Popel, *An immersed boundary lattice Boltzmann approach to simulate deformable liquid capsules and its application to microscopic blood flows*, Phys. Biol., **4**, 285, 2007.
8. M. M. Dupin, I. Halliday, C. M. Care, L. Alboul, and L. L. Munn, *Modeling the flow of dense suspensions of deformable particles in three dimensions*, Phys. Rev. E, **75**, 066707, 2007.
9. S. Melchionna, *A Model for Red Blood Cells in Simulations of Large-scale Blood Flows*, Macromol. Theory & Sim., **20**, 548, 2011.
10. F. Janoschek, F. Toschi, and J. Harting, *A simplified particulate model for coarse-grained hemodynamics simulations*, Phys. Rev. E, **82**, 056710, 2010.
11. S. Melchionna, *Incorporation of smooth spherical bodies in the Lattice Boltzmann method*, J. Comput. Phys., **230**, no. 10, 3966–3976, May 2011.
12. S. Succi, *The Lattice Boltzmann Equation for Fluid Dynamics and Beyond*, Oxford University Press, USA, 2001.
13. X. Shan, X.-F. Yuan, and H. Chen, *Kinetic Theory Representation of Hydrodynamics: A Way Beyond the Navier Stokes Equation*, J. Fluid Mech., **550**, 413, 2006.
14. R. Benzi, S. Succi, and M. Vergassola, *The lattice Boltzmann equation: theory and applications*, 1992.
15. Z. Guo, C. Zheng, and B. Shi, *Discrete lattice effects on the forcing term in the lattice Boltzmann method*, Phys. Rev. E, **65**, 046308, 2002.
16. J. F. Brady and G. Bossis, *Stokesian Dynamics*, Ann. Rev. Fluid Mech., **20**, 111, 1988.
17. C. S. Peskin, *The Immersed Boundary Method*, Acta Numer., **11**, 479, 2002.
18. L. G. Leal, *Advanced Transport Phenomena: Fluid Mechanics and Convective Transport Processes*, Cambridge University Press, 1 edition, June 2007.
19. F. Rioual, T. Biben, and C. Misbah, *Analytical analysis of a vesicle tumbling under a shear flow*, Phys. Rev. E, **69**, 061914, 2004.
20. B. Duenweg and A. Ladd, *Lattice Boltzmann simulations of soft matter systems*, Adv. Polym. Sci., **221**, 89, 2008.
21. S. R. Keller and R. Skalak, *Motion of a Tank-Treading Ellipsoidal Particle in a Shear Flow*, J. Fluid Mech., **120**, 27, 1982.
22. P. Olla, *Simplified Model for Red Cell Dynamics in Small Blood Vessels*, Phys. Rev. Lett., **82**, 453, 1999.
23. E. N. Lightfoot, *Transport Phenomena and Living Systems: Biomedical Aspects of Momentum and Mass Transport*, John Wiley & Sons Inc, Feb. 1974.
24. P. L. Maffettone and M. Minale, *Equation of change for ellipsoidal drops in viscous flow*, Journal of Non-Newtonian Fluid Mechanics, **78**, 227, 1998.
25. M. Bernaschi, S. Melchionna, S. Succi, M. Fyta, E. Kaxiras, and J.K. Sircar, *MUPHY: A parallel MUlti PHYsics/scale code for high performance bio-fluidic simulations*, Comp. Phys. Comm., **180**, 1495–1502, 2009.

26. M. Bernaschi, M. Fatica, S. Melchionna, S. Succi, and E. Kaxiras, *A flexible high-performance Lattice Boltzmann GPU code for the simulations of fluid flows in complex geometries*, Concurrency and Computation: Practice and Experience, **22**, no. 1, 1–14, 2010.
27. A. Peters, S. Melchionna, E. Kaxiras, J. Latt, J. Sircar, M. Bernaschi, M. Bisson, and S. Succi, *Multiscale Simulation of Cardiovascular flows on the IBM Bluegene/P: Full Heart-Circulation System at Red-Blood Cell Resolution*, in: Proceedings of the 2010 ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis, pp. 1–10, IEEE Computer Society. 2010.
28. P. Ahlrichs and B. Duenweg, *Simulation of a single polymer chain in solution by combining lattice Boltzmann and molecular dynamics*, J. Chem. Phys., **111**, 8225, 1999.
29. A. R. Pries, D. Neuhaus, and P. Gaehtgens, *Blood viscosity in tube flow: dependence on diameter and hematocrit*, Am. J. Physiol. Heart Circ. Physiol., **263**, 1770, 1992.
30. G. Bugliarello and J. Sevilla, *Velocity Distribution and other characteristics of steady and pulsatile blood flow in fine glass tubes*, Biorheol., **7**, 85, 1970.
31. S. Melchionna, M. Bernaschi, S. Succi, E. Kaxiras, F. J. Rybicki, D. Mitsouras, A. U. Coskun, and C. L. Feldman, *Hydrokinetic approach to large-scale cardiovascular blood flow*, Comput. Phys. Comm., **181**, 462–472, 2010.
32. Anthony J. C. Ladd, *Numerical Simulations of Particulate Suspensions Via a Discretized Boltzmann Equation. Part 1. Theoretical Foundation*, J. Fluid Mech., **271**, 285–309, 1994.
33. J. G. Gay and B. J. Berne, *Modification of the overlap potential to mimic a linear site–site potential*, J. Chem. Phys., **74**, 3316, 1981.
34. M. P. Allen and G. Germano, *Expressions for forces and torques in molecular simulations using rigid bodies*, Mol. Phys., **104**, 3225, 2006.

Simulations of Blood Flow on the Cell Scale

Dmitry A. Fedosov

Theoretical and Soft Matter Biophysics
Institute of Complex Systems and Institute for Advanced Simulation
Forschungszentrum Jülich, 52425 Jülich, Germany
E-mail: d.fedosov@fz-juelich.de

Red blood cells (RBCs) in various flows exhibit a rich dynamics due their deformability and govern rheological properties and flow characteristics of human blood. Using a mesoscopic RBC model which incorporates membrane shear elasticity, bending rigidity, and viscosity, we quantitatively predict the behavior of a single RBC in shear flow and the dependence of blood viscosity on shear rate and hematocrit. In shear flow, single RBCs respond by tumbling at low shear rates and tank-treading at high shear rates. In transitioning between these regimes, the membrane exhibits substantial deformation controlled largely by flexural stiffness. In RBC suspension (blood) under shear, not only the tumbling/tank-treading cell dynamics affects blood flow characteristics, but also RBC collective behavior and cell-cell aggregation interactions. RBC aggregation leads to reversible rouleaux structures and a tremendous increase of blood viscosity at low shear rates, and results in the presence of a yield stress. The non-Newtonian behavior of blood is analyzed and related to the suspension's microstructure, deformation and dynamics of single RBCs. The generality of these cell models suggests that they can easily be adapted to tune the properties of a much wider class of complex fluids including capsule and vesicle suspensions.

1 Introduction

Blood is circulated around the entire body performing a number of physiological functions. Its main functions are the transport of oxygen and nutrients to cells of the body, removal of waste products such as carbon dioxide and urea, and circulation of molecules and cells which mediate the organism's defense and immune response and play a fundamental role in the tissue repair process. Abnormal blood flow is often correlated with a broad range of disorders and diseases which include hypertension, anemia, atherosclerosis, malaria, and thrombosis. Understanding the rheological properties and dynamics of blood cells and blood flow is crucial for many biomedical and bioengineering applications. Examples include the development of blood substitutes, the design of blood flow assisting devices, and drug delivery. In addition, understanding of vital blood related processes in health and disease may aid in the development of new effective treatments.

Blood is a physiological fluid that consists of erythrocytes or red blood cells (RBCs), leukocytes or white blood cells (WBCs), thrombocytes or platelets, and plasma containing various molecules and ions. RBCs constitute approximately 45% of the total blood volume, WBCs around 0.7%, and the rest is taken up by blood plasma and its substances. One microliter of blood contains about 5 million RBCs, roughly 5 thousand WBCs, and approximately a quarter million platelets. Due to a high volume fraction of RBCs, the rheological properties of blood are mainly determined by their properties and interactions.

Modern rheometry techniques are able to reliably measure macroscopic properties of cell suspensions, for instance the bulk viscosity of blood[1–3]. At low shear rates the RBCs in whole blood have been observed to aggregate into structures called "rouleaux", which

resemble stacks of coins[1,4,5]. The aggregation process appears to be strongly correlated to the presence of the plasma proteins[4,5]. Experiments with washed RBCs re-suspended in pure saline to which fibrinogen was added progressively[4] showed a tremendous viscosity increase at low deformation rates with respect to fibrinogen concentration. In addition, such suspensions exhibit a yield stress[1,6,7], i.e., a threshold stress for flow to begin.

These experimental advances have not been accompanied by theoretical developments which can yield quantitative predictions of rheological and flow properties of blood. A number of theoretical and numerical analyses have sought to describe cell behavior and deformation in a variety of flows. Examples include models of ellipsoidal cells enclosed by viscoelastic membranes[8,9], numerical models based on shell theory[10–12], and discrete descriptions at a mesoscopic level[13–16]. Mesoscopic modeling of viscoelastic membranes is developing rapidly with a RBC membrane modeled as a network of viscoelastic springs in combination with a membrane flexural stiffness, and constraints on the surface area and volume[13–16]. However, recent theoretical and numerical studies focused mostly on the behavior of a single RBC in various flows[13,8,16]. Several studies have also been performed to simulate a suspension of multiple cells[17–19] in tube flow.

In this chapter, a theoretical analysis will be presented for a membrane network model exhibiting specified macroscopic membrane properties without parameter adjustment. RBC dynamics in shear flow showing tumbling and tank-treading will be studied in detail with a view to delineating the effect of the membrane shear moduli, bending rigidity, external, internal, and membrane viscosities. Comparison with available experiments will demonstrate that the computational model is able to accurately describe realistic RBC dynamics in shear flow. Comparison of the numerical simulations with theoretical predictions[8,9] will reveal discrepancies suggesting that the current theoretical models are only qualitatively accurate due to strong simplifications.

Moreover, we will examine blood rheological properties of modeled RBC suspension. In particular, we will investigate the effect of RBC aggregation on blood viscosity, reversible rouleaux formation, and yield stress in a RBC suspension[20]. In addition, we will establish the connection between the rheology of a cell suspension and its microscopic properties on a single-cell level, such as structure or arrangement, cell viscoelastic properties, and local dynamics. In conclusion, we will focus on the *quantitative* prediction of rheological properties and dynamics of single RBCs and blood flow.

2 Red Blood Cells

A healthy human RBC has a biconcave shape with an average diameter of approximately $7.82\ \mu m$. Fig. 1 shows a schematic of a RBC membrane which consists of a lipid bilayer with an attached cytoskeleton formed by a network of the spectrin proteins linked by short filaments of actin. The lipid bilayer is considered to be a nearly viscous and area preserving membrane[10], while RBC elasticity is attributed to the attached spectrin network, as is the integrity of the entire RBC when subjected to severe deformations in the capillaries as small as $3\ \mu m$. The RBC membrane encloses a viscous cytosol whose viscosity is several times larger than that of blood plasma under physiological conditions. Mechanical and rheological characteristics of RBCs and their dynamics are governed by: membrane elastic and viscous properties, bending resistance, and the viscosities of the external/internal fluids.

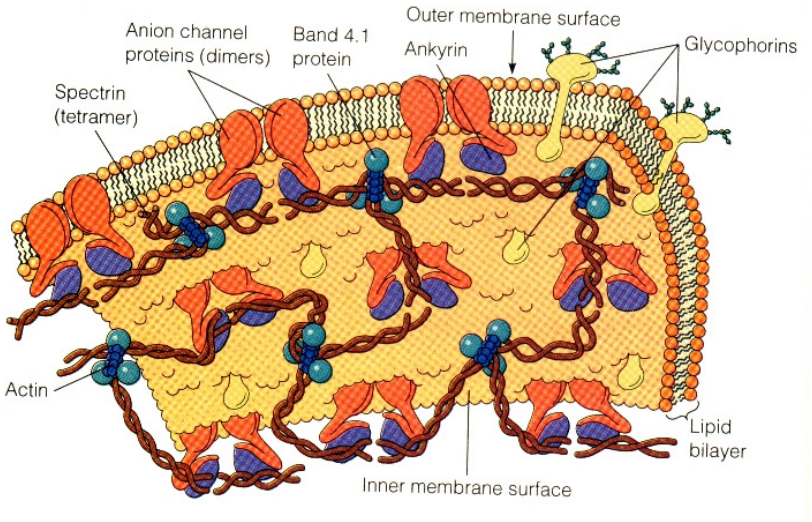


Figure 1. A schematic of the RBC membrane structure.

3 Methods and Models

In the model, the RBC membrane is represented by a viscoelastic network. The motion of the membrane and of the internal and external fluids is described by the method of dissipative particle dynamics (DPD)[21], a mesoscopic particle-based simulation technique, see Appendix for details.

3.1 Red Blood Cell Membrane

The RBC membrane is represented by N_v DPD particles with coordinates $\{\mathbf{x}_{i=1\ldots N_v}\}$ which are vertices of a two-dimensional triangulated network on the RBC surface[22, 16, 23], as shown in Fig. 2. The network has a fixed connectivity with the energy as follows

$$U(\{\mathbf{x}_i\}) = U_s + U_b + U_{a+v}, \tag{1}$$

where U_s is the spring's potential energy, U_b is the bending energy, and U_{a+v} corresponds to the area and volume conservation constraints. The U_s contribution provides membrane elasticity similar to that of a spectrin network of RBC membrane. A "dashpot" is attached to each spring, and therefore, the spring forces are a combination of conservative elastic forces and dissipative forces, which provide network viscous response similar to RBC membrane viscosity. The bending energy mimics bending resistance of the RBC membrane, while the area and volume conservation constraints mimic area-incompressibility of the lipid bilayer and incompressibility of a cytosol, respectively. Below, these energies are described in detail.

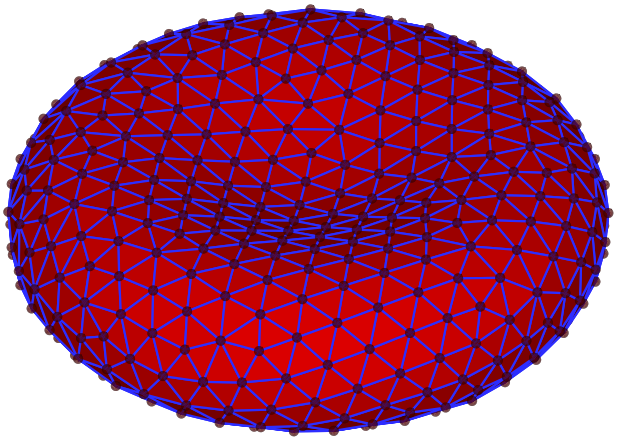


Figure 2. A sketch of a RBC membrane network.

The network nodes are connected by N_s springs with the potential energy as follows

$$U_s = \sum_{j \in 1...N_s} \left[\frac{k_B T l_m (3x_j^2 - 2x_j^3)}{4p(1 - x_j)} + \frac{k_p}{(n-1) l_j^{n-1}} \right], \tag{2}$$

where l_j is the length of the spring j, l_m is the maximum spring extension, $x_j = l_j / l_m$, p is the persistence length, $k_B T$ is the energy unit, k_p is the spring constant, and n is a power. The above equation includes the attractive wormlike chain potential and a repulsive potential for $n > 0$ such that a non-zero equilibrium spring length can be imposed. The performance of different spring models for the RBC membrane was studied in Ref. 23 in detail.

To incorporate the membrane viscosity into the RBC model a dissipative force is introduced for each spring. Following the general framework of the fluid particle model[24] we can define dissipative $\mathbf{F}_{ij}^D$ and random $\mathbf{F}_{ij}^R$ forces for each spring, where $i, j \in 1...N_v$ are a pair of two network vertices connected by a spring. Such forces satisfy the fluctuation-dissipation balance providing consistent temperature of the RBC membrane in equilibrium and are given by

$$\mathbf{F}_{ij}^D = -\gamma^T \mathbf{v}_{ij} - \gamma^C (\mathbf{v}_{ij} \cdot \mathbf{e}_{ij}) \mathbf{e}_{ij}, \tag{3}$$

$$\mathbf{F}_{ij}^R dt = \sqrt{2k_B T} \left(\sqrt{2\gamma^T} d\overline{\mathbf{W}}_{ij}^S + \sqrt{3\gamma^C - \gamma^T} \frac{tr[d\mathbf{W}_{ij}]}{3} \mathbf{1} \right) \cdot \mathbf{e}_{ij}, \tag{4}$$

where γ^T and γ^C are dissipative parameters and the superscripts T and C denote the "translational" and "central" components, $\mathbf{v}_{ij}$ is the relative velocity of spring ends, $tr[d\mathbf{W}_{ij}]$ is the trace of a random matrix of independent Wiener increments $d\mathbf{W}_{ij}$, and

$d\overline{\mathbf{W}_{ij}^S} = d\mathbf{W}_{ij}^S - tr[d\mathbf{W}_{ij}^S]\mathbf{1}/3$ is the traceless symmetric part. Note that the condition $3\gamma^C - \gamma^T \geq 0$ has to be satisfied.

The bending energy of the RBC membrane is given as follows

$$U_b = \sum_{j \in 1 \ldots N_s} k_b \left[1 - cos(\theta_j - \theta_0)\right], \tag{5}$$

where k_b is the bending constant, θ_j is the instantaneous angle between two adjacent triangles having the common edge j, and θ_0 is the spontaneous angle.

In addition, the RBC model includes the area and volume conservation constraints with the corresponding energy given by

$$U_{a+v} = \sum_{j \in 1 \ldots N_t} \frac{k_d (A_j - A_0)^2}{2A_0} + \frac{k_a (A - A_0^{tot})^2}{2A_0^{tot}} + \frac{k_v (V - V_0^{tot})^2}{2V_0^{tot}}, \tag{6}$$

where N_t is the number of triangles in the membrane network, A_0 is the triangle area, and k_d, k_a and k_v are the local area, global area and volume constraint coefficients, respectively. The terms A and V are the total RBC area and volume, while A_0^{tot} and V_0^{tot} are the specified total area and volume, respectively. More details on the RBC model can be found in Refs. 16, 23.

3.2 Membrane Macroscopic Properties

Several parameters must be chosen in the membrane network model to ensure a desired mechanical response. Fig. 3 depicts a network model and its continuum counterpart. To

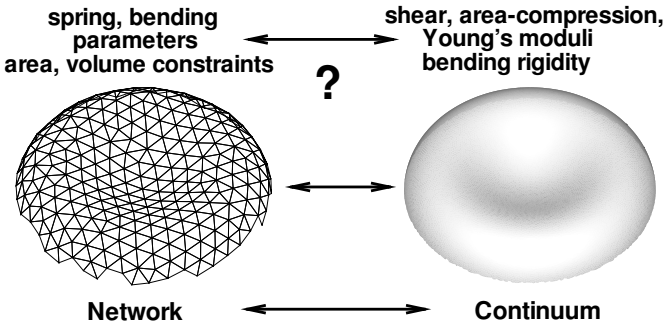


Figure 3. Illustration of a membrane network and corresponding continuum model.

circumvent ad-hoc parameter adjustment, we derive relationships between local model parameters and network macroscopic properties for an elastic hexagonal network. A similar analysis for a two-dimensional particulate sheet of equilateral triangles was presented in Refs. 25, 23.

Fig. 4 illustrates an element in a hexagonal network with vertex **v** placed at the origin of a local Cartesian system. Using the virial theorem[26], we find that the Cauchy stress

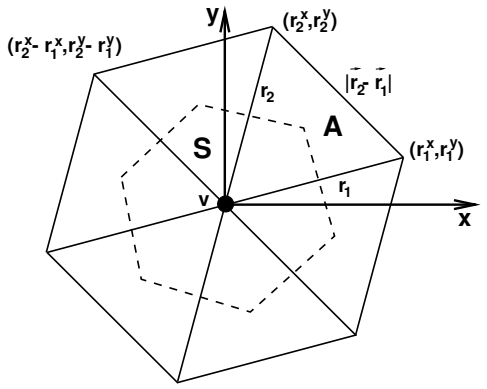


Figure 4. Illustration of an element in a hexagonal triangulation.

tensor at **v** is

$$\tau_{\alpha\beta} = -\frac{1}{S}\left[\frac{f(r_1)}{r_1} r_1^\alpha r_1^\beta + \frac{f(r_2)}{r_2} r_2^\alpha r_2^\beta + \frac{f(|\boldsymbol{r}_2 - \boldsymbol{r}_1|)}{|\boldsymbol{r}_2 - \boldsymbol{r}_1|}(r_2^\alpha - r_1^\alpha)(r_2^\beta - r_1^\beta)\right] - \left(\frac{k_a(A_0^{tot} - N_t A)}{A_0^{tot}} + \frac{k_d(A_0 - A)}{A_0}\right)\delta_{\alpha\beta}, \quad (7)$$

where α and β stand for x or y, $f(r)$ is the spring force, $A_0^{tot} = N_t A_0$, $S = 2A_0$, $\delta_{\alpha\beta}$ is the Kronecker delta, and S is the area of the hexagonal element centered at **v**.

3.2.1 Shear Modulus

The shear modulus is derived from the network deformation by applying a small engineering shear strain γ to the network element shown in Fig. 4. For instance, the deformation of a material vector $\boldsymbol{r}_1$ is then described as

$$\boldsymbol{r}_1' = \boldsymbol{r}_1 \cdot \boldsymbol{J} = \begin{bmatrix} r_1^x + \frac{1}{2} r_1^y \\ \frac{1}{2} r_1^x \gamma + r_1^y \end{bmatrix}, \quad (8)$$

where

$$\boldsymbol{J} = \begin{bmatrix} 1 & \gamma/2 \\ \gamma/2 & 1 \end{bmatrix} + O(\gamma^2) \quad (9)$$

is the linear strain tensor and $\boldsymbol{r}_1 = (r_1^x; r_1^y)$, as shown in Fig. 4. Because the shear deformation is area preserving, only spring forces in Eq. 7 contribute to the membrane shear modulus.

Expanding τ_{xy} in a Taylor series, we find that

$$\tau_{xy}' = \tau_{xy} + \left.\frac{\partial \tau_{xy}'}{\partial \gamma}\right|_{\gamma=0} \gamma + O(\gamma^2). \quad (10)$$

The linear shear modulus of the network is

$$\mu_0 = \left.\frac{\partial \tau'_{xy}}{\partial \gamma}\right|_{\gamma=0}. \tag{11}$$

For example, differentiating the first term of τ_{xy} in Eq. 7 yields

$$\frac{\partial}{\partial \gamma}\left(\frac{f(r'_1)}{r'_1} r_1^{x'} r_1^{y'}\right)_{\gamma=0} = \left(\frac{\partial \frac{f(r_1)}{r_1}}{\partial r_1}\frac{(r_1^x r_1^y)^2}{r_1} + \frac{f(r_1) r_1}{2}\right)_{r_1=l_0}, \tag{12}$$

where l_0 is the equilibrium spring length. Using the vector-product definition of the area of a triangle, we obtain

$$(r_1^x r_1^y)^2 + (r_2^x r_2^y)^2 + (r_2^x - r_1^x)^2 (r_2^y - r_1^y)^2 = 2A_0^2. \tag{13}$$

The linear shear modulus of the network model is

$$\mu_0 = \frac{\sqrt{3} k_B T}{4 p l_m x_0}\left(\frac{x_0}{2(1-x_0)^3} - \frac{1}{4(1-x_0)^2} + \frac{1}{4}\right) + \frac{\sqrt{3} k_p (n+1)}{4 l_0^{n+1}}, \tag{14}$$

where $x_0 = l_0/l_m$.

3.2.2 Area Compression and Young's Moduli

The linear elastic area compression modulus K is found from the in-plane pressure following a small area expansion as

$$p = -\frac{1}{2}(\tau_{xx} + \tau_{yy}) = \frac{3\,l}{4A} f(l) + \frac{(k_a + k_d)(A_0 - A)}{A_0}. \tag{15}$$

Defining the compression modulus as

$$K = -\left.\frac{\partial p}{\partial \log A}\right|_{A=A_0} = -\frac{1}{2}\left.\frac{\partial p}{\partial \log l}\right|_{l=l_0} = -\frac{1}{2}\left.\frac{\partial p}{\partial \log x}\right|_{x=x_0}, \tag{16}$$

and using Eqs. 15 and 16, we obtain

$$K = 2\mu_0 + k_a + k_d. \tag{17}$$

For the nearly constant-area membrane enclosing a red blood cell, the compression modulus is much larger than the shear elastic modulus μ_0.

The Young's modulus of the two-dimensional sheet is given by

$$Y = \frac{4K\mu_0}{K + \mu_0}. \tag{18}$$

As $K \to \infty$, we obtain $Y \to 4\mu_0$. To ensure a nearly constant area, we set $k_a + k_d \gg \mu_0$.

3.2.3 Bending Rigidity

Helfrich[27] proposed an expression for the bending energy of a lipid membrane,

$$E_c = \frac{k_c}{2} \iint (C_1 + C_2 - 2C_0)^2 \, dA + k_g \iint C_1 C_2 \, dA, \tag{19}$$

where C_1 and C_2 are the principal curvatures, C_0 is the spontaneous curvature, and k_c, k_g are bending rigidities. The second term on the right-hand side of Eq. 19 is constant for any closed surface.

A relationship between the bending constant, k_b, and the macroscopic membrane bending rigidity, k_c, can be derived for a spherical shell. Fig. 5 shows two equilateral triangles with edge length l_0 whose vertices lie on a sphere of radius R. The angle between the tri-

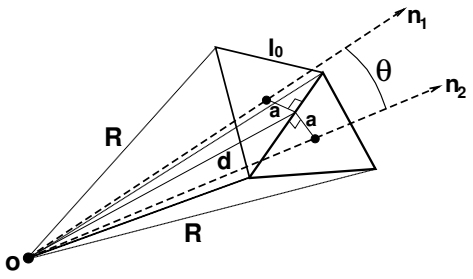


Figure 5. Illustration of two equilateral triangles on the surface of a sphere of radius R.

angle normals $\mathbf{n}_1$ and $\mathbf{n}_2$ is denoted by θ. In the case of a spherical shell, the total energy in Eq. 19 is found to be

$$E_c = 8\pi k_c \left(1 - \frac{C_0}{C_1}\right)^2 + 4\pi k_g = 8\pi k_c \left(1 - \frac{R}{R_0}\right)^2 + 4\pi k_g, \tag{20}$$

where $C_1 = C_2 = 1/R$ and $C_0 = 1/R_0$. In the network model, the bending energy of the triangulated sphere is

$$U_b = N_s \, k_b \, [1 - \cos(\theta - \theta_0)]. \tag{21}$$

Expanding $\cos(\theta - \theta_0)$ in a Taylor series around $\theta - \theta_0$ provides us with the leading term

$$U_b = \frac{1}{2} N_s k_b (\theta - \theta_0)^2 + O\left((\theta - \theta_0)^4\right). \tag{22}$$

With reference to Fig. 5, we find that $2a \approx \theta R$ or $\theta = l_0/(\sqrt{3}R)$, and $\theta_0 = l_0/(\sqrt{3}R_0)$.

For a sphere, $A = 4\pi R^2 \approx N_t A_0 = \sqrt{3} N_t l_0^2/4 = \sqrt{3} N_s l_0^2/6$, and $l_0^2/R^2 = 8\pi\sqrt{3}/N_s$. Finally, we obtain

$$U_b = \frac{1}{2} N_s k_b \left(\frac{l_0}{\sqrt{3}R} - \frac{l_0}{\sqrt{3}R_0}\right)^2 = \frac{N_s k_b l_0^2}{6R^2} \left(1 - \frac{R}{R_0}\right)^2 = \frac{4\pi k_b}{\sqrt{3}} \left(1 - \frac{R}{R_0}\right)^2. \tag{23}$$

Equating the macroscopic bending energy E_c to U_b for $k_g = -4k_c/3$ and $C_0 = 0$, we obtain $k_b = 2k_c/\sqrt{3}$ in agreement with the limit of a continuum approximation[28].

The spontaneous angle θ_0 is set according to the total number of vertices on the sphere, N_v. It can be shown that $\cos\theta = 1 - 1/[6(R^2/l_0^2 - 1/4)]$ and the number of sides is $N_s = 2N_v - 4$. The bending coefficient, k_b, and spontaneous angle, θ_0, are given by

$$k_b = \frac{2}{\sqrt{3}}\,k_c, \qquad \theta_0 = \arccos\left(\frac{\sqrt{3}(N_v-2)-5\pi}{\sqrt{3}(N_v-2)-3\pi}\right). \tag{24}$$

3.2.4 Membrane Viscosity

Since interparticle dissipative interaction is an intrinsic part of the DPD formulation, incorporating dissipative and random forces into springs fits naturally into the DPD scheme. The general framework of the fluid-particle model[24] provides us with Eqs. 3 and 4. These dissipative and random forces in combination with an elastic spring constitute a mesoscopic viscoelastic spring. To relate the membrane shear viscosity, η_m, to the model dissipative parameters γ^T and γ^C, an element of the hexagonal network shown in Fig. 4 is subjected to a constant shear rate, $\dot{\gamma}$. The shear stress τ_{xy} at short times can be approximated from the contribution of the dissipative force in Eq. 3,

$$\tau_{xy} = -\frac{1}{2A_0}\Big[\gamma^T\dot{\gamma}\left((r_y^1)^2+(r_y^2)^2+(r_y^2-r_y^1)^2\right)+\frac{\gamma^C\dot{\gamma}}{l_0^2}\Big(\,(r_x^1r_y^1)^2+(r_x^2r_y^2)^2 \\ +(r_x^2-r_x^1)^2(r_y^2-r_y^1)^2\,\Big)\Big] = \dot{\gamma}\,\sqrt{3}\,(\gamma^T+\frac{1}{4}\,\gamma^C\,). \tag{25}$$

The membrane viscosity is given by

$$\eta_m = \frac{\tau_{xy}}{\dot{\gamma}} = \sqrt{3}\,\left(\gamma^T+\frac{1}{4}\,\gamma^C\right). \tag{26}$$

This equation indicates that γ^T accounts for the largest portion of the membrane dissipation. Therefore, γ^C is set to its minimum value, $\frac{1}{3}\,\gamma^T$, in the simulations.

3.3 Membrane-Solvent Interfacial Conditions

The cell membrane encloses a viscous fluid and is surrounded by a liquid solvent. Fig. 6 shows a snapshot of a simulation in equilibrium, where red particles are membrane vertices, blue particles represent the external fluid, and green particles represent the internal fluid. To prevent mixing of the internal and external fluids, we require impenetrability. We also enforce no-slip boundary conditions at the membrane implemented by pairwise interactions between fluid particles and membrane nodes. Bounce-back reflection of fluid particles at the triangular plaquettes satisfies membrane impenetrability and better enforces no-slip compared to specular reflection. However, bounce-back reflection alone does not guarantee no-slip. In practice, it is necessary to properly set the DPD dissipative interactions between fluid particles and membrane vertices.

The continuum linear shear flow over a flat plate is used to determine the dissipative force coefficient γ for the fluid-membrane coupling. For the continuum, the total shear force on area A of the plate is $A\eta_0\dot{\gamma}$, where η_0 is the fluid viscosity and $\dot{\gamma}$ is the local

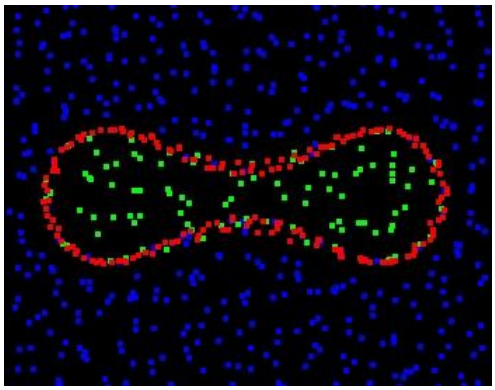


Figure 6. A slice through a sample equilibrium simulation. Red particles are membrane vertices, blue particles represent the external fluid, and green particles represent the internal fluid.

shear-rate. To mimic the membrane surface, wall particles are distributed over the plate to match the configuration of the cell network model. The force on a single wall particle in this system exerted by the surrounding fluid under shear can be expressed as

$$F_v = \iiint_{V_h} n\, g(r)\, F^D\, dV, \tag{27}$$

where F^D is the DPD dissipative force between fluid and wall particles, n is the fluid number density, $g(r)$ is the radial distribution function of fluid particles relative to the wall particles, and V_h is the half-sphere volume of fluid above the plate. Thus, the total shear force on the area A is equal to $N_A F_v$, where N_A is the number of plate particles residing in the area A. When conservative interactions between fluid particles and the membrane vertices are neglected, the radial distribution function simplifies to $g(r) = 1$. Setting $N_A F_v = A\eta_0\dot{\gamma}$ yields an expression for the dissipative force coefficient γ in terms of the fluid density and viscosity and the wall density, N_A/A. Near a wall where the half-sphere lies within the range of the linear wall shear flow, the shear rate cancels out. This formulation has been verified to enforce satisfactory no-slip boundary conditions for shear flow over a flat plate, and is an excellent approximation for no-slip at the membrane surface.

3.4 RBC Aggregation Interactions

For blood, the attractive cell-cell interactions are crucial for simulation of RBC aggregation into rouleaux. These forces are approximated phenomenologically with a Morse potential,

$$U_M(r) = D_e \left[e^{2\beta(r_0 - r)} - 2e^{\beta(r_0 - r)} \right], \tag{28}$$

where r is the separation distance, r_0 is the zero force distance, D_e is the well depth of the potential, and β characterizes the interaction range. The Morse potential interactions are implemented between every two vertices of separate RBCs if they are within a defined potential cutoff radius r_d. Even though the Morse potential in Eq. 28 contains a short-range repulsive force when $r < r_0$, such repulsive interactions cannot prevent two RBCs

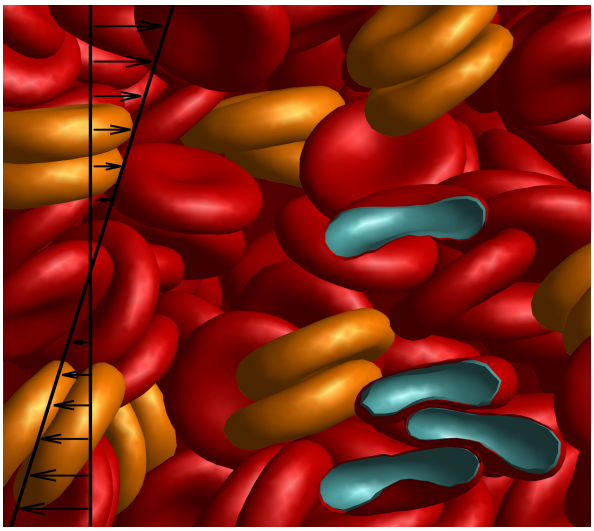


Figure 7. Simulation of whole blood under shear flow. RBCs are shown in red and in orange, where orange color depicts the rouleaux structures formed due to aggregation interactions between RBCs. The image also displays several cut RBCs with the inside drawn in cyan to illustrate RBC shape and deformability.

from an overlap. To guarantee no overlap among RBCs we employ a short range Lennard-Jones potential and specular reflections of RBC vertices on membranes of other RBCs. The specular reflections of RBC vertices on surfaces of other RBCs are necessary due to coarseness of the triangular network which represents the RBC membrane.

4 Simulation Results and Discussion

We present simulation results for the behavior of a single RBC in shear flow and discuss the effect of various membrane properties on RBC dynamics. We also study dense RBC suspension (blood) under shear and examine the blood viscosity with and without RBC aggregation, rouleaux formation, and yield stress. Finally, we establish a link between bulk blood properties, microstructure, and the flow behavior of single RBCs.

4.1 Simulation Setup and Parameters

A single RBC or the RBC suspension were subjected to linear shear flow with periodic Lees-Edwards boundary conditions[29] as shown in Fig. 7. The computational domain had the size of $5.6D_0 \times 4.0D_0 \times 3.4D_0$, where D_0 is the RBC diameter which is equal to about $7.82\ \mu m$ for a healthy RBC. In case of the RBC suspension, 168 RBCs and 117599 solvent particles were placed in the computational domain. The RBC membrane Young's modulus was set to $Y_0 = 269924\ k_BT/D_0^2$, which corresponds to $Y_0 = 18.9\ \mu N/m$ at physiological temperature of $T = 37^o\ C$. The RBC bending rigidity was assumed to be $k_c = 3\times10^{-19} J$, which is equal to approximately $70k_BT$ at $T = 37^o\ C$. The corresponding Föppl-von Kármán number $0.25Y_0D_0^2/k_c$ is therefore equal to approximately 963. The membrane

viscosity was set to be $12\eta_0$, where η_0 is the suspending fluid viscosity. The coefficients for the area and volume constraints were set large enough in order to closely approximate membrane and cytosol incompressibility. Coupling between the solvent and RBCs was performed through a dissipative force between fluid particles and membrane vertices.

Interactions between different RBCs included the short range repulsive Lennard-Jones potential with parameters $\epsilon = 10.0\ k_BT$ and $\sigma_{LJ} = 0.037\ D_0$. These repulsive interactions result in a thin layer next to a RBC membrane which cannot be accessed by other cells. This layer can be interpreted as a slight increase of the RBC volume. Therefore, the RBC volume was assumed to be about 10% larger than that of the triangulated network. The concentration of RBCs is called hematocrit and denoted as H_t. RBC aggregation interactions were mediated by the Morse potential with parameters $D_e = 3.0\ k_BT$, $r_0 = \sigma_{LJ}$, $\beta = 0.45\ \sigma_{LJ}^{-1}$, and $r_d = 3.7\ \sigma_{LJ}$. For more details see Ref. 20.

4.2 Single RBC in Shear Flow

Experimental observations have shown that RBCs tumble at low shear rates and exhibit a tank-treading motion at high shear rates[30–32,8]. Fischer[31] attributed this behavior to a minimum elastic energy state of the cell membrane. Cells can be made to tank-tread in the laboratory for several hours. When the flow is stopped, the cells relax to the original biconcave shape where attached microbeads recover their original relative position. It appears that tank-treading is possible only when a certain elastic energy barrier has been surpassed. Theoretical analyses have considered ellipsoidal cell models tank-treading along a fixed ellipsoidal path[8,9]. Our simulations show that the dynamics depends on the membrane shear modulus, shear rate, and viscosity ratio $\lambda = (\eta_i + \eta_m)/\eta_o$, where η_i, η_m, and η_o are the interior, membrane, and outer fluid viscosities.

For viscosity ratio $\lambda < 3$, the theory predicts tumbling at low shear rates and tank-treading motion at high shear rates[9]. The cells exhibit an unstable behavior in a narrow intermittent region around the tumbling-to-tank-treading transition where tumbling can be followed by tank-treading and *vice versa*. For $\lambda > 3$, stable tank-treading does not necessarily arise. RBCs with viscosity ratio $\lambda > 3$ have been observed to tank-tread while exhibiting a swinging motion with a certain frequency and amplitude about an average tank-treading axis. The reliability of the theoretical predictions will be judged by comparison with the results of our simulations.

A RBC is suspended in a linear shear flow. The viscosities of the external solvent and internal cytosol fluid are set to $\eta_o = \eta_i = 0.005\ Pa \cdot s$, while the membrane viscosity is set to $\eta_m = 0.022\ Pa \cdot s$. Fig. 8 presents information on the cell tumbling and tank-treading frequencies under different conditions. Experimental observations by Tran-Son-Tay et al.[30] and Fischer[32] are included for comparison. In the case of a purely elastic membrane with or without inner solvent (circles and squares), the numerical results significantly overpredict the tank-treading frequency compared with experimental measurements. The internal solvent viscosity could be further increased to improve agreement with experimental data. However, since the cytosol is a hemoglobin solution with a well-defined viscosity of about $0.005\ Pa \cdot s$[33], excess viscous dissipation must occur inside the membrane. The data plotted with triangles in Fig. 8 show good agreement with experimental data for increased membrane viscosity.

The tumbling frequency is nearly independent of the medium viscosities. Increasing the viscosity of the internal fluid or raising the membrane viscosity slightly shifts

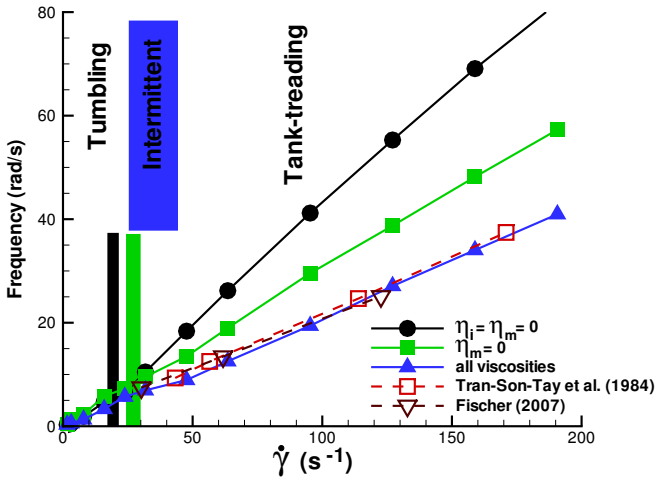


Figure 8. Tumbling and tank-treading frequency of a RBC in shear flow for $\eta_o = 0.005\ Pa \cdot s$, $\eta_i = \eta_m = 0$ (circles); $\eta_o = \eta_i = 0.005\ Pa \cdot s$, $\eta_m = 0$ (squares); $\eta_o = \eta_i = 0.005\ Pa \cdot s$, $\eta_m = 0.022\ Pa \cdot s$ (triangles).

the tumbling-to-treading threshold into higher shear rates through an intermittent regime. We estimate that the tank-treading energy barrier of a cell is approximately $E_c = 3$ to $3.5 \times 10^{-17}\ J$. In the theoretical model[9], the energy barrier was set to $E_c = 10^{-17}\ J$ to ensure agreement with experimental data. Membrane deformation during tank treading is indicated by an increase in the elastic energy difference with increasing shear rate to within about 20% of E_c.

An intermittent regime is observed with respect to the shear rate in all cases. Consistent with the experiments, the width of the transition zone broadens as the membrane viscosity increases. Similar results regarding intermittency were reported by Kessler et al.[34] for viscoelastic vesicles. We conclude that theoretical predictions of cell dynamics in shear flow are qualitatively correct at best due to the assumption of ellipsoidal shape and fixed ellipsoidal tank-treading path. Experiments[8] have shown and the present simulations have confirmed that the cell deforms along the tank-treading axis with strains of order $0.1-0.15$.

We have seen that a cell oscillates or swings around tank-treading axes with a certain frequency and amplitude. Fig. 9 presents graphs of the average tank-treading angle and swinging amplitude. The numerical results are consistent with experimental data in Ref. 8. The average swinging angle is larger for a purely elastic membrane without inner cytosol. The inclination angle is independent of the internal fluid and membrane viscosities and the swinging amplitude is insensitive to the fluid and membrane properties. The swinging frequency is exactly twice the tank-treading frequency.

4.3 Blood Viscosity

Blood viscosity was computed, with and without aggregation, as a function of the shear rate $\dot{\gamma}$ over the range $0.005s^{-1}$ to $1000.0s^{-1}$ in plane Couette flow. The shear rate and

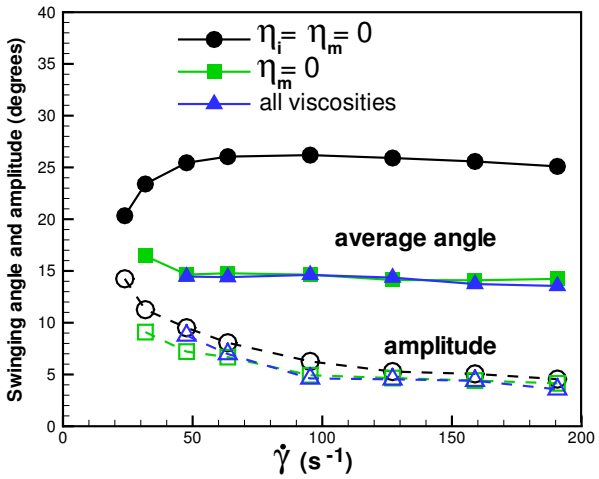


Figure 9. Graphs of the swinging average angle in degrees (filled symbols) and amplitude (open symbols) for (*a*) $\eta_o = 0.005\ Pa \cdot s$ and $\eta_i = \eta_m = 0$ (circles); (*b*) $\eta_o = \eta_i = 0.005\ Pa \cdot s$ and $\eta_m = 0$ (squares); (*c*) $\eta_o = \eta_i = 0.005\ Pa \cdot s$ and $\eta_m = 0.022\ Pa \cdot s$ (triangles).

the cell density in our simulations were verified to be spatially uniform. Fig. 10 shows the relative viscosity (RBC suspension viscosity normalized by η_0) against shear rate $\dot{\gamma}$ at hematocrit $H_t = 0.45$. The blood model predictions are in excellent agreement with the blood viscosities measured in three different laboratories[1–3]. The blood model, consisting only of RBCs in suspension, clearly captures the effect of aggregation on the viscosity at low shear rates, and suggests that cells and molecules other than RBCs have little effect on the viscosity, at least under healthy conditions. At intermediate shear rates, where aggregation is no longer relevant, shear thinning is due to a transition from tumbling to tank-treading motion, accompanied by strong cell deformations[20].

4.4 Reversible Rouleaux Formation

The formation of rouleaux in blood occurs in equilibrium and at sufficiently small shear rates, while large shear rates result in immediate dispersion of gentle RBC structures. Experimentally, aggregation is observed[1] to be a two-step process: the formation of short linear stacks with few RBCs, followed by their coalescence into long linear and branched rouleaux. As the shear rate increases the large rouleaux break up into smaller ones, and at higher values the suspension ultimately becomes one of mono-dispersed RBCs[35]. This process then reverses as the shear rate is decreased. This typical formation-destruction behavior of rouleaux is consistent with the results of our simulations as shown in Fig. 11. At low shear rates (left plot), the initially dispersed RBCs aggregate into large rouleaux of up to about 20 RBCs; as the shear rate is increased to moderate values (middle plot), these structures are reduced in size until at high rates (right plot) they are dispersed almost completely into individual RBCs. Reversibility is demonstrated by reduction of the shear rate to the formation value at which point individual RBCs begin to re-aggregate.

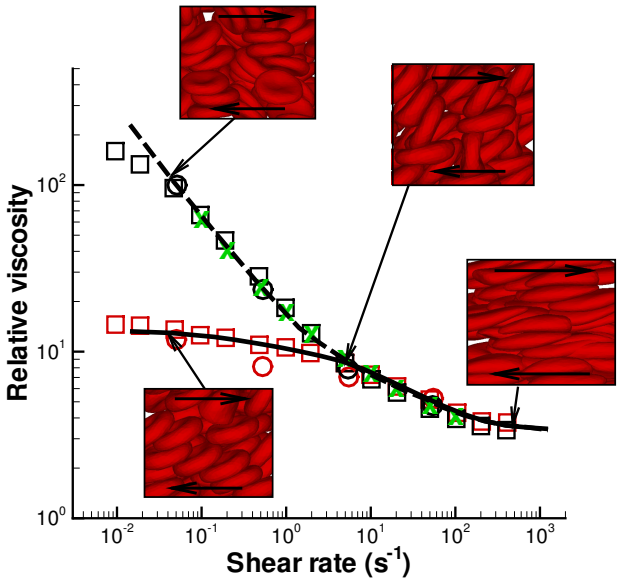


Figure 10. Validation of simulation results for whole blood and non-aggregating RBC suspension. Plot of non-Newtonian relative viscosity (the cell suspension viscosity normalized by η_0) as a function of shear rate at $H_t = 0.45$ and 37^oC: *simulated* curves are in black, and *experimental* points: Whole blood: green crosses - Merril et al.[1]; black circles - Chien et al.[2], black squares - Skalak et al.[3]. Non-aggregating RBC suspension: red circles - Chien et al.[2]; red squares - Skalak et al.[3].

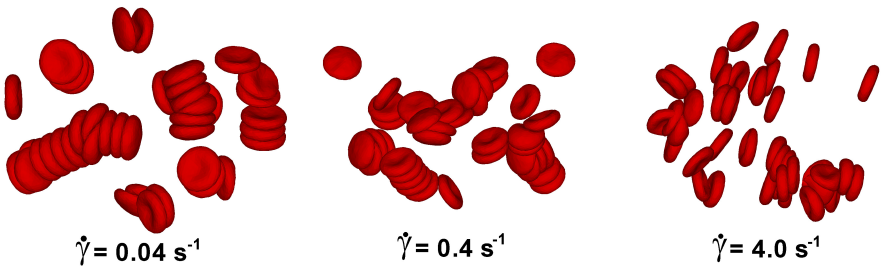


Figure 11. Visualization of aggregation. Simulated reversible rouleaux are formed by RBCs at $H_t = 0.1$. The left plot corresponds to low shear rates, middle plot to moderate share rates, and right plot to high shear rates as indicated with the shear rate values.

4.5 Yield Stress and Aggregation

Whole blood is believed to exhibit a yield stress, i.e. a threshold stress for flow to begin[1,6,7], which is often estimated by the extrapolation of measured shear stress to the zero shear rate on the basis of the Casson's equation[36],

$$\tau_{xy}^{1/2} = \tau_y^{1/2} + \eta^{1/2}\dot{\gamma}^{1/2}, \tag{29}$$

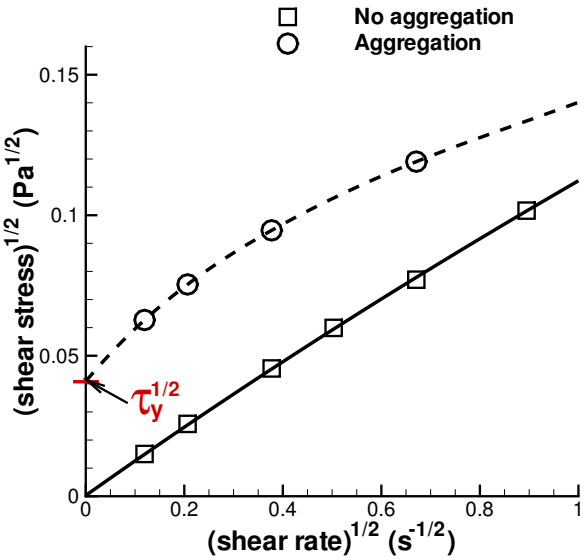


Figure 12. Correlation of aggregation with yield stress. Casson plots with a polynomial fit showing the extrapolated intercept τ_y for simulated suspensions with, dashed lines, and without aggregation, solid lines, at $H_t = 0.45$.

where τ_y is a yield stress and η is the suspension viscosity at large $\dot{\gamma}$. The assumptions of Casson's relation are likely to hold at least at low shear rates, which was successfully demonstrated for pigment-oil suspensions[36], Chinese ovary hamster cell suspensions[37], and blood[4]. Fig. 12 is a polynomial fit in Casson coordinates ($\dot{\gamma}^{1/2}$,$\tau_{xy}^{1/2}$) to the simulated data for a $H_t = 0.45$ suspension, which shows clearly that τ_y is non-zero for the aggregating RBC suspension, while τ_y is absent without cell aggregation. The yield stress for blood has previously been attributed to the presence of rouleaux in experiments reported in Refs. 1, 6, 7. Merrill et al.[1] found τ_y of healthy human blood to lie between 0.0015 and $0.005\ Pa$ at $H_t = 0.45$. Our simulation results in Fig. 12 fall into this range of the yield stress of whole blood.

4.6 Micro-to-Macro Link

The non-Newtonian nature of blood (e.g., shear thinning, yield stress) emerges from the interactions between cells and from their properties and dynamics. Therefore, we examined the structure and dynamics of the modeled suspensions on the level of single cells. We found null pair-correlations of RBC *centers of mass* for each direction (x, y, z), which indicates that the cell suspensions do not self-assemble or order themselves in any direction at $H = 45\%$. To examine the cell suspension's local microstructure, we calculate the radial distribution function (RDF) of RBC centers shown in Fig. 13(a). For the no-aggregation case, we find that no significant structures formed over the entire range of shear rates. At the lowest shear rate (red solid line) several small peaks in RDF indicate the presence of

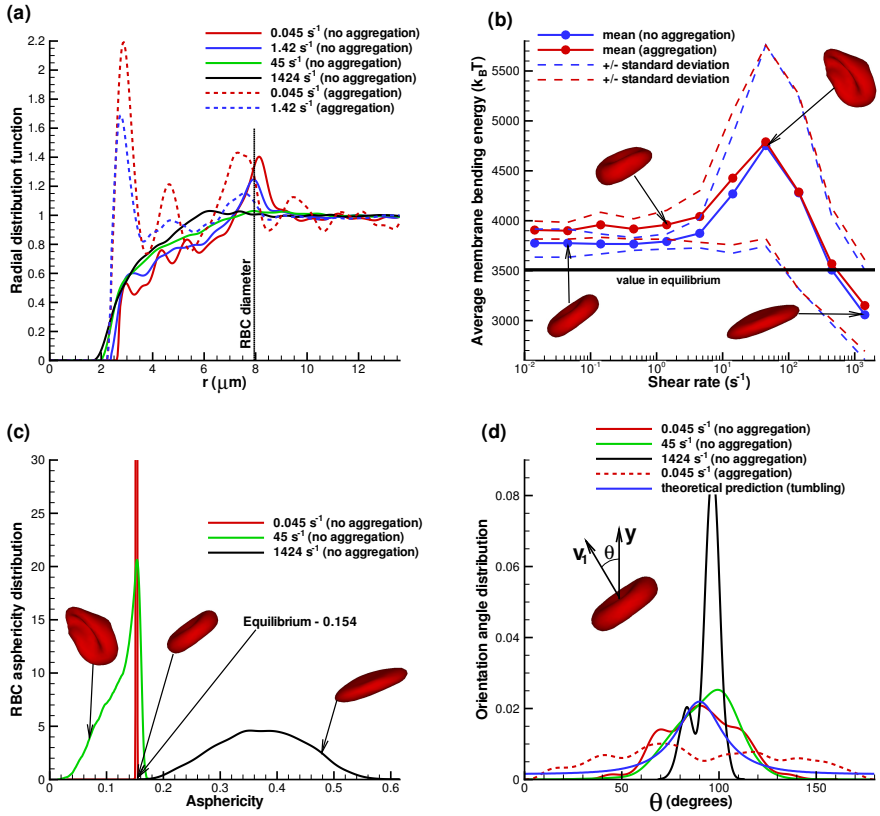


Figure 13. Structural and dynamical properties of RBC suspensions with $H = 45\%$. Snapshots show sample RBC conformations from simulations. (a) Radial distribution function showing cell suspension's structure. (b) Average membrane bending energy with respect to shear rate showing correlation between single cell deformation and dynamics. Dashed lines are the corresponding mean values plus/minus one standard deviation. (c) RBC asphericity distributions characterizing the deviation from a spherical shape as a function of shear rate. The asphericity is defined as $[(\lambda_1 - \lambda_2)^2 + (\lambda_2 - \lambda_3)^2 + (\lambda_3 - \lambda_1)^2]/(2R_g^4)$, where $\lambda_1 \leq \lambda_2 \leq \lambda_3$ are the eigenvalues of the gyration tensor and $R_g^2 = \lambda_1 + \lambda_2 + \lambda_3$. The asphericity for a single RBC in equilibrium is equal to 0.154. (d) Orientational angle distributions for various shear rates which illustrate single cell dynamics. The cell orientational angle is given by the angle between the eigenvector V_1 of the gyration tensor and the flow-gradient direction (y). Theoretical prediction showing the orientational angle distribution of a single tumbling RBC in shear flow is calculated from the theory in Ref. 8.

infrequent intermediate structures, since RBCs may have enough time to relax locally at very low shear rates. A larger peak of the red solid curve at $r = 8\mu m$, which is equal to the cell diameter, indicates that neighboring RBCs are often aligned with each other in the flow. As seen from the other solid curves (blue, green, and black), the correlations completely disappear at higher shear rates, and therefore the shear thinning behavior of a non-aggregating suspension is clearly not due to a change in microstructure. In contrast, several large peaks in the RDF function for the aggregating case at the lowest shear rate $\dot{\gamma} = 0.045\ s^{-1}$ (red dashed line) indicate the formation of rouleaux of 2 to 4 RBCs. In-

crease of the shear rate leads to the dispersion of rouleaux shown by the blue dashed curve in Fig. 13(a), where predominant RBC aggregates are formed by only two RBCs. At shear rates above approximately $5 - 10\ s^{-1}$ no difference in microstructure is detected between aggregating and non-aggregating cell suspensions. As a conclusion, the steep increase in viscosity of the aggregating blood at low shear rates is mainly due to the cell aggregation into rouleaux. In addition, rouleaux formation also provides a plausible explanation for the existence of yield stress, since with decrease of shear rate larger rouleaux structures are formed resulting in an eventual "solidification" of the suspension.

The dynamics of a single RBC in shear flow is characterized by the tumbling motion at low shear rates and membrane tank-treading at high shear rates[8, 15, 16]. The tumbling-to-tank-treading transition occurs within a certain range of intermediate shear rates, where a RBC may experience high bending deformations[16]. The deformation, orientation, and dynamics of cells within the suspension is illustrated in Figs. 13(b), (c), and (d). These plots show that cells in the suspension mostly tumble and retain their biconcave shape at low shear rates below $5\ s^{-1}$, which is confirmed by essentially no change in RBC bending energy and in its standard deviation (Fig. 13(b)), by the extremely narrow asphericity distribution around the equilibrium value of 0.154 (Fig. 13(c)), and by the wide orientational angle (θ) distribution in Fig. 13(d). Cell tumbling at low shear rates is slightly hindered in non-aggregating suspensions in comparison to tumbling of a single RBC in shear flow due to cell crowding, which results in sliding of cells over each other; this is shown by a higher peak in the orientational angle distribution (green curve) in Fig. 13(d) with respect to the theoretical prediction (blue curve). In contrast, RBC tumbling in aggregating suspensions appears to be nearly uniform, since RBCs tumble within multiple-cell rouleaux structures. At high shear rates, larger than about $200\ s^{-1}$, individual RBCs are subject to tank-treading motion illustrated by a narrow θ distribution (black line) in Fig. 13(d). At yet higher shear rates RBCs become strongly elongated as indicated by the RBC asphericity distribution in Fig. 13(c).

The most interesting and complex cell dynamics, however, occurs in the broad intermediate regime of shear rates between $5\ s^{-1}$ and $200\ s^{-1}$, where RBC aggregation interactions can be neglected. This range also corresponds to the main region of shear thinning for the non-aggregating cell suspension. In this range of shear rates, RBCs within the suspension experience severe deformations documented by a pronounced increase in the membrane bending energy and in its variation shown in Fig. 13(b). The asphericity distribution for $\dot{\gamma} = 45\ s^{-1}$ in Fig. 13(c) shows that RBCs attain on average a more spherical shape indicating transient folded conformations. This may result in a reduction of shear stresses due to collisional constraints of cell tumbling, and therefore in shear thinning. In addition, the transition of some cells to the tank-treading motion further reduces the shear stresses contributing to the viscosity thinning.

5 Summary

We have presented a mesoscopic model of RBCs implemented by the dissipative particle dynamics method. The spectrin cytoskeleton is represented by a network of interconnected viscoelastic springs comprising a membrane with elastic and viscous properties. The surface network accounts for bending resistance attributed to the lipid bilayer and incorporates local and global area constraints to ensure constant volume and surface area. The macro-

scopic properties of the membrane were related to the network parameters by theoretical analysis. RBC dynamics was simulated in shear flow, where a cell exhibits tumbling at low shear rates and tank-treading at high shear rates. A narrow intermittent region appears where these modes interchange. The model is able to quantitatively capture cell dynamics in shear flow. Comparison of the numerical results with existing theoretical predictions suggest that the latter suffers from oversimplification .

Results on the rheological properties of human blood suggest that the RBC suspension model is able to accurately predict shear-dependent viscosity of blood with and without aggregation interactions between RBCs. The RBC aggregation model was able to properly capture the assembly of RBCs into rouleaux structures. These simulations also confirmed that whole blood is a fluid with a non-zero yield stress. We have shown how single RBC characteristics and behavior contribute to the macroscopic properties of blood, which may not be possible to elucidate in experiments. The predictive capability of the current cell/-capsule suspension model can readily be extended to a variety of engineering and material science applications, which may aid in the development of new soft materials. Finally, such simulations of soft capsule suspensions are computationally demanding and are only feasible on massively parallel computers.

Acknowledgments

We would like to thank Gerhard Gompper, Bruce Caswell, and George E. Karniadakis for many fruitful discussions. Dmitry A. Fedosov acknowledges funding by the Humboldt Foundation. The computations were performed on JUROPA with a grant of computer time provided by the VSR of the Forschungszentrum Jülich.

Appendix

Dissipative Particle Dynamics

Dissipative particle dynamics (DPD)[21,38] is a mesoscopic particle method, where each particle represents a *molecular cluster* rather than an individual atom, and can be thought of as a soft lump of fluid. The DPD system consists of N point particles of mass m_i, position $\mathbf{r}_i$ and velocity $\mathbf{v}_i$. DPD particles interact through three forces: conservative ($\mathbf{F}^C_{ij}$), dissipative ($\mathbf{F}^D_{ij}$), and random ($\mathbf{F}^R_{ij}$) forces given by

$$\mathbf{F}^C_{ij} = F^C_{ij}(r_{ij})\hat{\mathbf{r}}_{ij}, \quad \mathbf{F}^D_{ij} = -\gamma\omega^D(r_{ij})(\mathbf{v}_{ij}\cdot\hat{\mathbf{r}}_{ij})\hat{\mathbf{r}}_{ij}, \quad \mathbf{F}^R_{ij} = \sigma\omega^R(r_{ij})\frac{\xi_{ij}}{\sqrt{dt}}\hat{\mathbf{r}}_{ij}, \quad (30)$$

where $\hat{\mathbf{r}}_{ij} = \mathbf{r}_{ij}/r_{ij}$, and $\mathbf{v}_{ij} = \mathbf{v}_i - \mathbf{v}_j$. The coefficients γ and σ define the strength of dissipative and random forces, respectively. In addition, ω^D and ω^R are weight functions, and ξ_{ij} is a normally distributed random variable with zero mean, unit variance, and $\xi_{ij} = \xi_{ji}$. All forces are truncated beyond the cutoff radius r_c. The conservative force is given by

$$F^C_{ij}(r_{ij}) = a_{ij}(1 - r_{ij}/r_c) \quad \text{for} \quad r_{ij} \le r_c, \quad (31)$$

where a_{ij} is the conservative force coefficient between particles i and j. The random and dissipative forces form a thermostat and must satisfy the fluctuation-dissipation theorem in order for the DPD system to maintain equilibrium temperature T[39]. This leads to

$$\omega^D(r_{ij}) = \left[\omega^R(r_{ij})\right]^2, \qquad \sigma^2 = 2\gamma k_B T, \tag{32}$$

where k_B is the Boltzmann constant. The choice for the weight functions is as follows

$$\omega^R(r_{ij}) = (1 - r_{ij}/r_c)^k \quad \text{for} \quad r_{ij} \leq r_c, \tag{33}$$

where k is an exponent. The time evolution of velocities and positions of particles is determined by the Newton's second law of motion

$$d\mathbf{r}_i = \mathbf{v}_i dt, \qquad d\mathbf{v}_i = \frac{1}{m_i} \sum_{j \neq i} \left(\mathbf{F}_{ij}^C + \mathbf{F}_{ij}^D + \mathbf{F}_{ij}^R\right) dt. \tag{34}$$

The above equations of motion are integrated using the modified velocity-Verlet algorithm[38].

References

1. E. W. Merrill, E. R. Gilliland, G. Cokelet, H. Shin, A. Britten, and JR. R. E. Wells, *Rheology of human blood near and at zero flow*, Biophys. J., **3**, 199–213, 1963.
2. S. Chien, S. Usami, H. M. Taylor, J. L. Lundberg, and M. I. Gregersen, *Effects of hematocrit and plasma proteins on human blood rheology at low shear rates*, J. App. Physiol., **21**, no. 1, 81–87, 1966.
3. R. Skalak, S. R. Keller, and T. W. Secomb, *Mechanics of blood flow*, J. Biomech. Engin., **103**, 102–115, 1981.
4. E. W. Merrill, E. R. Gilliland, T. S. Lee, and E. W. Salzman, *Blood Rheology: Effect of Fibrinogen Deduced by Addition*, Circ. Res., **18**, 437–446, 1966.
5. S. Chien, S. Usami, R. J. Kellenback, and M. I. Gregersen, *Shear-dependent interaction of plasma proteins with erythrocytes in blood rheology*, Am. J. Physiol., **219**, no. 1, 143–153, 1970.
6. G. Cokelet, E. W. Merrill, E. R. Gilliland, H. Shin, A. Britten, and JR. R. E. Wells, *The rheology of human blood-measurement near and at zero shear rate*, Transaction of the Society of Rheology, **7**, 303–317, 1963.
7. A. L. Copley, C. R. Huang, and R. G. King, *Rheogoniometric studies of whole human blood at shear rates from 1,000-0.0009* $\sec^{-1}$. *Part I. Experimental findings*, Biorheology, **10**, 17–22, 1973.
8. M. Abkarian, M. Faivre, and A. Viallat, *Swinging of red blood cells under shear flow*, Phys. Rev. Lett., **98**, 188302, 2007.
9. J. M. Skotheim and T. W. Secomb, *Red blood cells and other nonspherical capsules in shear flow: Oscillatory dynamics and the tank-treading-to-tumbling transition*, Phys. Rev. Lett., **98**, 078301, 2007.
10. Y. C. Fung, *Biomechanics: Mechanical properties of living tissues*, Springer-Verlag, New York, second edition, 1993.
11. C. D. Eggleton and A. S. Popel, *Large deformation of red blood cell ghosts in a simple shear flow*, Phys. Fluids, **10**, no. 8, 1834, 1998.

12. C. Pozrikidis, *Numerical Simulation of Cell Motion in Tube Flow*, Ann. Biomed. Engin., **33**, no. 2, 165–178, 2005.
13. H. Noguchi and G. Gompper, *Shape transitions of fluid vesicles and red blood cells in capillary flows*, Proc. Natl. Acad. Sci. USA, **102**, no. 40, 14159–14164, 2005.
14. M. M. Dupin, I. Halliday, C. M. Care, L. Alboul, and L. L. Munn, *Modeling the flow of dense suspensions of deformable particles in three dimensions*, Phys. Rev. E, **75**, no. 6, 066707, 2007.
15. I. V. Pivkin and G. E. Karniadakis, *Accurate coarse-grained modeling of red blood cells*, Phys. Rev. Lett., **101**, no. 11, 118105, 2008.
16. D. A. Fedosov, B. Caswell, and G. E. Karniadakis, *A multiscale red blood cell model with accurate mechanics, rheology, and dynamics*, Biophys. J., **98**, no. 10, 2215–2225, 2010.
17. Y. Liu and W. K. Liu, *Rheology of red blood cell aggregation by computer simulation*, J. Comp. Phys., **220**, 139–154, 2006.
18. J. L. McWhirter, H. Noguchi, and G. Gompper, *Flow-induced clustering and alignment of vesicles and red blood cells in microcapillaries*, Proc. Natl. Acad. Sci. USA, **106**, no. 15, 6039–6043, 2009.
19. J. B. Freund and M. M. Orescanin, *Cellular flow in a small blood vessel*, J. Fluid Mech., **671**, 466–490, 2011.
20. D. A. Fedosov, B. Pan, W. Caswell, G. Gompper, and G. E. Karniadakis, *Predicting human blood viscosity in silico*, Proc. Natl. Acad. Sci. USA, **108**, 11772–11777, 2011.
21. P. J. Hoogerbrugge and J. M. V. A. Koelman, *Simulating microscopic hydrodynamic phenomena with dissipative particle dynamics*, Europhys. Lett., **19**, no. 3, 155–160, 1992.
22. D. E. Discher, D. H. Boal, and S. K. Boey, *Simulations of the erythrocyte cytoskeleton at large deformation. II. Micropipette aspiration*, Biophys. J., **75**, no. 3, 1584–1597, 1998.
23. D. A. Fedosov, B. Caswell, and G. E. Karniadakis, *Systematic coarse-graining of spectrin-level red blood cell models*, Computer Meth. Appl. Mech. Engin., **199**, 1937–1948, 2010.
24. P. Espanol, *Fluid particle model*, Phys. Rev. E, **57**, no. 3, 2930–2948, 1998.
25. M. Dao, J. Li, and S. Suresh, *Molecularly based analysis of deformation of spectrin network and human erythrocyte*, Materials Sci. Engin. C, **26**, 1232–1244, 2006.
26. M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids*, Clarendon Press, Oxford, 1987.
27. W. Helfrich, *Elastic properties of lipid bilayers: theory and possible experiments*, Z. Naturforschung C, **28**, 693–703, 1973.
28. J. Lidmar, L. Mirny, and D. R. Nelson, *Virus shapes and buckling transitions in spherical shells*, Phys. Rev. E, **68**, no. 5, 051910, 2003.
29. A. W. Lees and S. F. Edwards, *The computer study of transport processes under extreme conditions*, J. Phys. C, **5**, 1921–1928, 1972.
30. R. Tran-Son-Tay, S. P. Sutera, and P. R. Rao, *Determination of RBC membrane viscosity from rheoscopic observations of tank-treading motion*, Biophys. J., **46**, no. 1, 65–72, 1984.

31. T. M. Fischer, *Shape memory of human red blood cells*, Biophys. J., **86**, no. 5, 3304–3313, 2004.
32. T. M. Fischer, *Tank-Tread Frequency of the Red Cell Membrane: Dependence on the Viscosity of the Suspending Medium*, Biophys. J., **93**, no. 7, 2553–2561, 2007.
33. G. R. Cokelet and H. J. Meiselman, *Rheological comparison of hemoglobin solutions and erythrocyte suspensions*, Science, **162**, 275–277, 1968.
34. S. Kessler, R. Finken, and U. Seifert, *Swinging and tumbling of elastic capsules in shear flow*, J. Fluid Mech., **605**, 207–226, 2008.
35. Q. Zhao, L. G. Durand, L. Allard, and G. Cloutier, *Effects of a sudden flow reduction on red blood cell rouleau formation and orientation using RF backscattered power*, Ultrasound Med. Biol., **24**, 503–511, 1998.
36. N. Casson, "A flow equation for pigmented oil suspension of printing ink", in: Rheology of Disperse Systems, C. C. Mill, (Ed.), pp. 84–104. Pergamon Press, New York, 1992.
37. A. Iordan, A. Duperray, and C. Verdier, *Fractal approach to the rheology of concentrated suspensions*, Phys. Rev. E, **77**, 011911, 2008.
38. R. D. Groot and P. B. Warren, *Dissipative particle dynamics: Bridging the gap between atomistic and mesoscopic simulation*, J. Chem. Phys., **107**, no. 11, 4423–4435, 1997.
39. P. Espanol and P. Warren, *Statistical mechanics of dissipative particle dynamics*, Europhys. Lett., **30**, no. 4, 191–196, 1995.

Introduction to Parallel Computing

Bernd Mohr

Institute for Advanced Simulation
Jülich Supercomputing Centre
Forschungszentrum Jülich, 52425 Jülich, Germany
E-mail: b.mohr@fz-juelich.de

The major parallel programming models for scalable parallel architectures are the message passing model and the shared memory model. This article outlines the main concepts of these models as well as the industry standard programming interfaces MPI and OpenMP. To exploit the potential performance of parallel computers, programs need to be carefully designed and tuned. We will discuss design decisions for good performance as well as programming tools that help the programmer in program tuning.

1 Introduction

Many applications like numerical simulations in industry and research as well as commercial applications such as query processing, data mining, and multi-media applications require more compute power than provided by sequential computers. Current hardware architectures offering high performance do not only exploit parallelism within a single processor via multiple CPU cores but also apply a medium to large number of processors concurrently to a single computation. High-end parallel computers currently (2012) deliver up to 10 Petaflop/s (10^{15} floating point operations per second). Parallel programming is required to fully exploit the compute power of the multiple cores.

This article concentrates on programming numerical applications on parallel computer architectures introduced in Sec. 1.1. Parallelization of those applications centers around selecting a decomposition of the data domain onto the processors such that the workload is well balanced and the communication between processors is reduced (Sec. 1.2)[4].

The parallel implementation is then based on either the message passing or the shared memory model (Sec. 2). The standard programming interface for the message passing model is MPI (Message Passing Interface)[8–12], offering a complete set of communication routines (Sec. 3). OpenMP[13–15] is the standard for directive-based shared memory programming and will be introduced in Sec. 4.

Since parallel programs exploit multiple threads of control, debugging is even more complicated than for sequential programs. Sec. 5 outlines the main concepts of parallel debuggers and presents TotalView[21] and DDT[3], the most widely available debuggers for parallel programs.

Although the domain decomposition is key to good performance on parallel architectures, program efficiency also heavily depends on the implementation of the communication and synchronization required by the parallel algorithms and the implementation techniques chosen for sequential kernels. Optimizing those aspects is very system dependent and thus, an interactive tuning process consisting of measuring performance data and applying optimizations follows the initial coding of the application. The tuning process is supported by programming model specific performance analysis tools. Sec. 6 presents basic performance analysis techniques.

1.1 Parallel Architectures

A *parallel computer* or *multi-processor system* is a computer utilizing more than one processor. A common way to classify parallel computers is to distinguish them by the way how processors can access the system's main memory because this influences heavily the usage and programming of the system.

In a *distributed memory architecture* the system is composed out of single-processor nodes with local memory. The most important characteristic of this architecture is that access to the local memory is faster than to remote memory. It is the challenge for the programmer to assign data to the processors such that most of the data accessed during the computation are already in the node's local memory. Two major classes of distributed memory computers can be distinguished:

No Remote Memory Access (NORMA) computers do not have any special hardware support to access another node's local memory directly. The nodes are only connected through a computer network. Processors obtain data from remote memory only by exchanging messages over this network between processes on the requesting and the supplying node. Computers in this class are sometimes also called **Network Of Workstations (NOW)** or **Clusters Of Workstations (COW)**.

Remote Memory Access (RMA) computers allow to access remote memory via specialized operations implemented by hardware, however the hardware does not provide a global address space, i.e., a memory location is not determined via an address in a shared linear address space but via a tuple consisting of the processor number and the local address in the target processor's address space.

The major advantage of distributed memory systems is their ability to scale to a very large number of nodes. Today (2012), systems with more than 700,000 cores have been built. The disadvantage is that such systems are very hard to program.

In contrast, a *shared memory architecture* provides (in hardware) a global address space, i.e., all memory locations can be accessed via usual load and store operations. Access to a remote location results in a copy of the appropriate cache line in the processor's cache. Therefore, such a system is much easier to program. However, shared memory systems can only be scaled to moderate numbers of processors, typically 64 or 128. Shared memory systems are further classified according to the quality of the memory accesses:

Uniform Memory Access (UMA) computer systems feature one global shared memory subsystem which is connected to the processors through a central bus or memory switch. All of the memory is accessible to all processors in the same way. Such a system is also often called **Symmetrical Multi Processor (SMP)**.

Non Uniform Memory Access (NUMA) computers are more scalable by physically distributing the memory but still providing a hardware implemented global address space. Therefore access to memory local or close to a processor is faster than to remote memory. If such a system has additional hardware which also ensures that multiple copies of data stored in different cache lines of the processors is kept coherent, i.e., the copies always do have the same value, then it is called a **Cache-Coherent Non Uniform Memory Access (ccNUMA)** system. ccNUMA systems offer the abstraction of a shared linear address space resembling physically shared memory systems. This abstraction simplifies the task of program development but does not necessarily facilitate program tuning.

While most of the early parallel computers were simple single processor NORMA systems, today's large parallel systems are typically *hybrid systems*, i.e., shared memory NUMA nodes with a moderate number of multi-core processors are connected together to form a distributed memory cluster system. To further increase their compute power, current high-end systems also deploy additional so-called *accelerators* attached to their nodes. These are often special versions of graphics processing units (GPU) with enhanced floating-point performance and error-correcting memory which are little shared-memory systems themselves further increasing the overall system complexity.

1.2 Data Parallel Programming

Applications that scale to a large number of processors usually perform computations on large data domains. For example, crash simulations are based on partial differential equations that are solved on a large finite element grid and molecular dynamics applications simulate the behavior of a large number of particles. Other parallel applications apply linear algebra operations to large vectors and matrices. The elemental operations on each object in the data domain can be executed in parallel by the available processors.

The scheduling of operations to processors is determined by a *domain decomposition*[5] specified by the programmer. Processors execute those operations that determine new values for elements stored in local memory (owner-computes rule). While processors execute an operation, they may need values from other processors. The domain decomposition has thus to be chosen so that the distribution of operations is balanced and the communication is minimized. The third goal is to optimize single node computation, i.e., to be able to exploit the processor's pipelines and the processor's caches efficiently.

A good example for the design decisions taken when selecting a domain decomposition is Gaussian elimination[1]. The main structure of the matrix during the steps of the algorithm is outlined in Fig. 1.

The goal of this algorithm is to eliminate all entries in the matrix below the main diagonal. It starts at the top diagonal element and subtracts multiples of the first row from the second and subsequent rows to end up with zeros in the first column. This operation is repeated for all the rows. In later stages of the algorithm the actual computations have to be done on rectangular sections of decreasing size. If the main diagonal element of the current row is zero, a pivot operation has to be performed. The subsequent row with the maximum value in this column is selected and exchanged with the current row.

A possible distribution of the matrix is to decompose its columns into blocks, one block for each processor. The elimination of the entries in the lower triangle can then be performed in parallel where each processor computes new values for its columns only. The main disadvantage of this distribution is that in later computations of the algorithm only a subgroup of the processors is actually doing any useful work since the computed rectangle is getting smaller.

To improve load balancing, a cyclic column distribution can be applied. The computations in each step of the algorithm executed by the processors differ only in one column.

In addition to load balancing also communication needs to be minimized. Communication occurs in this algorithm for broadcasting the current column to all the processors since it is needed to compute the multiplication factor for the row. If the domain decomposition is a row distribution, which eliminates the need to communicate the current column, the current row needs to be broadcast to the other processors.

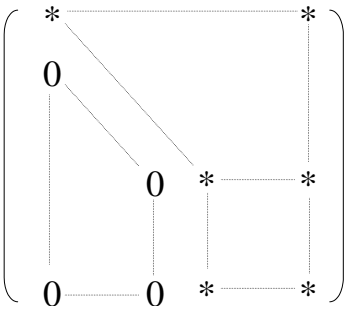


Figure 1. Structure of the matrix during Gaussian elimination.

If we consider also the pivot operation, communication is necessary to select the best row when a row-wise distribution is applied since the computation of the global maximum in that column requires a comparison of all values.

Selecting the best domain decomposition is further complicated due to optimizing single node performance. In this example, it is advantageous to apply BLAS3[2] operations for the local computations. These operations make use of blocks of rows to improve cache utilization. Blocks of rows can only be obtained if a block-cyclic distribution is applied, i.e., columns are not distributed individually but blocks of columns are cyclically distributed.

This discussion makes clear, that choosing a domain decomposition is a very complicated step in program development. It requires deep knowledge of the algorithm's data access patterns as well as the ability to predict the resulting communication.

2 Programming Models

Programming parallel computers is almost always done via the so-called *Single Program Multiple Data* (SPMD) model. SPMD means that the same program (executable code) is executed on all processors taking part in the computation, but it computes on different parts of the data which were distributed over the processors based on a specific domain decomposition. If computations are only allowed on specific processors, this has to be explicitly programmed by using conditional programming constructs (e.g., with `if` or `where` statements). There are two main programming models, *message passing* and *shared memory*, offering different features for implementing applications parallelized by domain decomposition.

2.1 Message Passing

The message passing model is based on a set of processes with private data structures. Processes communicate by exchanging messages with special send and receive operations. It is a natural fit for programming distributed memory machines but also can be used on shared memory computers. The domain decomposition is implemented by developing a code describing the local computations and local data structures of a single process. Thus, global arrays have to be split up and only the local part has to be allocated in a process. This handling of global data structures is called *data distribution*. Computations on the

global arrays also have to be transformed, e.g., by adapting the loop bounds, to ensure that only local array elements are computed. Access to remote elements has to be implemented via explicit communication, temporary variables have to be allocated, messages have to be constructed and transmitted to the target process. he standard programming interface for the message passing model is MPI (Message Passing Interface)[8–12], offering a complete set of communication routines (see next Sec.).

2.2 Shared Memory

The shared memory model is based on a set of threads that is created when parallel operations are executed. This type of computation is also called *fork-join parallelism*. Threads share a global address space and thus access array elements via a global index. The main parallel operations are *parallel loops* and *parallel sections*. Parallel loops are executed by a set of threads also called a *team*. The iterations are distributed among the threads according to a predefined strategy. This scheduling strategy implements the chosen domain decomposition. Parallel sections are also executed by a team of threads but the tasks assigned to the threads implement different operations. This feature can for example be applied if domain decomposition itself does not generate enough parallelism and whole operations can be executed in parallel since they access different data structures.

In the shared memory model, the distribution of data structures onto the node memories is not enforced by decomposing global arrays into local arrays, but the global address space is distributed onto the memories by the operating system. For example, the pages of the virtual address space can be distributed cyclically or can be assigned at first touch. The chosen domain decomposition thus has to take into account the granularity of the distribution, i.e., the size of pages, as well as the system-dependent allocation strategy.

While the domain decomposition has to be hard-coded into the message passing program, it can easily be changed in a shared memory program by selecting a different scheduling strategy for parallel loops.

Another advantage of the shared memory model is that automatic and incremental parallelization is supported. While automatic parallelization leads to a first working parallel program, its efficiency typically needs to be improved. The reason for this is that parallelization techniques work on a loop-by-loop basis and do not globally optimize the parallel code via a domain decomposition. In addition, dependence analysis, the prerequisite for automatic parallelization, is limited to access patterns known at compile time. The biggest disadvantage of this model is that it can only be used on shared memory computers.

In the shared memory model, a first parallel version is relatively easy to implement and can be incrementally tuned. In the message passing model instead, the program can be tested only after finishing the full implementation. Subsequent tuning by adapting the domain decomposition is usually time consuming.

OpenMP[13–15] is the standard for directive-based shared memory programming and will be introduced in Sec. 4.

2.3 Programming Accelerators

Currently, programming of accelerator devices is not supported by traditional techniques like MPI or OpenMP, but requires yet another level of parallel programming. Most

widespread are either low-level, non-portable techniques like Compute Unified Device Architecture (CUDA)[26,27] or Open Computing Language (OpenCL)[28] or high-level approaches based on pragmas like HMPP[29] or OpenACC[30]. Pragmas are meta-information added in the application source code that do not change the semantic of the original code. They provide a portable way to specify the remote execution (RPC) of functions or regions of code on GPUs and many-core accelerators as well as the transfer of data to and from the target device memory. They offer an incremental way of migrating applications by first declaring and generating kernels of critical computations, then by managing data transfers and finally by optimizing kernel performance and data synchronization.

3 MPI

The Message Passing Interface (MPI)[8–12] was mainly developed between 1993 and 1997. It is a community standard which standardizes the calling interface for a communication and synchronization function library. It provides Fortran 77, Fortran 90, C and C++ language bindings. It includes routines for point-to-point communication, collective communication, one-sided communication, parallel IO, and dynamic task creation. Currently, almost all available open-source and commercial MPI implementations support the 2.0 standard with the exception of dynamic task creation, which is only implemented by a few. In 2008 and 2009, updates and clarifications of the standard were published as Version 2.1 and 2.2 and work has begun to define further enhancements (version 3.x). For a simple example see the appendix.

3.1 MPI Basic Routines

MPI consists of more than 320 functions. But realistic programs can already be developed based on no more than six functions:

MPI_Init initializes the library. It has to be called at the beginning of a parallel operation before any other MPI routines are executed.

MPI_Finalize frees any resources used by the library and has to be called at the end of the program.

MPI_Comm_size determines the number of processors executing the parallel program.

MPI_Comm_rank returns the unique process identifier.

MPI_Send transfers a message to a target process. This operation is a blocking send operation, i.e., it terminates when the message buffer can be reused either because the message was copied to a system buffer by the library or because the message was delivered to the target process.

MPI_Recv receives a message. This routine terminates if a message was copied into the receive buffer.

3.2 MPI Communicator

All communication routines depend on the concept of a *communicator*. A communicator consists of a process group and a communication context. The processes in the process group are numbered from zero to process count - 1. The process number returned by

MPI_Comm_rank is the identification in the process group of the communicator which is passed as a parameter to this routine.

The communication context of the communicator is important in identifying messages. Each message has an integer number called a *tag* which has to match a given selector in the corresponding receive operation. The selector depends on the communicator and thus on the communication context. It selects only messages with a fitting tag and having been sent relative to the same communicator. This feature is very useful in building parallel libraries since messages sent inside the library will not interfere with messages outside if a special communicator is used in the library. The default communicator that includes all processes of the application is MPI_COMM_WORLD.

3.3 MPI Collective Operations

Another important class of operations are *collective operations*. Collective operations are executed by a process group identified via a communicator. All the processes in the group have to perform the same operation. Typical examples for such operations are:

MPI_Barrier synchronizes all processes. None of the processes can proceed beyond the barrier until all the processes started execution of that routine.

MPI_Bcast allows to distribute the same data from one process, the so-called *root* process, to all other processes in the process group.

MPI_Scatter also distributes data from a root process to a whole process group, but each receiving process gets different data.

MPI_Gather collects data from a group of processes at a root process.

MPI_Reduce performs a global operation on the data of each process in the process group. For example, the sum of all values of a distributed array can be computed by first summing up all local values in each process and then summing up the local sums to get a global sum. The latter step can be performed by the reduction operation with the parameter MPI_SUM. The result is delivered to a single target processor.

3.4 MPI IO

Data parallel applications make use of the IO subsystem to read and write big data sets. These data sets result from replicated or distributed arrays. The reasons for IO are to read input data, to pass information to other programs, e.g., for visualization, or to store the state of the computation to be able to restart the computation in case of a system failure or if the computation has to be split into multiple runs due to its resource requirements.

IO can be implemented in three ways:

Sequential IO A single node is responsible to perform the IO. It gathers information from the other nodes and writes it to disk or reads information from disk and scatters it to the appropriate nodes. Whereas this approach might be feasible for small amounts of data, it bears serious scalability issues, as modern IO subsystems can only be utilized efficiently with parallel data streams and aggregated waiting time increases rapidly at larger scales.

Private IO Each node accesses its own files. The big advantage of this implementation is that no synchronization among the nodes is required and very high performance can

be obtained. The major disadvantage is that the user has to handle a large number of files. For input the original data set has to be split according to the distribution of the data structure and for output the process-specific files have to be merged into a global file for post-processing.

Parallel IO In this implementation all the processes access the same file. They read and write only those parts of the file with relevant data. The main advantages are that no individual files need to be handled and that reasonable performance can be reached. The parallel IO interface of MPI provides flexible and high-level means to implement applications with parallel IO.

Files accessed via MPI IO routines have to be opened and closed by collective operations. The open routine allows to specify hints to optimize the performance such as whether the application might profit from combining small IO requests from different nodes, what size is recommended for the combined request, and how many nodes should be engaged in merging the requests.

The central concept in accessing the files is the *view*. A view is defined for each process and specifies a sequence of data elements to be ignored and data elements to be read or written by the process. When reading or writing a distributed array the local information can be described easily as such a repeating pattern. The IO operations read and write a number of data elements on the basis of the defined view, i.e., they access the local information only. Since the views are defined via runtime routines prior to the access, the information can be exploited in the library to optimize IO.

MPI IO provides blocking as well as nonblocking operations. In contrast to blocking operations, the nonblocking ones only start IO and terminate immediately. If the program depends on the successful completion of the IO it has to check it via a test function. Besides the collective IO routines which allow to combine individual requests, also non-collective routines are available to access shared files.

3.5 MPI Remote Memory Access

Remote memory access (RMA) operations (also called *one-sided communication*) allow to access the address space of other processes without participation of the other process. The implementation of this concept can either be in hardware, such as in the CRAY T3E, or in software via additional threads waiting for requests. The advantages of these operations are that the protocol overhead is much lower than for normal send and receive operations and that no polling or global communication is required for setting up communication.

In contrast to explicit message passing where synchronization happens implicitly, accesses via RMA operations need to be protected by explicit synchronization operations.

RMA communication in MPI is based on the *window concept*. Each process has to execute a collective routine that defines a window, i.e., the part of its address space that can be accessed by other processes.

The actual access is performed via *put* and *get* operations. The address is defined by the target process number and the displacement relative to the starting address of the window for that process.

MPI also provides special synchronization operations relative to a window. The MPI_Win_fence operation synchronizes all processes that make some address ranges accessible to other processes. It is a collective operation that ensures that all RMA operations

started before the fence operation terminate before the target process executes the fence operation and that all RMA operations of a process executed after the fence operation are executed after the target process executed the fence operation. There are also more fine grained synchronization methods available in the form of the General Active Target Synchronization or via locks.

4 OpenMP

OpenMP[13–15] is a directive-based programming interface for the shared memory programming model. It consists of a set of directives and runtime routines for Fortran 77 (published 1997), for Fortran 90 (2000), and a corresponding set of pragmas for C and C++ (1998). In 2005, a combined Fortran, C, and C++ standard (Version 2.5) was published, which was updated in 2008 (Version 3.0) and 2011 (Version 3.1).

Directives are special comments that are interpreted by the compiler. Directives have the advantage that the code is still a sequential code that can be executed on sequential machines (by ignoring the directives/pragmas) and therefore there is no need to maintain separate sequential and parallel versions.

Directives start and terminate parallel regions. When the master thread hits a parallel region a team of threads is created or activated. The threads execute the code in parallel and are synchronized at the beginning and the end of the computation. After the final synchronization the master thread continues sequential execution after the parallel region. The main directives are:

!$OMP PARALLEL DO specifies a loop that can be executed in parallel. The DO loop's iterations can be distributed among the set of threads according to various scheduling strategies including STATIC(CHUNK), DYNAMIC(CHUNK), and GUIDED(CHUNK). STATIC(CHUNK) distribution means that the set of iterations are consecutively distributed among the threads in blocks of CHUNK size (resulting in block and cyclic distributions). DYNAMIC(CHUNK) distribution implies that iterations are distributed in blocks of CHUNK size to threads on a first-come-first-served basis. GUIDED (CHUNK) means that blocks of exponentially decreasing size are assigned on a first-come-first-served basis. The size of the smallest block is determined by CHUNK size.

!$OMP PARALLEL SECTIONS starts a set of sections that are each executed in parallel by a team of threads.

!$OMP PARALLEL introduces a code region that is executed redundantly by the threads. It has to be used very carefully since assignments to global variables will lead to conflicts among the threads and possibly to nondeterministic behavior.

!$OMP DO / FOR is a work sharing construct and may be used within a parallel region. All the threads executing the parallel region have to cooperate in the execution of the parallel loop. There is no implicit synchronization at the beginning of the loop but a synchronization at the end. After the final synchronization all threads continue after the loop in the replicated execution of the program code.

The main advantage of this approach is that the overhead for starting up the threads is eliminated. The team of threads exists during the execution of the parallel region and need not be built before each parallel loop.

!$OMP SECTIONS is also a work sharing construct that allows the current team of threads executing the surrounding parallel region to cooperate in the execution of the parallel sections.

!$OMP TASK is only available with the new version 3.0 of the standard and greatly simplifies the parallelization on non-loop constructs by allowing to dynamically specify portions of the programs which can run independently.

Program data can either be shared or private. While threads do have their own copy of private data, only one copy exists of shared data. This copy can be accessed by all threads. To ensure program correctness, OpenMP provides special synchronization constructs. The main constructs are *barrier synchronization* enforcing that all threads have reached this synchronization operation before execution continues and *critical sections*. Critical sections ensure that only a single thread can enter the section and thus, data accesses in such a section are protected from race conditions. For example, a common situation for a critical section is the accumulation of values. Since an accumulation consists of a read and a write operation unexpected results can occur if both operations are not surrounded by a critical section. For a simple example of an OpenMP parallelization see the appendix.

5 Parallel Debugging

Debugging parallel programs is more difficult than debugging sequential programs not only since multiple processes or threads need to be taken into account but also because program behavior might not be deterministic and might not be reproducible. These problems are not solved by current state-of-the-art commercial parallel debuggers. They only deal with the first problem by providing menus, displays, and commands that allow to inspect individual processes and execute commands on individual or all processes.

Two widely used debuggers are TotalView from Rogue Wave Software[21] and DDT from Allinea[3]. They provide breakpoint definition, single stepping, and variable inspection for parallel programs via an interactive interface. The programmer can execute those operations for individual processes and groups of processes. They also provides some means to summarize information such that equal information from multiple processes is combined into a single information and not repeated redundantly. They also support MPI and OpenMP programs on many platforms.

6 Parallel Performance Analysis

Performance analysis is an iterative subtask during program development. The goal is to identify program regions that do not perform well. Performance analysis is structured into three phases:

Measurement: Performance analysis is done based on information on runtime events gathered during program execution. The basic events are, for example, cache misses, termination of a floating point operation, start and stop of a subroutine or message passing operation. The information on individual events can be summarized during program execution (*profiling*) or individual trace records can be collected for each event (*tracing*).

Analysis: During analysis the collected runtime data are inspected to detect *performance problems*. Performance problems are based on *performance properties*, such as the existence of message passing in a program region, which have a condition for identifying it and a severity function that specifies its importance for program performance. Current tools support the user in checking the conditions and the severity by a visualization of the program behavior. Future tools might be able to automatically detect performance properties based on a specification of possible properties. During analysis the programmer applies a threshold. Only performance properties whose severity exceeds this threshold are considered to be performance problems.

Ranking: During program analysis the severest performance problems need to be identified. This means that the problems need to be ranked according to the severity. The most severe problem is called the *program bottleneck*. This is the problem the programmer tries to resolve by applying appropriate program transformations.

Current techniques for performance data collection are *profiling* and *tracing*. Profiling collects summary data only. This can be done via *sampling*. The program is regularly interrupted, e.g., every 10 ms, and the information is added up for the source code location which was executed in this moment. For example, the UNIX profiling tool *prof* applies this technique to determine the fraction of the execution time spent in individual subroutines.

A more precise profiling technique is based on *instrumentation*, i.e., special calls to a *monitoring library* are inserted into the program. This can either be done in the source code by the compiler or specialized tools, or can be done in the object code. While the first approach allows to instrument more types of regions, for example, loops and vector statements, the latter allows to measure data for programs where no source code is available. The monitoring library collects the information and adds it to special counters for the specific region.

Tracing is a technique that collects information for each event. This results, for example, in very detailed information for each instance of a subroutine and for each message sent to another process. The information is stored in specialized trace records for each event type. For example, for each start of a send operation, the time stamp, the message size and the target process can be recorded, while for the end of the operation, the time stamp and bandwidth are stored.

The trace records are stored in the memory of each process and are written to a trace file either when the buffer is filled up or when the program terminates. The individual trace files of the processes are merged together into one trace file ordered according to the time stamps of the events.

Profiling has the advantage to be of moderate size while trace information tends to be very large. The disadvantage of profiling is that it is not fine grained; the behavior of individual instances of subroutines can for example not be investigated since all the information has been summed up.

Widely used performance tools include TAU[19,20] from the University of Oregon, Vampir[22,23] from the Technical University of Dresden, HPCToolkit[25] from Rice University, and Scalasca[17,18] from the Jülich Supercomputing Centre.

7 Summary

This article gave an overview of parallel programming models as well as programming tools. Parallel programming will always be a challenge for programmers. Higher-level programming models and appropriate programming tools only facilitate the process but do not make it a simple task.

While programming in MPI offers the greatest potential performance, shared memory programming with OpenMP is much more comfortable due to the global style of the resulting program. The sequential control flow among the parallel loops and regions matches much better with the sequential programming model all the programmers are trained for.

Although programming tools were developed over years, the current situation seems not to be very satisfying. Program debugging is done per thread, a technique that does not scale to larger numbers of processors. Performance analysis tools do also suffer scalability limitations and, in addition, the tools are complicated to use. The programmers have to be experts for performance analysis to understand potential performance problems, their proof conditions, and their severity. In addition they have to be experts for powerful but also complex user interfaces.

Future research in this area has to try to automate performance analysis tools, such that frequently occurring performance problems can be identified automatically. First automatic tools are already available: ParaDyn[7] from the University of Wisconsin-Madison, Persicope[6] from the Technical University Munich, and Scalasca[17,18] from the Jülich Supercomputing Centre.

A second important trend that will effect parallel programming in the future is the move towards more heterogeneous architectures: more and more machines employ one or more accelerators like GPUs in addition to the multi-core processors within a node of large distributed-memory clusters. This introduces a 3-level parallelism hierarchy (machine - node - accelerator) each requiring a different programming model, e.g. combining message passing between the individual SMP nodes, shared memory programming within a node, plus an accelerator-specific programming model like CUDA. This *hybrid* programming model will lead to even more complex programs and program development tools have to be enhanced to be able to help the user in developing these codes.

A promising approach to reduce complexity in parallel programming in the future are so-called *partitioned global address space* (PGAS) languages[16], such as Unified Parallel C (UPC)[24] or Co-array Fortran (CAF) which provide simple means to distribute data and communicate implicitly via efficient one-sided communication. CAF is part of the latest Fortran 2008 standard (ISO/IEC 1539-1:2010).

Appendix

This appendix provides three versions of a simple example of a scientific computation. It computes the value of π by numerical integration:

$$\pi = \int_0^1 f(x)dx \quad \text{with} \quad f(x) = \frac{4}{1+x^2}$$

This integral can be approximated numerically by the midpoint rule:

$$\pi \approx \frac{1}{n}\int_1^n f(x_i) \quad \text{with} \quad x_i = \frac{(i-0.5)}{n} \quad \text{for} \quad i = 1, \ldots, n$$

Larger values of the parameter n will give us more accurate approximations of π. This is not, in fact, a very good way to compute π, but it makes a good example because it has the typical, complete structure of a numerical simulation program (initialization - loop-based calculation - wrap-up), and the whole source code fits one one page or slide.

To parallelize the example, each process/thread computes and adds up the areas for a different subset of the rectangles. At the end of the computation, all of the local sums are combined into a global sum representing the value of π.

Sequential and OpenMP Version of Example Program

The following listing shows the corresponding implementation of the π integration example using OpenMP. As OpenMP is based on directives (which are plain comments in a non-OpenMP compilation mode), it is at the same time also a sequential implementation of the example.

```
1  program pi_omp
2  implicit none
3  integer          :: i, n
4  double precision :: f, x, sum, pi, h
5
6  open(1, file="pi.dat")
7  read(1,*) n
8
9  h = 1.0d0 / n
10 sum = 0.0d0
11 !$omp parallel do private(i,x) reduction(+:sum)
12 do i = 1, n
13    x = (i - 0.5d0) * h
14    sum = sum + (4.d0/(1.d0 + x*x))
15 end do
16 pi = h * sum
17
18 write(*, fmt="(A, F16.12)") "Value of pi is ", pi
19 end program
```

The OpenMP directive in line 11 declares the following do-loop as parallel resulting in a concurrent execution of loop iterations. As the variables `i` and `x` are used to store values during the execution of the loop, they have to be declared private, so that each thread executing iterations has its own copy. The variable `h` is only read, so it can be shared. Finally, it is specified that there is a reduction (using addition) over the variable `sum`.

MPI Version of Example Program

The following listing shows a Fortran90 implementation of the π numerical integration example parallelized with the help of MPI.

```
1  program pi_mpi
2  implicit none
3  include 'mpif.h'
4  integer          :: i, n, ierr, myrank, numprocs
5  double precision :: f, x, sum, pi, h, mypi
6
7  call MPI_Init(ierr)
8  call MPI_Comm_rank(MPI_COMM_WORLD, myrank, ierr)
9  call MPI_Comm_size(MPI_COMM_WORLD, numprocs, ierr)
10
11 if ( myrank == 0 ) then
12    open(1, file="pi.dat")
13    read(1,*) n
14 end if
15
16 call MPI_Bcast(n, 1, MPI_INTEGER, 0, MPI_COMM_WORLD, ierr)
17
18 h = 1.0d0 / n
19 sum = 0.0d0
20 do i = myrank+1, n, numprocs
21    x = (i - 0.5d0) * h
22    sum = sum + (4.d0/(1.d0 + x*x))
23 end do
24 mypi = h * sum
25
26 call MPI_Reduce(mypi, pi, 1, MPI_DOUBLE_PRECISION, &
27                 MPI_SUM, 0, MPI_COMM_WORLD, ierr)
28
29 if ( myrank == 0 ) then
30   write(*, fmt="(A, F16.12)") "Value of pi is ", pi
31 endif
32
33 call MPI_Finalize(ierr)
34 end program
```

First, the MPI system has to be initialized (lines 7 to 9) and terminated (line 33) with the necessary MPI calls. Next, the input of parameters (line 11 to 14) and the output of results (lines 29 to 31) has to be restricted so that it is only executed by one processor. Then, the input has to be broadcasted to the other processors (line 16). The biggest (and most complicated) change is to program the distribution of work and data. The do-loop in line 20 has to be changed so that each processor only calculates and summarizes its part of the distributed computations. Finally, the reduce call in lines 26/27 collects the local sums and delivers the global sum to processor 0.

As one can see, because of the need to explicitly program all aspects of the parallelization, the MPI version is almost twice as long as the OpenMP version. Although this is clearly more work, it gives a programmer much more ways to express and control parallelism. Also, the MPI version will run on all kinds of parallel computers, while OpenMP is restricted to the shared memory architecture.

References

1. D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, Prentice-Hall (1989).
2. J. J. Dongarra, J. Du Croz, I. S. Duff, and S. Hammarling, A set of Level 3 Basic Linear Algebra Subprograms, *ACM Trans. Math. Soft.*, 16:1–17, (1990).
3. Allinea: *DDT*, `http://allinea.com/`.
4. I. Foster, *Designing and Building Parallel Programs*, Addison Wesley (1994).
5. G. Fox, *Domain Decomposition in Distributed and Shared Memory Environments*, International Conference on Supercomputing June 8-12, 1987, Athens, Greece, Lecture Notes in Computer Science 297, edited by C. Polychronopoulos (1987).
6. M. Gerndt and M. Ott, Automatic Performance Analysis with Periscope, *Concurrency and Computation: Practice and Experience*, Vol. 22, No. 6, 736–748 (2010).
7. B. P. Miller, M. D. Callaghan, J. M. Cargille, J. K. Hollingsworth, R. B. Irvine, K. L. Karavanic, K. Kunchithapadam, and T. Newhall, The Paradyn Parallel Performance Measurement Tool, *IEEE Computer*, Vol. 28, No. 11, 37–46 (1995).
8. MPI Forum: *Message Passing Interface*, `http://www.mpi-forum.org/`.
9. M. Snir, S. Otto, S. Huss-Lederman, D. Walker, and J. Dongarra, *MPI - the Complete Reference, Volume 1, The MPI Core*, 2nd ed., MIT Press (1998).
10. W. Gropp, S. Huss-Lederman, A. Lumsdaine, E. Lusk, B. Nitzberg, W. Saphir, and M. Snir, *MPI - the Complete Reference, Volume 2, The MPI Extensions*, MIT Press (1998).
11. W. Gropp, E. Lusk, A. , Skjellum, *Using MPI, 2nd Edition*, MIT Press (1999).
12. W. Gropp, E. Lusk, R. Thakur, *Using MPI-2: Advanced Features of the Message Passing Interface*, MIT Press (1999).
13. OpenMP Forum: *OpenMP Standard*, `http://www.openmp.org/`.
14. L. Dagum and R. Menon, *OpenMP: An Industry-Standard API for Shared-memory Programming*, IEEE Computational Science & Engineering, 5(1):46–55 (1998).
15. B. Chapman, G. Jost, R. van der Pas, *Using OpenMP: Portable Shared Memory Parallel Programming*, MIT Press (2007).

16. C. Coarfa, Y. Dotsenko, J. Mellor-Crummey, F. Cantonnet, T. El-Ghazawi, A. Mohanty, Y. Yao, An Evaluation of Global Address Space Languages: Co-Array Fortran and Unified Parallel C, *Proceedings of the ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPOPP 2005)*, ACM, (2005).
17. B. J. N. Wylie, M. Geimer, F. Wolf, Performance measurement and analysis of large-scale parallel applications on leadership computing systems, *Scientific Programming*, 16(2–3):167–181, (2008).
18. M. Geimer, F. Wolf, B. J. N. Wylie, B. Mohr, A scalable tool architecture for diagnosing wait states in massively parallel applications, *Parallel Computing*, 35(7):375–388, (2009).
19. S. Shende, A. Malony, A. Morris, S. Parker, J. de St. Germain, Performance evaluation of adaptive scientic applications using TAU, *Parallel Computational Fluid Dynamics – Theory and Applications*, Elsevier, 421–428 (2008).
20. S. Shende, A. Malony, The TAU parallel performance system, *Intl. Journal of High Performance Computing Applications*, 20, 287–331, SAGE Publications, (2006).
21. Rogue Wave Software: *TotalView*, `http://www.roguewave.com/`.
22. H. Brunst, D. Hackenberg, G. Juckeland, H. Rohling, Comprehensive Performance Tracking with Vampir 7, *Tools for High Performance Computing 2009*, LNCS, 17–29, Springer, (2010).
23. A. Knüpfer, H. Brunst, J. Doleschal, M. Jurenz, M. Lieber, H. Mickler, M. S. Müller, W. E. Nagel, The Vampir Performance Analysis Tool-Set, *Tools for High Performance Computing 2008*, LNCS, 139–155, Springer, (2008).
24. UPC Consortium, *UPC Language Specifications, v1.2*, Lawrence Berkeley National Lab Tech Report LBNL-59208, (2005).
25. L. Adhianto, S. Banerjee, M. Fagan, M. Krentel, G. Marin, J. Mellor-Crummey, and N .R. Tallent, HPCToolkit: Tools for performance analysis of optimized parallel programs, *Concurrency and Computation: Practice and Experience*, 22(6):685–701, (2010).
26. R. Farber, *CUDA Application Design and Development*, ISBN-13: 978-0123884268, (2011).
27. J. Nickolls, I. Buck, M. Garland, K. Skadron, Scalable Parallel Programming with CUDA, *ACM Queue* 6(2):40–53, (2008).
28. R. Tsuchiyama, T. Nakamura, T. Iizuka, A. Asahara, S. Miki, *OpenCL Programming Book*, Fixstars Corporation, 2010.
29. R. Dolbeau, S. Bihan and F. Bodin, *HMPP: A Hybrid Multi-core Parallel Programming Environment*, 1st Workshop on General Purpose Processing on Graphics Processing Units, Boston, (2007).
30. OpenACC Application Program Interface, `http://www.openacc-standard.org/`.

Scalability of $\mu\varphi$ and the Parallel Algebraic Multigrid Solver of DUNE-ISTL

Olaf Ippisch and Markus Blatt

Interdisciplinary Center for Scientific Computing
Heidelberg University, 69120 Heidelberg, Germany
E-mail: {Olaf.Ippisch,Markus.Blatt}@iwr.uni-heidelberg.de

The scalability of the application $\mu\varphi$ (MuPhi), a numerical solver for Richards' equation, was tested on the BlueGene/P type parallel computer JUGENE in Jülich at the Extreme Scaling Workshop 2011[1]. The arising linear equation systems were solved with DUNE-ISTL using a BiCGStab solver in combination with an algebraic multigrid preconditioner. We present scaling results for the computation as well as for file I/O up to 294'849 cores and 150 billion unknowns and discuss implementation details for JUGENE.

1 Introduction

Richards' equation is a second order partial differential equation (PDE) describing water flow in partially water saturated porous media:

$$\frac{\partial\theta(\psi_m,\vec{x})}{\partial t} - \nabla\cdot\{K(\theta,\vec{x})\cdot[\nabla\psi_m - \rho_w g\vec{e}_z]\} + r_w = 0.$$

Here, ψ_m is the matrix potential, ρ_w the density of water, g gravitational acceleration, $\vec{e}_z$ the unity vector in the vertical and r_w a source sink term. $\theta(\psi_m,\vec{x})$ is the volumetric water content and $K(\theta,\vec{x})$ is the hydraulic conductivity function. Both are highly nonlinear, spatially heterogeneous material functions of ψ_m.

For time dependent problems Richards' equation is a non-linear (probably degenerated) parabolic equation. For steady-state saturated porous media it is an elliptic PDE (θ and K are then spatially variable constants).

Richards' equation is solved using a cell-centred finite-volume scheme with full upwinding in space and an implicit Euler scheme in time. Linearisation of the nonlinear equations is done by an inexact Newton method with line search.

2 Algebraic Multigrid as Parallel Linear Solver

Most of the computation time is consumed by solving the sparse linear systems arising in the Newton method. For this we use the biconjugate gradient stabilized method[8] preconditioned by a massively parallel algebraic multigrid solver based on aggregation[2].

We decompose the unknowns between the processes such that each unknown is owned by exactly one process. Each process stores whole rows of the matrix corresponding to unknowns owned by the process. Additionally each process has to store all unknowns j with $a_{ij} \neq 0$ for an unknown i it owns. All rows corresponding to rows not owned are set to $a_{jj} = 1$ and $a_{ij} = 0$ for $j \neq i$. Note that for the matrix-vector product we are able to compute the unknowns owned by a process with a local matrix-vector product. For a more detailed description see Ref. 3.

During the setup phase each process computes the aggregates for the unknowns it owns. After communicating this information it sets up the prolongation and coarse level matrix locally. Once the number of unknowns is below a given threshold we agglomerate the data onto fewer processes for a better ratio between communication and computation time. As a side effect we thus facilitate aggregation across process boundaries. The new data decomposition is computed by ParMETIS[6] using the graph of the communication pattern for the matrix-vector product. Unfortunately the parallel algorithm of ParMETIS cannot handle our graph on the full machine. Therefore we have to use the recursive bisection graph partitioning algorithm sequentially on one process. This procedure is repeated on coarser levels until the whole linear system of the coarsest level is stored on one process and is directly solved using SuperLU[4]. As a smoother we use hybrid Gauss-Seidel[10].

3 Test Cases

We formulated two test cases for a groundwater flow problem and for a (time dependent) unsaturated flow problem, respectively. For the test cases the material parameters were either

- homogeneous (thus for the groundwater flow problem the Laplace equation was solved), scenario homog
- heterogeneous with a heterogeneity which was the same on each node, scenario block
- heterogeneous with a structure which was much larger, scenario large

The heterogeneity was created as an equally weighted sum of two log-normal autocorrelated Gaussian random fields with different correlation length (100 voxel horizontally and 20 voxel vertically for the coarse scale and 2 voxel for the fine scale). The random numbers were generated with the Quantim image processing library[9] and had a mean of 0 and a variance of 1.5 for the groundwater test case and 1.0 for the unsaturated test case.

The hydraulic parameters were scaled with the exponential of these values according to the principle of Miller similarity[7]. A van Genuchten/Mualem model was used for the basic curve with the parameters of a medium sand:

Parameter	θ_s	θ_r	K_s	n	α	τ
Value	0.34	0.0	40.0 cm/h	2.0	5.0	0.5

For the groundwater test case Dirichlet boundary conditions were used at the west and east boundary imposing an average pressure gradient of 1 m/m. The potential at the lower edge of the east boundary was set to the size of the domain (thus the whole domain was water saturated). No-flow boundary conditions were used at all other boundaries. Initial condition was a full-saturation at hydraulic equilibrium. The size of one grid element was 0.1 m in all directions. We tested weak scalability in each step doubling the size of the domain using 80^3 grid cells per process resulting in 147 billion unknowns on 287'496 processes.

For the unsaturated test case no-flow boundary conditions were used at all side boundaries, a constant potential of zero was given at the bottom and a constant flux of 2 mm/h was applied at the top. Initial condition was hydraulic equilibrium with a potential of 0

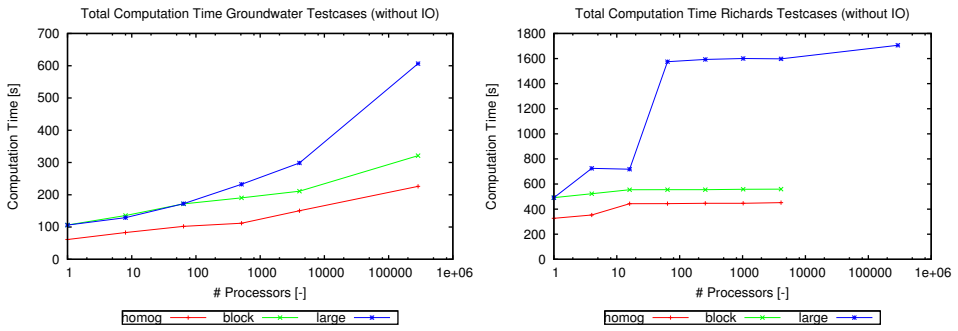


Figure 1. Total computation time without I/O for the different groundwater (left) and unsaturated (right) test scenarios.

cm at the bottom. The size of a grid element was $0.01 \times 0.01 \times 0.01$ m. One time step of one hour was simulated. In the weak scalability test we used $64 \times 64 \times 128$ grid cells per process. The domain was only increased in the horizontal direction. Up to 294'849 processes were used.

4 Results

Besides performing simulations for the test cases we were able to speed up the data output (by a factor of two) and to get rid of a long delay at the start of the program resulting from dynamical linking of the system libraries. With static linking the delay vanished completely.

The linear solver with an algebraic multigrid preconditioner has an expected complexity of $O(N \log N)$, where N is the number of unknowns. The computation time (Fig. 1) shows for the groundwater test case the expected linear rise in a logarithmic plot of total computation time against the number of unknowns/processes. For the large scenario there is a non-linear increase in the number of iterations needed at the beginning until the full structure (which was periodic with a length of 1024^3) is resolved. Afterwards the computation time rises again linear, but with a higher slope.

For the unsaturated test case the total computation time is even constant for all scenarios as soon as the structure is fully resolved. This is a consequence of the strict diagonal dominance of the matrix due to the additional time derivative term.

In an analysis of the computation time per step of defect calculation, matrix assembly, generation of the coarse grid hierarchy of the algebraic multigrid solver (coarsening) and application of the linear solver, respectively, most components of the code scale perfectly with a parallel efficiency close to one (Fig. 2). There is a decrease in the efficiency of the linear solver per step for the unsaturated laplace scenario. However, the iteration time for the laplace case is at the beginning smaller than for all other scenarios. With increasing number of processes/unknowns the iteration time approaches the time needed in the other scenarios.

For the groundwater test case the remaining unknowns are redistributed to fewer processes if the number of unknowns becomes too small. Finally on the coarsest level the

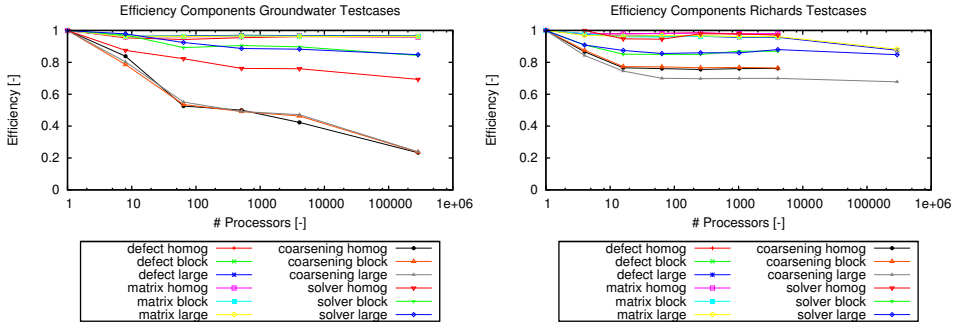


Figure 2. Efficiency per step of the different components of the code for the groundwater (left) and unsaturated (right) test scenarios.

system was solved exactly with SuperLU on one processes. The efficiency of this redistribution was difficult to achieve as ParMETIS could not be run in parallel (see above) and thus a sequential complexity was introduced. However, after an initial steep decrease it seems to scale perfectly well reaching an efficiency of 25 per cent at 287.496 processes.

For the unsaturated test case the redistribution was not necessary. The coarsening was performed as long as possible and the remaining linear equation system was solved iteratively with a parallel BiCGStab solver. This resulted in a much better efficiency of the coarsening but was only possible as the flow was mainly vertical with little horizontal coupling and the unknowns in the vertical direction were always completely on one process.

The reading of structures with as many points as the unknowns in the computation (up to 150 billion) was performed with parallel HDF5 (only for the block and large scenarios). For the block scenarios the structure was read sequentially by one process and broadcasted to the other processes, for the large scenario each process read its own hyperslab. For the unsaturated large scenario this took 1559 seconds, which is not really satisfying. The efficiency for block scenarios was better (Fig. 3). For the groundwater large scenario a pre-partitioned structure data in the SIONlib[5] format was used which reduced the time for structure input to 93 seconds.

Output was also written with the SIONlib library. The output time increased strongly with the number of processes. However, this was also due to the limited bandwidth of the I/O subsystem. For the unsaturated large scenario with 294'849 cores for the output of the potential 2.3 Terrabyte were written in 123 seconds, which corresponds to 19 GB/s which is close to the system limit. For the groundwater large scenario with 287'496 cores 8.8 Terrabyte of data was written (additionally including the flux field, the RT0-coefficients of the flux field and the volumetric water content) in 568 seconds (corresponding to 15 GB/s).

Crucial for the good scalability of our model and solver are

- a domain decomposition resulting in a communication pattern, which requires mostly local communications with few neighbour processes

- accumulation of communication so that only few large messages are sent instead of many small ones

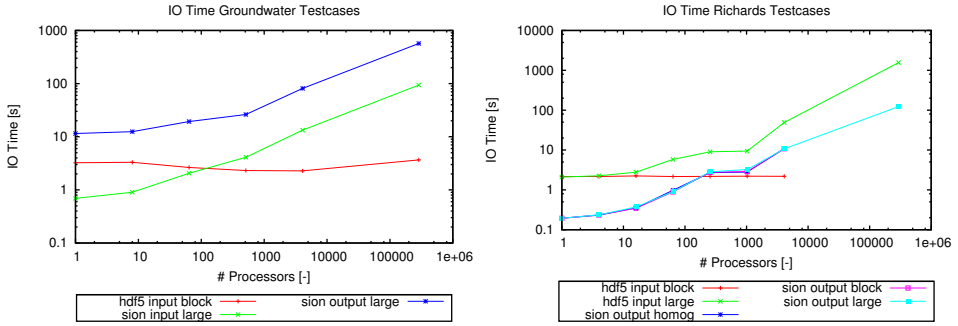


Figure 3. Time for input of structures with and the output of the results for the groundwater (left) and unsaturated (right) testcases. Please note the double-logarithmic scaling.

- the good scalability of the algebraic multigrid solver with the number of unknowns
- the use of IO libraries really exploiting the features of the parallel file system (GPFS)

A minor bottle neck remains the sequential computation of the new data decomposition with ParMETIS as the parallel version needs too much memory.

Acknowledgements

We thank the Jülich Supercomputing Centre for the possibility to take part in the Jülich BlueGene/P Extreme Scaling Workshop 2011. This enabled us to test our code on more processors than ever and to discover and fix some scalability bottle necks.

References

1. B. Mohr and W. Frings. Jülich BlueGene/P Extreme Scaling Workshop 2011, Technical Report, FZJ-JSC-IB-2011-2, p. 21-26, 2011.
2. M. Blatt. *A Parallel Algebraic Multigrid Method for Elliptic Problems with Highly Discontinuous Coefficients*. PhD thesis, Ruprecht–Karls–Universität Heidelberg, 2010. `http://www.ub.uni-heidelberg.de/archiv/10856/`.
3. M. Blatt and P. Bastian. On the generic parallelisation of iterative solvers for the finite element method. *Int. J. Comput. Sci. Engrg.*, 4(1):56–69, 2008.
4. J. W. Demmel, J. R. Gilbert, and X. S. Li. An asynchronous parallel supernodal algorithm for sparse gaussian elimination. *SIAM J. Matrix Anal. Appl.*, 20:915–952, July 1999.
5. W. Frings. SIONlib: Scalable I/O library for parallel access to task-local files. `http://www2.fz-juelich.de/jsc/sionlib`.
6. G. Karypis and V. Kumar. A parallel algorithm for multilevel graph partitioning and sparse matrix ordering. *J. Parallel Distrib. Comput.*, 48(1):71–95, 1998.
7. E. E. Miller and R. D. Miller. Physical theory for capillary flow phenomena. *J. Appl. Phys.*, 27:324–332, 1956.

8. H. A. van der Vorst. BI-CGSTAB: a fast and smoothly converging variant of bi-cg for the solution of nonsymmetric linear systems. *SIAM J. Sci. Stat. Comput.*, 13:631–644, March 1992.
9. H.-J. Vogel. Quantim4, C/C++ library for scientific image processing. `http://www.quantim.ufz.de`.
10. U. M. Yang. On the use of relaxation parameters in hybrid smoothers. *Numer. Linear Algebra Appl.*, 11(2–3):155–172, 2004.

Highly Parallel Geometric Multigrid Algorithm for Hierarchical Hybrid Grids*

Björn Gmeiner, Tobias Gradl, Harald Köstler, and Ulrich Rüde

Chair for System Simulation
Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), 91058 Erlangen, Germany
E-mail: {bjoern.gmeiner, tobias.gradl, harald.koestler, ruede}@informatik.uni-erlangen.de

Current supercomputers are approaching a million cores and their compute power and amount of memory enable us to solve linear systems with more than 10^{12} degrees of freedom. However, this forces us to partition our problem into a large number of sub-problems that can be treated in parallel. Parts of the algorithm that do not permit such high degrees of parallelism thus easily become a bottleneck. Additionally, the performance analysis and debugging of programs for such a high number of cores become challenging tasks in themselves. Our Hierarchical Hybrid Grids framework is capable of solving elliptic partial differential equations discretized with finite elements on a compromise between structured and unstructured grids by a geometric multigrid method. It is designed to run highly parallel and was adapted within this project to run on the JUGENE located in Jülich. We present scaling results and discuss the specifics of an efficient implementation of our software on Blue Gene/P systems.

1 Introduction

A variety of applications have a high demand on algorithms which are able to provide very high spatial resolutions finite element (FE). Large-scale example include seismic sea wave and earthquake simulations or weather predictions. In much smaller scales, direct numerical simulations provide insight to highly turbulent phenomena in fluid mechanics. Here, it is only possible to calculate a domain of some cubic centimeters at very high Reynold numbers. But there are also interesting issues at other scales, e.g. in acoustics: An important design goal for concert halls is to acquire excellent room acoustics. A direct approach to capture the acoustics in such buildings is to solve the pressure equation. If an average concert hall is simulated at a reasonable resolution, we already can fill the complete memory of today's largest supercomputers.

A major aim of our report is to show that it is still possible to design a relatively flat, but still efficient multigrid (MG) algorithm on current highly parallel supercomputers. Flat means in this context, that we have a multigrid algorithm with up to seven grid levels and thus a relatively large number of unknowns on the coarsest grid.

Our test machine is a Blue Gene/P cluster located in Jülich, which has 73 728 nodes. Each node is equipped with four compute cores. Besides other software approaches e.g.[1–3], our framework Hierarchical Hybrid Grids (HHG) implements, a highly parallel, multigrid variant for such machines.

We present the HHG approach on Blue Gene/P, which is different in terms of its architecture and degree of parallelism. The next section introduces the problems with the coarsest grids and ways to treat it. Furthermore, some parallel issues are discussed, which arose in our implementations.

*Reprint from NIC Symposium 2012, NIC Series Volume 45, p. 323–330, ISBN 978-3-89336-758-0.

Section three presents strong and weak scaling results. In the strong scaling, we try to push our MG to its limits on the underlying computer architecture. We achieved a speedup of 50 for a scaling from N to 96*N cores. The weak scaling shows a good parallel efficiency up to $292\,912$ cores, while the overhead for calculations on the coarsest grid stays reasonable.

1.1 Multigrid Algorithm

Discretizing a second-order differential equation with finite elements leads to a system of equations. An important property of the system matrix is sparsity. The sparsity pattern reduces the number of operations *per iteration* over the domain for iterative solvers to a complexity of O(N) for N unknowns. However, an increasing number of iterations are required for solvers like Gauss-Seidel (GS) or Conjugated Gradient (CG) with growing problem size. MG is a strategy to change this behaviour, such that the number of operations to solve a system is linearly dependent on the number of unknowns. This allows us to solve large systems in reasonable time. In the following, we give a quick overview on the multigrid method. For an introduction, we refer the reader to Ref. 4.

An observation for local acting iterative smoothers, like GS or Jacobi solvers, is that they smooth high frequent errors very well. In contrast, low frequent errors are reduced extremely slow. In order to resolve this shortcoming, MG uses a second effect: Let us consider a function on a fine grid, which has low frequencies. When we transfer this function to a coarser grid, the frequencies of the function increase with respect to the mesh size. Thus, we have changed the frequency of a function by changing its discretization. In MG this idea is applied to low frequencies of the error by transferring the error to coarser grids.

For the linear case, we can use the *error equation* (1) to calculate the error e_k on the coarser grids, where N_k is the discretization matrix, r_k the residual, f_k the right hand side, and v_k the approximated solution. The subindex k denotes the level of the grid. For a two-grid case k is the fine grid and $k-1$ is the coarse grid. Afterwards the solution is corrected by the error on the fine grid (see Fig. 1).

$$N_k e_k = r_k = f_k - N_k v_k \tag{1}$$

1.2 Hierarchical Hybrid Grids

The HHG framework[5,6] uses a hybrid discretization strategy by combining structured and unstructured grids. A coarse input FE mesh is organized into the grid primitives vertices, edges, faces, and volumes. This grid is unstructured and thus provides geometric flexibility. The primitives are then refined in a structured way, resulting in semi-structured meshes (see Fig. 2). The generated structured regions are stored in a directly addressed way into the memory to allow an efficient execution. Moreover, the regularity of the resulting grids may be exploited in such a way that it is no longer necessary to explicitly assemble the global discretization matrix. In particular, given an appropriate input grid, the discretization matrix may be defined implicitly using stencils that are constant for each structured patch. Hence, HHG is designed to have low memory consumption as well as hardware efficient execution and a high degree of parallelism. This approach allows to solve elliptic partial differential equations with a very high resolution. HHG supports different point- and line-wise relaxation schemes for the smoothing procedure.

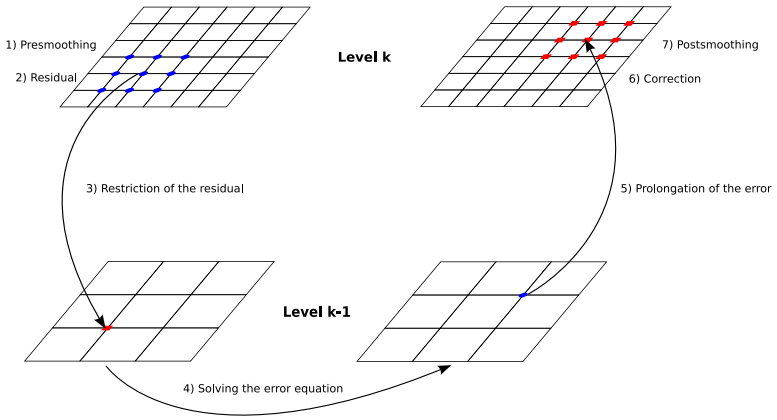


Figure 1. Multigrid two-grid cycle.

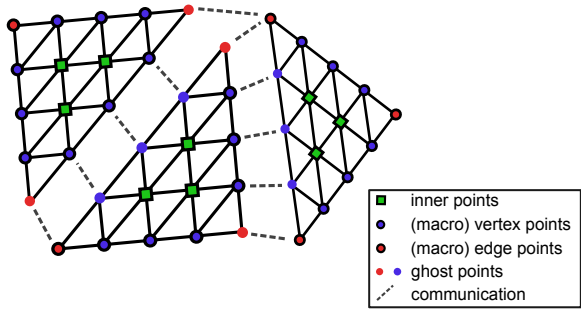


Figure 2. Space partitioning in HHG primitives.

2 Grid Partitioning and the Coarsest Grids

From the theoretical point of view, MG has a linear complexity of $O(N)$ with respect to the number of unknowns for sparse systems. For serial runs, this observation can also be made in practical use. In parallel settings we are able to achieve this behaviour up to a certain degree that seems to be limited by the increasing amount of communication cost for the coarsest levels. Estimates for the decreasing volume to surface ratio for multigrid hierarchy are given in[7]. As an example, let us assume we want to utilize $300,000$ cores, having one process per core. With static grid partitioning the coarsest grid would consist of $300,000$ elements. At the latest at this stage there are two possibilities to treat this grid: Proceed with the construction of new coarser grid levels by collocating elements on fewer number of cores (agglomeration) or stop at this stage and apply any iterative or direct solver[8].

In our MG algorithm, we decided to go for the second approach. This strategy is also known as flat multigrid, truncated cycle, or U-cycle. For a regular tetrahedral grid it is too drastic to end up with one element per process, like in our example. This would lead to up to four local unknowns per process and at least 44 ghost points. HHG refines each input

element twice to generate the coarsest multigrid level, so that the minimal size per process is 64 elements (up to 35 local unknowns), which provide a reasonable volume to surface ratio.

In the case of grid partitioning two rules have to be considered, when dealing with a very high degree of parallelism. First, when constructing the grids in a setup phase, often global acting algorithms have to be used, e.g. to find communication neighbours or to perform global numbering. Let us assume an operation between two input values costs 10 processor cycles and let n_P be the number of processors. Algorithms of complexity of $O(n_P)$ and $O(n_P log_2(n_P))$ would need $3 \cdot 10^6$ and $5.5 \cdot 10^7$ cycles for $n_P = 3 \cdot 10^5$, respectively. A modern processor can perform these operations in far below one second. However, while for an $O(n_P^2)$ class it is often still possible with $n_P = 10 \cdot 10^4$ to treat our example in 10^9 cycles (around one second), $n_P = 3 \cdot 10^5$ processors would require 10^{11} cycles. This is in the range of a minute on current hardware. It can be acceptable, but our assumptions are quite optimistic. If there is more than one sub-grid per processor, an operation is more expensive, or thinking about the next generations of supercomputers a complexity of $O(n_P^2)$ has to be avoided by choosing an other algorithm or parallelization, if possible. In HHG we had to reduce the complexity for the mesh construction to $O(n_P log_2(n_P))$, since the setup times took much longer than the solving phase.

Second, we have to consider main memory requirements. Since we read in an unstructured input mesh to allow geometric flexibility, basic logical information of the mesh can easily grow up to more than hundreds of MB. This is already the case for a few million elements, which the coarsest grid might have when we use a flat multigrid algorithm. Thus one should consider to hold the coarse grid structure in main memory of each processor just in the setup phase, or to avoid that at all. Every instance of HHG only stores its own part of the coarsest grid to avoid memory problems.

Another issue is, if the problem solution on the coarsest grid of a truncated cycle can be approximated by a non-optimal solver in reasonable time. In this report, CG is used as a solver for the coarsest grid. To give an estimation of the necessary number of CG steps, we assume the following:

- The required number of CG steps is proportional to the diameter of the domain.
- We assume that high frequency error are eliminated by smoothing.
- One CG step for the diameter one with one coarse grid point is sufficient.

Consequently, our simple estimation for the required number of CG-steps n_{CG} for l multigrid levels and d dimensions (here $d = 3$) is

$$n_{CG} \approx \frac{\sqrt[d]{N}}{2^{l-1}}. \tag{2}$$

However, in practice this is only a rough estimation, multiplied by a constant ($c \approx 1$) because of the third assumption.

3 Scaling on JUGENE

Next we present weak and strong scaling results of our multigrid algorithm. Hereby push HHG to the limits of current CPU parallelism. All our calculations in this report were

performed on the super computer JUGENE in Jülich consisting of $73\,728$ compute nodes or sockets. Each socket is equipped with IBM's Blue Gene/P (BGP) quad-core processor. The following results are measured by using VN (virtual node) mode on the JUGENE. This means each node executes 4 tasks, sharing 2 GB of main memory.

$$\begin{aligned} \Delta u &= f \text{ in } \Omega\,, \\ u &= 0 \text{ on } \partial\Omega \end{aligned} \tag{3}$$

As a test problem we use Poisson's equation (3) with a right hand side f. It is a hard test problem from a performance and scalability point of view, since it has a low computation to communication ratio compared to more complex PDEs.

3.1 Strong Scaling

To have smaller runtimes per time step, it is interesting not to utilize the full main memory. Fixing the problem size, but increasing the number of processors is known as strong scaling. For computational steering it can be useful to reduce the runtime of one time step to the order of 0.05 seconds to achieve real-time behaviour. In an implicit time stepping, one time step per frame would be feasible. In this section, we want to evaluate the potential of a multigrid algorithm to reach this goal.

In our experimental setting, we solved $2.14 \cdot 10^9$ unknowns using 512 to $49\,152$ cores. For the strong scaling experiment, we use the following multigrid components:

- V(3,3) row-wise red-black GS smoothing,
- 60 steps of a CG solver to approximate the coarsest grid,
- Five multigrid levels,
- Direct coarse grid approximation,
- Linear interpolation.

It can be shown by local fourier analysis[9], that three pre- and post-smoothing steps are necessary to achieve the "classical" multigrid convergence of 0.1 for tetrahedral grids. Fig. 3 shows that HHG achieves a good strong scaling behaviour over a wide range of cores.

Increasing the number of cores by a factor of 96 we are able to reduce the initial time from 7.95 seconds per V-cycle to 0.16 seconds. A perfect strong scaling would result in 0.08 seconds per V-cycle. Thus we can clearly observe that the communication overhead impacts the performance. For the largest run, we have quite small memory arrays of around 383 KB for each variable (right hand side, unknowns, residual) including ghost points. In this data volume, the ghost points require 69 KB. Thus, the volume to surface ratio is quite small. Furthermore, latency of the messages has a larger impact, since more messages per time have to be sent. A significant part of the time (about 38%) is spent for the approximation on the coarsest grid.

With a similar setup, but using V(1,1) cycles and 40 CG steps on the coarsest grid, one cycle takes 0.07 seconds. For an implicit real-time application, additional time is required for a visualization pipeline (post-processing) and the time stepping itself. However, it should be possible to tune the solving further, by e.g. optimizing the coarsest grid solver and the other multigrid components, having less unknowns per core, or utilizing stronger processors. So, highly parallel real-time simulations with a multigrid algorithm seem to be challenging but feasible on current hardware.

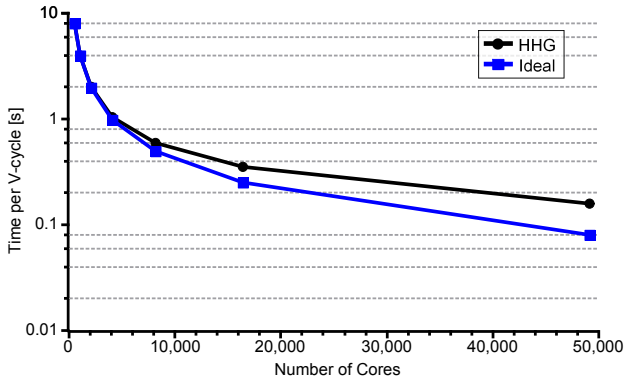


Figure 3. Strong scaling behaviour of HHG on PowerPC 450 cores of a Blue Gene/P located at Jülich. This test case was performed starting from 512 cores and solving a system of $2.14 \cdot 10^9$ unknowns.

3.2 Weak Scaling

This section discusses some effects of solving large linear systems by MG with up to $294\,912$ compute cores. The setup of the weak scaling experiments corresponds to the strong scaling from the previous section except for:

- the number of CG steps depends on the size of the coarsest grid,
- we apply a constant numbers of MG levels:
  - Six for up to $262\,144$ cores and seven for $294\,912$ cores,
- we have 12 structured regions per core for six MG levels and
  we have 1 structured region per core for seven MG levels.

The parallel efficiency is reflected in Tab. 1 by the time per V-cycle. Utilizing the whole machine, HHG achieved a parallel efficiency of 69.2%. In the limit HHG solved up to 10^{12} degrees of freedom (DoF).

When fixing the number of levels for V-cycles, the number of required CG steps on the coarsest grid grows from 15 to 180 CG steps. However, for the largest runs the time on the coarsest mesh is around 12.5% of the total time for one V-cycle only. A comparison of the work done on the coarsest grid between two different MG cycles is given in Fig. 4. Our prediction is calculated by Eq. 2. In our setup an F-cycle is very similar to the prediction. An F-cycles is similar to a full-multigrid cycle and requires a better coarse grid approximation than a V-cycle, i.e. for a V-cycle approximately only half of the CG-steps are necessary. Moreover, we have to keep in mind that the number of CG-steps between the restarts is different. Restart means that only the approximated solution of the previous step is available, but no additional information like previous search directions. In our case, there are five times more CG-restarts for the F-cycle than for the V-cycle. However, the measured number of CG-steps is in the same order of magnitude as predicted. One reason for deviations from the prediction in the measurements is the shape of the domain. At many points in Fig. 4, the domain in one or two directions is twice as large as in the others. Furthermore, we do not consider a spherically shaped domain.

The full machine run uses seven instead of six levels. In our semi-structured approach, a structured region cannot be shared by multiple processors. Thus, we are not able to

Cores	Struct. Regions	DoF $(\cdot 10^6)$	CG	Time (s)
128	1 536	535	15	5.64
256	3 072	1 071	20	5.66
512	6 144	2 142	25	5.69
1024	12 288	4 287	30	5.71
2048	24 576	8 577	45	5.75
4096	49 152	17 159	60	5.92
8192	98 304	34 326	70	5.86
16384	196 608	68 669	90	5.91
32768	393 216	137 355	105	6.17
65536	786 432	274 744	115	6.41
131072	1 572 864	549 555	145	6.42
262144	3 145 728	1 099 176	180	6.52
294912	294 912	824 365	110	3.80

Table 1. Weak scaling behaviour of HHG on PowerPC 450 cores of a Blue Gene/P at Jülich.

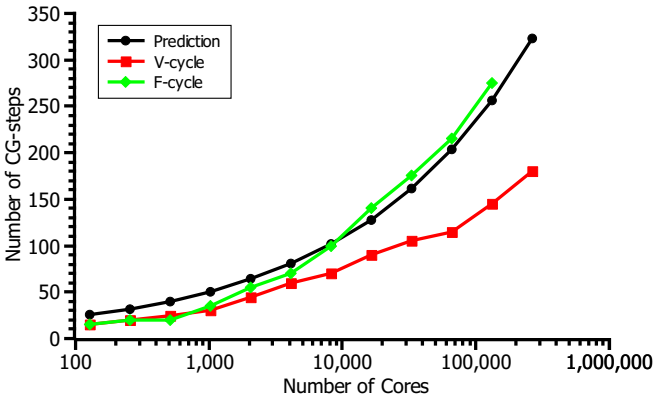


Figure 4. Required number of CG-steps on the coarsest grid per cycle with increasing numbers of compute cores.

utilize full main memory. The additional level reduces the number of CG iterations. The performance in terms of solved unknowns per second increases by 14% due to larger inner loops for the finest grids. Here, the overall achieved performance using the full machine is 59.8 TFLOP/s in the solving phase.

4 Conclusions and Future Work

We presented scaling results for geometric multigrid within the HHG software on the JUGENE supercomputer located at Jülich. We addressed scalability problems and communication overhead created by the coarsest grid in the multigrid hierarchy. A careful implementation results in excellent scalability results, i.e. the coarse grids do not seriously effect

the overall parallel performance. To this end we explored and analyzed the weak scaling of numerical experiments with up to 10^{12} unknowns.

Next we plan to compare our results to another geometric multigrid framework on different architectures. In this context hybrid parallelization and GPU acceleration could be interesting issues. Apart from that we are extending HHG to treat the Poisson problem occurring in vortex particle direct numerical simulations.

Acknowledgements

The work was supported by the Kompetenznetzwerk für Technisch-Wissenschaftliches Hoch- und Höchstleistungsrechnen in Bayern (KONWIHR) and the International Doctorate Program (IDK) within the Elite Network of Bavaria.

We would like to thank Jutta Docter and Bernd Mohr, Jülich Supercomputing Centre (JSC) for their excellent support during the largest runs.

We are grateful to the Jülich Supercomputing Centre for providing the computational resources on JUGENE.

References

1. H.J. Bungartz, M. Mehl, T. Neckel, and T. Weinzierl, *The PDE framework Peano applied to fluid dynamics: an efficient implementation of a parallel multiscale fluid dynamics solver on octree-like adaptive Cartesian grids*, Computational Mechanics, **46**, no. 1, 103–114, 2010.
2. R. Falgout, V. Henson, J. Jones, and U. Yang, *Boomer AMG: A Parallel Implementation of Algebraic Multigrid*, Tech. Rep. UCRL-MI-133583, Lawrence Livermore National Laboratory, 1999.
3. R.S. Sampath and G. Biros, *A Parallel Geometric Multigrid Method for Finite Elements on Octree Meshes*, SIAM Journal on Scientific Computing, **32**, 1361, 2010.
4. W. Briggs, V. Henson, and S. McCormick, *A Multigrid Tutorial*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, USA, 2nd edition, 2000.
5. B. Bergen, T. Gradl, F. Hülsemann, and U. Rüde, *A Massively Parallel Multigrid Method for Finite Elements*, Computing in Science and Engineering, **8**, no. 6, 56–62, 2006.
6. T. Gradl, C. Freundl, H. Köstler, and U. Rüde, *Scalable Multigrid*, in: High Performance Computing in Science and Engineering. Garching/Munich 2007, pp. 475–483, 2008.
7. F. Hülsemann, M. Kowarschik, M. Mohr, and U. Rüde, "Parallel geometric multigrid", in: Numerical Solution of Partial Differential Equations on Parallel Computers, pp. 165–208. Springer-Verlag, Berlin, Heidelberg, New York, 2005.
8. D. Xie and L.R. Scott, *The parallel u-cycle multigrid method*, in: Proceedings of the 8th Copper Mountain Conference on Multigrid Methods. Citeseer, 1996.
9. R. Wienands and W. Joppich, *Practical Fourier Analysis for Multigrid Methods*, vol. 5 of *Numerical Insights*, Chapmann and Hall/CRC Press, Boca Raton, Florida, USA, 2005.

Schriften des Forschungszentrums Jülich
IAS Series

---

1. **Three-dimensional modelling of soil-plant interactions: Consistent coupling of soil and plant root systems**
   by T. Schröder (2009), VIII, 72 pages
   ISBN: 978-3-89336-576-0
   URN: urn:nbn:de:0001-00505

2. **Large-Scale Simulations of Error-Prone Quantum Computation Devices**
   by D. B. Trieu (2009), VI, 173 pages
   ISBN: 978-3-89336-601-9
   URN: urn:nbn:de:0001-00552

3. **NIC Symposium 2010**
   Proceedings, 24 – 25 February 2010 | Jülich, Germany
   edited by G. Münster, D. Wolf, M. Kremer (2010), V, 395 pages
   ISBN: 978-3-89336-606-4
   URN: urn:nbn:de:0001-2010020108

4. **Timestamp Synchronization of Concurrent Events**
   by D. Becker (2010), XVIII, 116 pages
   ISBN: 978-3-89336-625-5
   URN: urn:nbn:de:0001-2010051916

5. **UNICORE Summit 2010**
   Proceedings, 18 – 19 May 2010 | Jülich, Germany
   edited by A. Streit, M. Romberg, D. Mallmann (2010), iv, 123 pages
   ISBN: 978-3-89336-661-3
   URN: urn:nbn:de:0001-2010082304

6. **Fast Methods for Long-Range Interactions in Complex Systems**
   Lecture Notes, Summer School, 6 – 10 September 2010**,** Jülich, Germany
   edited by P. Gibbon, T. Lippert, G. Sutmann (2011), ii, 167 pages
   ISBN: 978-3-89336-714-6
   URN: urn:nbn:de:0001-2011051907

7. **Generalized Algebraic Kernels and Multipole Expansions for Massively Parallel Vortex Particle Methods**
   by R. Speck (2011), iv, 125 pages
   ISBN: 978-3-89336-733-7
   URN: urn:nbn:de:0001-2011083003

8. **From Computational Biophysics to Systems Biology (CBSB11)**
   Proceedings, 20 - 22 July 2011 | Jülich, Germany
   edited by P. Carloni, U. H. E. Hansmann, T. Lippert, J. H. Meinke, S. Mohanty, W. Nadler, O. Zimmermann (2011), v, 255 pages
   ISBN: 978-3-89336-748-1
   URN: urn:nbn:de:0001-2011112819

9. **UNICORE Summit 2011**
Proceedings, 7 - 8 July 2011 | Toruń, Poland
edited by M. Romberg, P. Bała, R. Müller-Pfefferkorn, D. Mallmann (2011), iv, 150 pages
ISBN: 978-3-89336-750-4
URN: urn:nbn:de:0001-2011120103

10. **Hierarchical Methods for Dynamics in Complex Molecular Systems**
Lecture Notes, IAS Winter School, 5 – 9 March 2012, Jülich, Germany
edited by J. Grotendorst, G. Sutmann, G. Gompper, D. Marx (2012), vi, 540 pages
ISBN: 978-3-89336-768-9
URN: urn:nbn:de:0001-2012020208

The focus of this Winter School was on hierarchical methods for dynamical problems having primarily in mind systems described in terms of many atoms or molecules. One end of relevant time scales certainly is nonadiabatic quantum dynamics methods, which operate on the subfemtosecond time scale but influence dynamical events that are orders of magnitude slower. Examples for such phenomena might be photoinduced switching of individual molecules, which results into large-amplitude relaxation in liquids or photodriven phase transitions of liquid crystals. On the other end of the relevant time scales methods are important to investigate and understand the non-equilibrium dynamics of complex fluids, with typical time scales in the range from microseconds to seconds. Examples are the flow of polymer solutions, or the flow of blood through microvessels.

The Lecture Notes contain state-of-the-art information on methodological foundations and methods coming from materials science, soft matter, life science and fluid dynamics. In addition to introducing discipline-specific methods, modern numerical algorithms and parallel programming techniques are presented in detail.

This publication was edited at the Jülich Supercomputing Centre (JSC) which is an integral part of the Institute for Advanced Simulation (IAS). The IAS combines the Jülich simulation sciences and the supercomputer facility in one organizational unit. It includes those parts of the scientific institutes at Forschungszentrum Jülich which use simulation on supercomputers as their main research methodology.

IAS Series
Volume 10
ISBN 978-3-89336-768-9

JÜLICH
FORSCHUNGSZENTRUM