

Multiscale Modelling Methods for Applications in Materials Science

Lecture Notes

edited by Ivan Kondov, Godehard Sutmann

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

Multiscale Modelling Methods for Applications in Materials Science

edited by
Ivan Kondov, Godehard Sutmann

CECAM Tutorial, 16 – 20 September 2013
Forschungszentrum Jülich
Lecture Notes

organized by
Karlsruhe Institute of Technology
Forschungszentrum Jülich

Schriften des Forschungszentrums Jülich

IAS Series

Volume 19

ISSN 1868-8489

ISBN 978-3-89336-899-0

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 2013

Schriften des Forschungszentrums Jülich
IAS Series Volume 19

ISSN 1868-8489
ISBN 978-3-89336-899-0

Persistent Identifier: [urn:nbn:de:0001-2013090204](http://nbn:de:0001-2013090204)
Resolving URL: <http://www.persistent-identifier.de/?link=610>

Neither this book nor any part of it may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, microfilming, and recording, or by any information storage and retrieval system, without permission in writing from the publisher.

Preface

Current advances in multiscale modelling of materials promise scientific and practical benefits including simple physical interpretation based on analysis of the underlying sub-models, as well as an improved computational scaling and acceptable amount of produced data, which make the simulation of large and complex real-world materials feasible. These developments give rise to an unprecedented predictive power of multiscale models allowing a reliable computation of macroscopic materials properties from first principles with sufficient accuracy. However, the development of methods which efficiently couple multiple scales in materials science is still a challenge, since (i) proper coupling schemes have to be developed which respect the physical and chemical descriptions on the different scales; (ii) boundary conditions for e.g. mechanics, thermodynamics or hydrodynamics have to be respected and (iii) error control and numerical stability have to be guaranteed. In addition to these physical and numerical requirements, multiscale modelling poses serious challenges to the practical realization of coupled applications due to the complex organization of interfaces between the sub-models and heterogeneity of computational environments. Therefore, both integrative and coordination actions, such as the Max-Planck Initiative *Multiscale Materials Modelling of Condensed Matter*, FP7 projects MAPPER and MMM@HPC, or the CECAM node MM1P *Multiscale Modelling from First Principles*, have been initiated which bundle the expertise of different groups (in fields such as quantum chemistry, molecular dynamics, coarse-grained modelling methods and finite element analysis) and move forward both the theoretical understanding as well as the practical implementation of a multiscale simulation environment.

The knowledge of and the experience with novel multiscale techniques, such as sequential/hierarchical modelling or hybrid methods, as well as modelling tools should be disseminated to a larger number of groups in the materials science and physics community. Since the topic of *multiscale modelling in materials science* is still underdeveloped in university courses, it is essential to provide tutorials by established experts to young scientists working in multiscale simulations or starting in the field. In particular, postgraduate students and postdoctoral researchers entering the field are addressed by this tutorial.

Past winter schools like *Multiscale Simulation Methods in Molecular Sciences* (2009) or *Hierarchical Methods for Dynamics in Complex Molecular Systems* (2012), organized at Forschungszentrum Jülich focused on dynamical aspects in molecular systems on different time scales. They addressed non-adiabatic quantum dynamics, including descriptions of photo-induced processes, up to non-equilibrium dynamics of complex fluids, while still keeping the atomistic scale in the classical, quantum mechanical and mixed quantum-classical descriptions. In the present tutorial *Multiscale Modelling Methods for Applications in Materials Science* we emphasize on methodologies encompassing not only the dynamical aspects but also steady-state or/and equilibrium properties on the meso- and macroscopic scales treated for example by coarse-grained and finite-elements methods. Moreover, this tutorial predominantly addresses modelling of systems with modern high-profile applications with industrial importance, such as materials for energy conversion and storage and for next generation electronics, which are not restricted to molecular systems. The lecture notes collected in this book reflect the course of lectures presented in the tutorial and include twelve chapters subdivided into two parts. The lecture notes in the first part *Methods* provide a comprehensive introduction to the underlying methodology, which

assume some background knowledge in various of the theoretical methods and computational techniques employed in multiscale modelling, e.g. quantum mechanics, statistical physics, theoretical chemistry, theoretical solid state physics, Monte Carlo and molecular dynamics simulations. The lectures particularly explain the physical interrelations between different scales and introduce best practices in combining the methods. The contributions in the second part of the lecture notes, entitled *Applications and Tools* illustrate the combination of different approaches to treat high-profile applications, such as coarse graining of polymers and biomolecules, and modelling of organic light-emitting diodes, electrochemical energy storage devices (Li-ion batteries and fuel cells) and energy conversion devices (organic electronics and carbon nanodevices). Furthermore, an introduction is given to modern tools and platforms for the technical implementation of such applications, e.g. to UNICORE or Simulink®.

The multiscale modelling often requires novel implementations and practical approaches for performing computer simulations in an increasingly complex software environment. In contrast to standard approaches that have been used for many years in the community, these new approaches, based e.g. on the UNICORE middleware, expose the physics aspects of the models to the modelling scientist while hiding the technical complexity of the underlying computer infrastructures. First practical experiences with such an approach is essential to strengthen the acquired knowledge during the lecture parts. The MMM@HPC project (www.multiscale-modelling.eu) has developed a UNICORE-based integrated platform for multiscale materials modelling which is demonstrated during one of the hands-on sessions, organized in cooperation with the MMM@HPC project. In further hands-on sessions the gained knowledge from the lectures is practiced for selected applications.

We cordially thank the lecturers for their great effort in writing the lecture notes in due time so that the book could be edited and printed in advance to the tutorial bringing maximal benefit for the participants. Further thanks go to the instructors who prepared the exercises for the hands-on sessions. We are very grateful to the tutorial's secretaries Elke Bielitz and Britta Hoßfeld as well as to Monika Marx for compiling the lecture notes manuscripts and producing a high-quality book. We gratefully acknowledge financial support from Forschungszentrum Jülich, CECAM (www.cecarn.org) and the MMM@HPC project funded by the 7th Framework Programme of the European Commission within the Research Infrastructures with grant agreement number RI-261594.

Karlsruhe and Jülich,
September 2013

Ivan Kondov
Godehard Sutmann

Contents

Methods

Introduction to Multiscale Modelling of Materials

<i>James A. Elliott</i>	1
1 Introduction	1
2 Molecular Modelling Methods	4
3 Coarse-Grained Modelling Methods	12
4 Conclusions and Outlook	17

Introduction to Modelling, Scalability and Workflows with DL_POLY

<i>Ilian T. Todorov, Laurence J. Ellison, Michael A. Seaton, William Smith</i>	21
1 Introduction	21
2 Software	21
3 Molecular Structures	22
4 Force Field	22
5 Integration Algorithms	24
6 Parallelisation	25
7 Electrostatics	27
8 Scalability and Performance	28
9 I/O Files and Performance	29
10 GridBeans and Workflows	31
11 Concluding Remarks	36

Atomistic Simulations Using the Approximate DFT Method DFTB+: Applications to Nanomaterials and Bio-Systems

<i>Thomas Frauenheim, Bálint Aradi</i>	41
1 Introduction	41
2 Theory of DFTB	42
3 Example: Bulk Amorphous Oxides	47
4 Summary	53

Wavelets For Electronic Structure Calculations

Thierry Deutsch, Luigi Genovese

55

1	Introduction	55
2	Atomistic Simulations	55
3	Pseudopotentials	63
4	Kohn-Sham DFT with Daubechies Wavelets	64
5	Overview of the Method	71
6	Treatment of Kinetic Energy	72
7	Calculation of Hartree Potential	78
8	Calculation of Forces	90
9	Preconditioning	92
10	Orthogonalization	93
11	Parallelization	95
12	Performance Results	96
13	Conclusions	98

Elmer Finite Element Solver for Multiphysics and Multiscale Problems

Mika Malinen, Peter R  back

101

1	Introduction	101
2	Solving a Coupled Problem with the Solver of Elmer	102
3	The Key Capabilities of the Solver	106
4	Applying Elmer to Multiscale Problems	110
5	Concluding Remarks	113

Modeling Charge Distributions and Dielectric Response Functions of Atomistic and Continuous Media

Martin H. M  ser

115

1	Introduction	115
2	General Aspects of Charge-Equilibration Approaches	116
3	Bottom-Up Motivation of Charge-Equilibration Models	121
4	Top-Down Approach to Charge-Equilibration Models	123
5	Applications	129
6	Conclusions	133

Applications and Tools

Systematic Coarse Graining of Polymers and Biomolecules

<i>Roland Faller</i>	135
1 Introduction	135
2 Fundamentals and Theoretical Basis of Different Coarse-Graining Techniques	137
3 Examples	141
4 Conclusions	148
5 Acknowledgements	148

Theory and Simulation of Charge Transport in Disordered Organic Semiconductors

<i>Peter A. Bobbert</i>	151
1 Hopping Transport	151
2 The Disorder Energy Landscape	153
3 The Master Equation	154
4 Master-Equation Calculations for a Complete Device	155
5 Master-Equation Calculations with Periodic Boundary Conditions	156
6 Solving the Master Equation Iteratively	157
7 The Drift-Diffusion Equation	158
8 Monte Carlo	159
9 Example: a Hole-Only Device	160
10 Transients	161
11 Uncorrelated or Correlated Disorder?	163
12 Random-Resistor Network	163
13 Percolation Theory and Scaling Ansatz	164
14 Determining the Critical Conductance	166
15 Application of the Scaling Expression to Different Hopping Models	168
16 Effect of Lattice Disorder	171
17 Carrier-Concentration Dependence of the Mobility	173
18 Temperature Dependence of the Mobility	174
19 Monte Carlo Modeling of Electronic Processes in a White Multilayer OLED	175
20 Concluding Remarks	179

Multiscale Modeling Methods for Electrochemical Energy Conversion and Storage

<i>Alejandro A. Franco</i>	183
1 Introduction	183
2 Modeling Experiments and Experimenting Models	194
3 Multiscale Models of EPGs: Examples and Practice	205
4 Conclusions and Challenges	235

Multiscale Transport Methods for Exploring Nanomaterials and Nanodevices		251
<i>Frank Ortmann, Stephan Roche</i>		
1	Introduction: State of the Art of Computational Approaches for Nanodevice Simulation	251
2	Kubo-Transport Methodology	254
3	Landauer Transport Approach	258
4	Ab initio Methods for Material Parameters	259
5	New Electronics Features of Chemically-Modified Graphene-Based Materials: Mobility Gaps	261
6	Limits of Ballistic Transport in Silicon Nanowires	263
7	Organic Semiconductors	266
8	Conclusion and Perspective	271
9	Acknowledgements	272
Electronic Structure of Organic/Organic Interfaces: A Quantum-Chemical Insight		277
<i>Jérôme Cornil</i>		
1	Introduction	277
2	Interface Dipole: Charge Transfer and Polarization Components	278
3	TTF/TCNQ Model Systems	279
4	Extended TTF-TCNQ Stacks	284
5	C ₆₀ / Pentacene Complexes	285
6	Energy Landscape around Organic/Organic Interfaces	288
7	Conclusions	292
UNICORE Rich Client User Manual		295
<i>Bastian Demuth, Lara Flörke, Björn Hagemeier, Michael Rambadt, Daniel Mallmann, Mathilde Romberg, Rajveer Saini, Bernd Schuller on behalf of the UNICORE Team</i>		
1	Introduction	295
2	A Brief History of UNICORE	296
3	Installation and Startup	296
4	Basic Usage Guide	297
5	Concluding Remarks and Further Information	317
6	Glossary	318

Introduction to Multiscale Modelling of Materials

James A. Elliott

Department of Materials Science and Metallurgy,
University of Cambridge, 27 Charles Babbage Road,
Cambridge, CB3 0FS, UK
E-mail: jae1001@cam.ac.uk

As computing power continues to increase at a relentless pace,¹ it is tempting to consider the simulation of large and/or complex systems using brute force atomistic simulation methods alone. However, even extrapolating from current state-of-the-art methodologies, it would still take well over a century of continued exponential growth in computing resources to achieve parity with ‘real time’ simulations of experimental systems of macroscopic size and, in any case, the sheer amount of data produced would overwhelm any attempt at detailed scientific analysis. Therefore, it is imperative that we now seek to exploit the regions of overlap between well-established techniques for electronic structure calculations, molecular dynamics, mesoscopic simulations and continuum modelling to allow efficient multiscale simulations of increasingly complex condensed phase systems. In this lecture, I will introduce the concept of multiscale modelling in Materials Science, in which there have been significant technical and scientific advances over the last decade,² enabling novel fields of application from nanotechnology to biomineralization.

In these accompanying notes, I will first briefly introduce some of the basic techniques used in multiscale modelling, including both molecular and mesoscopic particle dynamics, elementary principles of coarse-graining, and finite element analysis, before focusing, in the lecture itself, on several recent research highlights from my own group’s work. The first focus area will be pharmaceutical³ and biocomposite materials,⁴ where multiscale models have been applied to formulation of powders for drug tableting, and to investigate the mechanical properties of biomineral structures such as bone. In particular, the structure and properties of the collagen matrix depend greatly on confinement by solvent and mineral phase. The second area will be carbon nanomaterials, where I will present a model for predicting the strength of yarn-like carbon nanotube fibres,⁵ and relate this to earlier molecular dynamics simulations of single, double and multi-wall nanotube bundles under hydrostatic pressure.⁶ These show bundles containing nanotubes with a range of geometries ranging from cylindrical to fully collapsed, depending on diameter and number of walls. I will describe the implications for shear stress transfer between nanotubes in bundles in the context of improving mechanical properties of macroscopic assemblies of nanotubes. I will conclude by discussing the future potential of multiscale modelling in materials science over next 5 years.

1 Introduction

One of the major driving forces for the increasing usage and applicability of computational modelling in the physical and biological sciences over last decade has been the exponential growth in available computing power, often represented in the form of Moore’s law, which originally related to the observed doubling time of the number of transistors in microprocessors (as shown in Figure 1), although similar relationships can also be found for many other performance metrics, such as memory size, storage capacity and cost per floating point operation. Of course, Moore’s ‘law’ is simply an observation, rather than a law of nature, but with the recent advent of multiple core CPUs and low-cost, massively parallel GPUs for scientific computing there is evidence that the rate of increase of computing

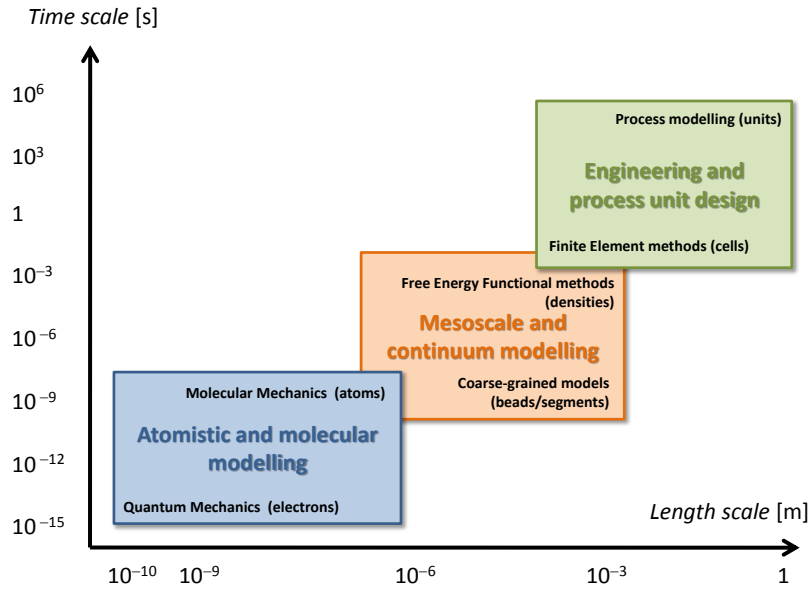


Figure 2. Hierarchy of multiscale modelling techniques as function of their applicable ranges of time and length scales, with fundamental entities (electrons, particles, etc.) given in parentheses. Adapted from Elliott (2011).²

simplifying assumptions of periodicity or long-range order. At the other end of the scale are engineering and process models, which are essentially based on solution of continuum partial differential equations, and can encompass macroscopic systems. Again, these are very mature, and routinely applied. However, there is a large gap between these two regimes, which has increasingly become filled by what are known as mesoscale methods, that include coarse-grained discrete particle simulations (where each particle represents a group of atoms) and density functional methods based on free energies (*e.g.* phase field method), although there is clearly some overlap between classification of such methods depending on the exact application. This field of mesoscale modelling is somewhat less mature, and there is no unique prescription for how to move from lower level to higher levels. In some cases, it is best to pass parameters calculated from quantum or atomic scale to macroscopic models (*hierarchical* approach), and in others it is better to carry out explicitly linked simulations (*hybrid* approach).

In the remainder of these lecture notes, I will briefly outline some of the basic simulation algorithms in the hierarchy shown in Figure 2, and discuss their relative advantages and disadvantages. These will be illustrated in the lecture by some real examples from research work in my group. However, I will not discuss in detail methods for calculating potential energies and gradients thereof, either by electronic structure methods (*e.g.* density functional or molecular orbital theory) or by using classical parameterized force fields, except when required to illustrate the applications of the algorithms themselves.

2 Molecular Modelling Methods

The origins of the so-called ‘molecular mechanics’⁷ (MM) approach to atomistic modelling, whereby classical, semi-empirical potential energy functions are used to approximate the behaviour of molecular systems, can be rationalised in part by considering the history of computer simulation as an extension of the tradition of mechanical model building that preceded it. The crystallographer J. D. Bernal describes in his Bakerian Lecture⁸ how, during the early 1950s, he built a structural model of a simple monatomic liquid, which at that time could not be described by existing theories of solids and gases, from an array of spheres randomly coordinated by rods of varying length. However, the emergence of mechanical, and subsequently electronic, digital computers enabled a much less labour intensive approach to modelling.

2.1 Molecular dynamics (MD)

In 1957, Alder and Wainwright published the first computer simulation of ‘hard’ spheres moving in a periodic box.⁹ Hard, in this sense, means that the spheres were forbidden from overlapping, rather like macroscopic steel ball bearings. Although this might not seem like a very realistic model for a liquid, these simulations eventually lead to the important (and initially controversial) conclusion that it is the harsh short-range repulsive forces between atoms in a liquid that are primarily responsible for the freezing transition, whereas the influence of the longer range attractive forces is somewhat less important. Nowadays, even though we can carry out MD simulations of complex macromolecules and charged particles with continuously varying, more realistic interaction potentials, the underlying mechanical analogies of the formalism are still evident. However, the algorithms for simulating systems evolving with a continuous potential interaction are rather different from those used in the first ‘impulsive’ MD simulations developed by Alder and Wainwright.

2.1.1 Impulsive and continuous-time conservative MD

In a system with hard particles, the dynamics evolves ballistically between particle impacts, with a characteristic time that depends on the frequency of collisions, τ , which is around 0.2 ns for Ar at 298 K. However, for a system where the force on each particle can be calculated as a gradient of a continuous potential energy function (or from the Hellman-Feynman theorem in the case of *ab initio* MD), then we can solve Newton’s equations of motion numerically using some finite difference scheme, a process that is referred to as integration. This means that we advance the system by some small, discrete time step, Δt , recalculate the forces and velocities, and then repeat the process iteratively. Provided that Δt is small enough, this produces an acceptable approximate solution to the continuous equations of motion.

However, the choice of time step length is crucial: too short and phase space is sampled inefficiently, too long and the energy will fluctuate wildly and the simulation may become catastrophically unstable. The instabilities are caused by the motion of atoms being extrapolated into regions where the potential energy is prohibitively high, *e.g.* if there is any atomic overlap. A good rule of thumb is that the time step should be an order of magnitude less than the period of the fastest motion in the system, which for macromolecules

is usually bond stretching (*e.g.* C–H stretch period is approximately 11 fs, so a time step of 1 fs is often used). Clearly, we would like to make the time step as long as possible without producing instability, as this gives us the largest amount of simulated time per unit of computer time.

The first MD simulations of a fluid of Lennard-Jones particles were carried out by Verlet in 1967,¹⁰ using an integration algorithm that is still widespread today and known as the Störmer-Verlet method. Although there are many alternative methods now available, the Störmer-Verlet method is rather straightforward to derive and has some attractive symmetry properties that lead to good long-time energy conservation. It is based on a Taylor expansion of the atomic positions, $\mathbf{r}(t)$, at times $t + \Delta t$ and $t - \Delta t$, *i.e.* extrapolating both forwards and backwards in time from current time, t , by some finite difference, Δt .

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \mathbf{v}(t)\Delta t + \frac{1}{2} \mathbf{a}(t)\Delta t^2 + O(\Delta t^3) \quad (1)$$

$$\mathbf{r}(t - \Delta t) = \mathbf{r}(t) - \mathbf{v}(t)\Delta t + \frac{1}{2} \mathbf{a}(t)\Delta t^2 - O(\Delta t^3) \quad (2)$$

where the first and second time derivatives of position $\mathbf{r}(t)$ with respect to time are written as velocity, $\mathbf{v}(t)$, and acceleration, $\mathbf{a}(t)$. Adding these two series (1) and (2) together, all the odd order terms in Δt cancel, and we are left with:

$$\mathbf{r}(t + \Delta t) + \mathbf{r}(t - \Delta t) = 2\mathbf{r}(t) + \mathbf{a}(t)\Delta t^2 + O(\Delta t^4) \quad (3)$$

Rearranging this expression for the new particle positions, we obtain:

$$\mathbf{r}(t + \Delta t) = 2\mathbf{r}(t) - \mathbf{r}(t - \Delta t) + \mathbf{a}(t)\Delta t^2 + O(\Delta t^4) \quad (4)$$

The accelerations at each time step, $\mathbf{a}(t)$, are obtained from Newton's second law using the known atomic masses and summing over all the forces given by the gradient of the potential energy, and so the new atomic positions at time $t + \Delta t$ can be obtained from the current positions at time, t , and previous time, $t - \Delta t$. Equation (4) has the features that it is accurate (locally) to terms of order Δt^4 (although the global accuracy is quadratic) and, more importantly, upon a change of variable from Δt to $-\Delta t$, the equation remains invariant. This time-reversal symmetry matches exactly the microscopic reversibility of the continuous-time particle dynamics, and helps to ensure that, even though numerical errors in the actual particle positions accumulate exponentially, the total energy of the system is still a conserved quantity. Numerical integrators which do not have this symmetry property, such as some predictor-corrector methods, may have problems with long-term energy conservation even though they have a higher numerical accuracy at each time step.¹¹ Note that the atomic velocities at each time step are not calculated explicitly, but can be obtained by averaging the positions between adjacent steps:

$$\mathbf{v}(t) = [\mathbf{r}(t + \Delta t) - \mathbf{r}(t - \Delta t)]/(2\Delta t) + O(\Delta t^2) \quad (5)$$

2.1.2 Extended ensembles for MD

Since the forces in Section 2.1.1 are derived from the gradient of a scalar potential, and Newton’s third law is strictly obeyed, the dynamics necessarily conserve the total energy of system, generating configurations in the microcanonical or NVE thermodynamic ensemble. However, most real systems of interest exist under conditions of constant temperature (canonical, or NVT ensemble) or constant temperature and pressure (isothermal-isobaric, or NpT ensemble). There are various methods for modifying MD to simulate systems at constant temperature and/or pressure, based on both deterministic and stochastic techniques or a hybrid of the two. The simplest but least accurate method of achieving constant temperature is to artificially scale the atomic velocities to drive the instantaneous system temperature, $T(t)$, related to mean kinetic energy of particles, towards some target equilibrium temperature, T_{eq} . In the Berendsen method,¹² this is achieved by setting the rate of change of instantaneous temperature equal to:

$$\frac{dT(t)}{dt} = \frac{1}{\tau_t} (T_{\text{eq}} - T(t)) \quad (6)$$

where τ_t is a parameter controlling the rate of energy flow between system and an external heat reservoir (often referred to as the thermostat ‘relaxation time’). As a rule of thumb, setting $\Delta t/\tau_t < 0.01$ usually results in a slow enough relaxation to produce a stable equilibrium temperature within a few tens of picoseconds. However, if τ_t is too short, then the temperature will fluctuate wildly, and if τ_t is too long, then simulation will take a long time to reach equilibrium.

Given that the instantaneous temperature can be estimated at each time step from the mean atomic kinetic energy:

$$T(t) \approx \frac{1}{3Nk_B} \sum_{i=1}^N m_i v_i^2 \quad (7)$$

then by substitution into equation (6), some simple algebra shows that if velocities are rescaled by a factor of λ , then the scale factor required at each step to obtain the desired target temperature is given by:

$$\lambda^2 = 1 + \frac{\Delta t}{\tau} \left(\frac{T_{\text{ext}}}{T(t)} - 1 \right) \quad (8)$$

Although fast and simple to implement, the Berendsen method suffers from the fact that fictitious forces are effectively applied to the atoms to change their velocities and, more seriously, that the fluctuations in kinetic energy of atoms are not correctly reproduced, especially for smaller systems. As a result, it is typically only used during equilibration period of MD simulation. A superior method for producing MD configurations which correctly sample the canonical ensemble was developed by Nosé,¹³ and later refined by Hoover,¹⁴ and is based on the concept of an extended Lagrangian. It is well-known that Newton’s equations can be reformulated variationally in terms of a Lagrangian, L :

$$L = \frac{1}{2} \sum_i^N m_i \dot{\mathbf{x}}_i^2 - V(\mathbf{x}) \quad (9)$$

from which follows Newton's second law, $\mathbf{F}_i = \dot{\mathbf{p}}_i$, by substitution of (9) into the Euler-Lagrange equation:

$$\frac{d}{dt} \frac{\partial L}{\partial \dot{\mathbf{x}}_i} - \frac{\partial L}{\partial \mathbf{x}_i} = 0 \quad (10)$$

The advantage of Lagrangian approach is that it uses generalized coordinates, which can include variables representing the external heat reservoir, and also be used to derive their corresponding equations of motion. For the Nosé-Hoover thermostat, the modified form of Newton's second law is given by:

$$\dot{\mathbf{p}}_i = \mathbf{F}_i - \zeta \mathbf{p}_i \quad (11)$$

where ζ is a frictional coefficient that evolves in time so as to minimise the difference between the instantaneous kinetic and equilibrium temperatures according to the expression:

$$\dot{\zeta} = \frac{1}{\tau_t} \{T(t)/T_{\text{eq}} - 1\} \quad (12)$$

where, similar to above, τ_t is the thermostat relaxation time.

The Nosé-Hoover thermostat correctly reproduces energy fluctuations of the system in the canonical ensemble but, similarly to the Berendsen method above, introduces fictitious forces (note the modified form of Newton's law) on atoms that can interfere with momentum transport on a local scale. Alternative thermostats based on pairwise thermalisation of energy, such as that used in dissipative particle dynamics (see section 3.1.1), can improve this situation, especially for small systems. Also, the Nosé-Hoover thermostat is prone to poor equilibration due to resonant energy transfer between system and reservoir, resulting in temperature oscillations that do not die away with time. In order to avoid this, it is common to couple many, sometimes thousands, of thermostats with different relaxation times into a so-called Nosé-Hoover chain, which has been shown to improve ergodicity of simulations of small, stiff systems which start far from equilibrium.¹⁵ The introduction of stochastic forces through coupling to a Langevin-type thermostat can also be effective in improving ergodicity, at the expense of a loss of continuity of the dynamics.

The Berendsen and Nosé-Hoover methods may also be extended to allow fluctuations in the cell volume for a periodic system, thereby allowing a constant pressure or stress to be maintained. The details of equations of motion will not be described here, but suffice it to say that great care must be taken to check whether isotropic or anisotropic variations in cell parameters are required to accurately simulate the system of interest. Completely isotropic cell box fluctuations are only suitable for elastically isotropic systems (*e.g.* liquids or glasses). However, in the case that fully anisotropic cell box fluctuations (Parinello-Rahman method¹⁶) are required, *e.g.* for low-symmetry crystals, the equations of motion for cell parameters can become very involved, with independent barostat relaxation times required for each direction in order to achieve a stable equilibrium pressure.

2.1.3 Enhanced sampling methods in MD

While the standard MD methods described in Sections 2.1.1 and 2.1.2 allow for simulations in a variety of useful thermodynamic ensembles, they are fundamentally limited by the time scale of the integration process, which is governed by the product of time step length (limited by relaxation of fastest motions in system) and total number of steps (limited by CPU power). In order to address this weakness, a number of sampling methods have been developed which focus on enhancing the probability of rare events – *i.e.* those that would occur with only vanishingly small probability over the time scale of a canonical MD simulation under normal conditions of temperature and pressure. A good example is the diffusion of atomic silver on a flat Ag(100) surface, where the real time between each hopping event is of order 10 μ s,¹⁷ which may take around one week of CPU time! The artificial acceleration of rare events allows the extension of atomistic modelling time scales up to the micro or even millisecond range, whereas the extension of length scales is better handled by coarse-grained methods described in Section 3.

Time acceleration in dynamical simulations can be achieved using some of the ideas of umbrella sampling (see section 2.2.3) in combination with either a bias potential or transition state theory (TST),¹⁸ which gives a simple Arrhenius form for the transition rate if the energy is assumed to vary harmonically near the minimum and barrier regions for all the degrees of freedom other than the reaction coordinate direction. For example, in the hyperdynamics scheme conceived by Voter,¹⁹ the original potential energy surface (PES), $V(x)$, is augmented by a bias potential, ΔV , which is zero at the dividing surface between the two energy minima, and acts to increase the frequency of barrier crossing by ‘filling in’ the areas of low energy. In the regions where the bias potential is non-zero, the effective simulation time passes more quickly by a factor of $\exp(\beta\Delta V)$. The ratio of accumulated hypertime to the standard MD clock time is known as the ‘boost’, and can be as large as 10^6 if an appropriate form of biasing potential is chosen. Unfortunately, it is not easy to find a general method of specifying the bias potential, and this area still is a topic of ongoing research.²⁰

A technique related to hyperdynamics is the metadynamics method of Laio and Parinello,^{21,22} in which a series of Gaussian functions are added to the PES in order to flatten it and force the atom to explore other regions of phase space. Metadynamics enables the rapid exploration of free energy surfaces in some chosen set of coordinates, but there is no direct connection to a timescale and so any dynamics is largely fictitious. Laio and Gervasio²² give the analogy of a walker trapped in an empty swimming pool at night who, from time to time, drops packets of sand on ground as they wander in the darkness. Given a sufficient supply of sand, they will eventually escape and, if they are able to remember where they dropped the sand, be able to reconstruct a negative image of the pool. The advantages of the metadynamics method over hyperdynamics are that it requires no *a priori* knowledge of the bias potential, and that the sum of Gaussians deposited up to a particular time provides an unbiased estimate of the free energy in the region explored during the simulation.²²

Two other accelerated dynamics methods also developed by Voter and co-workers that do not rely on biasing the PES are Parallel Replica Dynamics (PRD) and Temperature Accelerated Dynamics (TAD).^{23,24} In PRD, the canonical dynamics of a single system is replicated on a number (say, M) of processors running in parallel. Having allowed the replicas to become locally uncorrelated (*i.e.* randomised within the original basin of

attraction), the dynamics of all M systems are then monitored until, it is hoped, a transition occurs in a single one corresponding to a rare event. The simulation clock is then advanced by the elapsed time summed over all M replicas, and the replication process is continued from the replica which made the transition, allowing for a short period in which correlated dynamic events could occur. TAD, on the other hand, can be thought of as very similar to “on-the-fly” Kinetic Monte Carlo (KMC),²⁵ in which the barriers are constructed during the course of the simulation. It is based on the concept of raising the temperature of the system to enable rare events to occur more frequently, whilst at the same time preventing the system from evolving along diffusion pathways only accessible at high temperature. By confining the system to its local basin of attraction, the TAD method essentially tries to find all possible escape routes at high temperatures, and then selects the one with the shortest time to occur at low temperatures. Compared to other accelerated dynamics methods, TAD is the most approximate, relying heavily on the assumption of harmonic TST, whereas PRD is the most accurate.

2.2 Monte Carlo (MC)

MC methods derive their name from the association of statistical sampling methods with games of chance, such as those played in the famous casinos of Monte Carlo. Although the usage “Monte Carlo” was coined relatively recently (1949) and intimately associated with the use of computers, statistical sampling methods are in fact much older than this, as explained in Section 2.2.1. The goal of a typical MC simulation is to calculate the expectation value of some mechanical quantity Q (*e.g.* internal energy), which is defined by an average over all microstates of the system weighted with their Boltzmann probability:

$$\langle Q \rangle = \frac{1}{Z} \sum_i Q_i \exp(-\beta E_i) \quad (13)$$

where $\beta = 1/(k_B T)$, E_i is the energy of microstate i and $Z = \sum_i \exp(-\beta E_i)$ is the partition function.

Why not simply enumerate all the microstates and calculate the expectation value directly? Well, to borrow an illustration from Newman and Barkema,²⁶ a litre of gas at standard temperature and pressure contains of order 10^{22} molecules, each with a typical velocity of 100 m s^{-1} giving a de Broglie wavelength of around 10^{-10} m . Thus, the total number of microstates of this system is of order $(10^{27})^{10^{22}}$, ignoring the effects of indistinguishability, which is completely beyond enumeration even using a computer. However, if we choose only a small subset of M simulated states, selected according to a Boltzmann probability distribution, then the desired expectation value reduces to an arithmetic average over all sampled states, given by:

$$\langle Q \rangle = \frac{1}{M} \sum_{i=1}^M Q_i \quad (14)$$

We may ask if it is valid to average over such an infinitesimal portion of phase space. However, even real physical systems are sampling only a tiny fraction of their total number

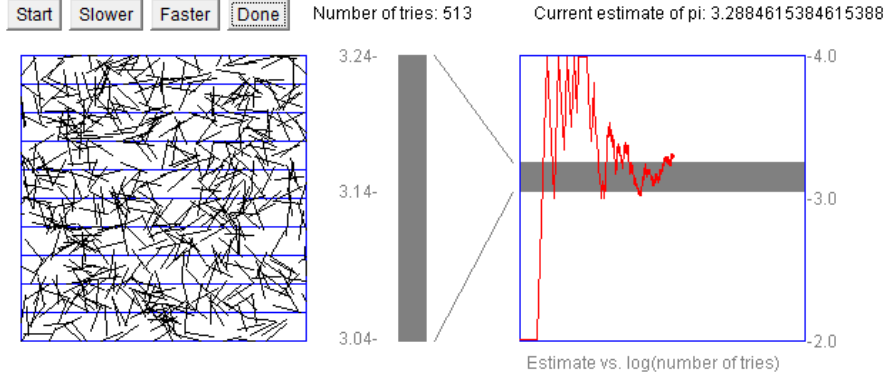


Figure 3. Screenshot of applet showing convergence of Buffon's needle experiment to estimate value of π after more than 500 throws of needle. <http://www.angelfire.com/wa/hurben/buff.html>

of microstates during the time we can make physical measurements on them. For example, in our aforementioned litre of gas, the molecules are undergoing collisions at a rate of roughly 10^9 collisions per second. This means that the system is changing microstates at a rate of 10^{31} per second, so it would require of order $10^{10^{23}}$ times the current predicted lifetime of the universe for it to move through every state! Therefore, it should not be surprising that we can perform reasonable calculations by considering a small, but representative, fraction of these states.

2.2.1 Statistical sampling methods

As mentioned above, the idea of using statistical sampling to estimate deterministic quantities predates considerably the existence of digital computers. Perhaps the most well-known example is the estimation of π by repeatedly dropping a needle onto a surface ruled with equally spaced lines. The experiment is named after Georges-Louis Leclerc, Comte de Buffon, who showed in 1777 that if a needle of length l is thrown at random onto lines of spacing d then the probability that the needle lands intersecting a line is $2l/(\pi d)$, provided that $d \geq l$. Laplace then pointed out in 1820 that if a needle is thrown down N times, and lands on a line M of those times, then an estimate for π is given by:

$$\lim_{N \rightarrow \infty} (2Nl/Md) \quad (15)$$

Figure 3 shows the behaviour of expression (15) to π for system with $l = d$, simulated by a JAVA applet, after more than 500 throws of the needle. As can be seen in the graph on right-hand side, the convergence is rather poor, with the standard error decreasing only as $1/\sqrt{N}$, and therefore this is not recommended as a method to calculate an accurate value of π . Nevertheless, it demonstrates the principle of Monte Carlo by which a deterministic quantity can be estimated from an average over states generated *via* a stochastic process, subject to certain acceptance rules.

2.2.2 Metropolis MC

In order to sample the canonical distribution, as required in Section 2.2, the Metropolis algorithm can be used to generate an appropriate ensemble of states *via* a Markov process.^{25,27} Instead of simply new accepting states on the basis of their absolute Boltzmann factor, the Metropolis algorithm is characterised by having an acceptance probability of unity if the new state has a lower energy than the initial state, which results in a much more efficient simulation in most cases. The algorithm can be summarised as follows:

1. Start with a system in (an arbitrarily chosen) state μ and evaluate the energy of current state, E_μ
2. Generate a new state ν by a small *ergodic* perturbation to state μ , and evaluate energy of new state, E_ν
3. If $E_\nu - E_\mu < 0$ then accept the new state. If $E_\nu - E_\mu > 0$ then accept the new state with probability $\exp[-\beta(E_\nu - E_\mu)]$
4. Return to step 2 and repeat until equilibrium is achieved (*i.e.* states appear with their correct Boltzmann probabilities at temperature T)

The Metropolis method can be illustrated by application to a simple model for ferromagnet, due to Ising,²⁸ which consists of a number of two-state (up/down) spins on a periodic lattice. Each spin can interact with its nearest neighbour, and also with an external magnetic field, according to the following Hamiltonian:

$$H = -\varepsilon \sum_{i,j} s_i s_j - B \sum_i s_i \quad (16)$$

where $s_i = \pm 1$ is the spin of state i , ε is the exchange energy and B is the strength of applied external field.

The lattice is initialized in a random ($T = \infty$) or completely ordered ($T = 0$) configuration, and then potential new states are generated by flipping single spin states (*i.e.* changing the sign of a particular s_i chosen at random). This guarantees ergodicity, as every microstate of the system is accessible, in principle, *via* this procedure. It is also very important, as in the case of molecular dynamics, to ensure microscopic reversibility at each step, in order to obtain the true canonical average quantities.

Since the 2D Ising model is exactly solvable by statistical mechanics techniques, it is possible to compare the predicted results of magnetization and heat capacity against theory.²⁶ It is found that Metropolis MC has problems with convergence around the Curie temperature, T_C , due to large fluctuations in the magnetization near the phase transition point. This can be improved by, for example, swapping clusters of spin states instead of single spin states.²⁵ Provided we generate these clusters probabilistically, the algorithm is still ergodic, and requires many fewer MC steps per lattice site for equilibration, also giving much better performance around the Curie point.

2.2.3 Enhanced sampling methods in MC

As seen above, there are many situations in which standard canonical MD or MC simulations are inadequate to calculate certain desired thermodynamic quantities. An obvious

example is entropies and free energies, which cannot be computed directly as averages over configurations or time as they are related to the volume of phase space accessible to the system. Moreover, the conventional methods sample only sparsely from unstable regions of configuration space, such as near to a transition point, giving rise to large statistical errors in the free energy differences calculated by comparing simulations of the two phases separately in thermal equilibrium. By introducing an additional weighting function to bias the Boltzmann distribution used in standard MC, the system can be guided to sample more frequently from the normally unstable regions, resulting in a more accurate estimate of the free energy.

The concept of “umbrella sampling” was originally developed by Torrie and Valleau²⁹ in order to calculate free energy differences in systems undergoing large changes in configuration, such as a first order phase transition (if the changes are not too large, more straightforward methods such as thermodynamic integration or Widom particle insertion can be used instead¹¹). However, the choice of weighting functions must be determined *ad hoc*, and the most efficient scheme is not always obvious. More recently, an adaptive form of umbrella sampling was developed by Wang and Landau,³⁰ which is related to the multi-canonical method of Berg and Neuhaus³¹ in which the histogram of sampled states is first flattened and then reweighted to enable the correct Boltzmann distribution to be deduced at any temperature within the sampled range. However, the objective of the Wang-Landau method is to determine the full density of states (DOS) by performing a random walk in configurational space with a probability proportional to the reciprocal of the density of states.

The thermodynamic reweighting method is illustrated in Figure 4 for an isolated lattice polymer chain of length 100 segments, where the $\log(\text{DOS})$ is shown as a function of the number of polymer-polymer nearest-neighbour contacts, and each contact contributes an energy, ϵ . The five dashed black lines show canonical histograms for the system at five different temperatures, and the dashed coloured line shows the total histogram taking into account the appropriately reweighted contributions from all temperatures. Once the DOS distribution is known to sufficient precision, all other thermodynamic quantities can easily be derived, and the inset to Figure 4 shows a plot of heat capacity versus reduced temperature showing transitions (*i.e.* peaks in the heat capacity) from an extended coil at high temperatures to a compact ‘crystal’ at low temperatures.³²

3 Coarse-Grained Modelling Methods

In order to move from atomistic to mesoscopic length scales, it is necessary to integrate out any redundant degrees of freedom, a process known as “coarse-graining”, which can be achieved either by forcing atoms onto a lattice or by grouping them into larger particles or ‘beads’ (often referred to as “mapping”). Conversely, the process of “reverse-mapping” refers to the restoration of full atomistic detail. These groups of atoms interact through some effective potential derived either through a systematic fitting procedure or by *ad hoc* parameterization of soft, repulsive potentials designed to reproduce the potential-of-mean-force between the centres-of-mass of the groups of atoms. In general, there is no systematic procedure for coarse-graining that is applicable across all classes of material, but many groups have developed methods for particular systems,^{33–36} and semi-automated

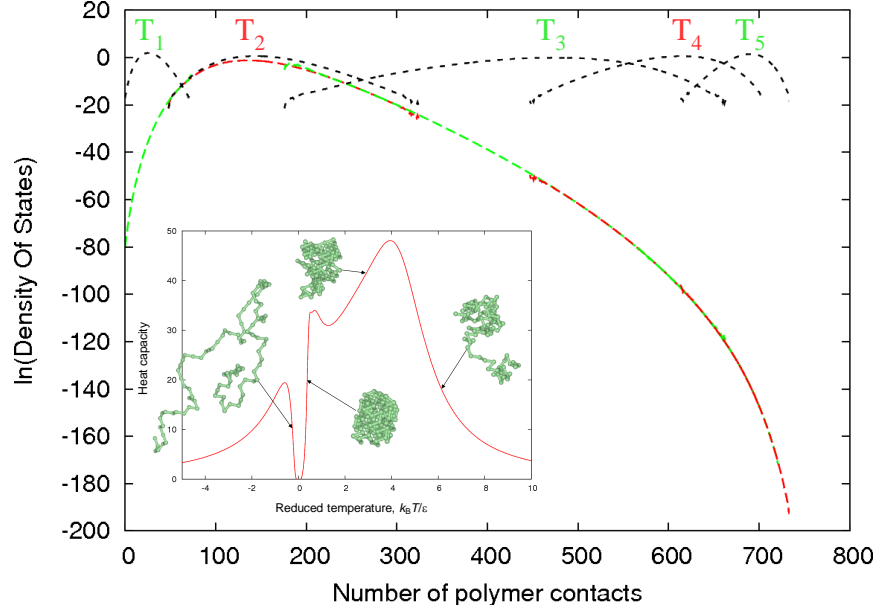


Figure 4. Reconstruction of density of states for a lattice chain polymer by histogram reweighting, with heat capacity versus reduced temperature shown as inset.

tools, such as the VOTCA (Versatile Object-oriented Toolkit for Coarse-graining Applications³⁷) package are available.

An example of a polymer chain coarse-grained onto a face-centred cubic lattice was shown in Figure 4. The processes of mapping and reverse-mapping enables the calibration of a generic lattice chain to a particular system, with specific molecular chemistry, by matching of the densities, end-to-end distance and radial distribution functions.³⁸ Furthermore, due to the greatly reduced number of configurations, the lattice chain provides a much more computationally convenient framework for simulating larger systems or bulk phase behaviour.³⁹ The length scale of mesoscopic simulations can now even be extended to model colloids or powders through the use of dissipative particle dynamics (DPD) and discrete (or distinct) element modelling (DEM)^{40,41} in which each particle is considered to be either a whole or part constituent of a granular medium, interacting *via* elastic and dissipative force contact laws. This brings us finally almost to the top of the hierarchy of modelling techniques shown in Figure 2, making contact with the Finite Element Method (FEM),⁴² commonly used for stress analysis and heat or mass transfer problems in engineering.

3.1 Coarse-grained particle methods

3.1.1 Dissipative particle dynamics

A popular mesoscale method for simulating soft materials, such as polymer and liquids, is Dissipative Particle Dynamics (DPD), first developed by Hoogerbrugge and Koelman^{43,44} for modelling the flow of hard spheres in suspension, and reformulated on a rigorous thermodynamic basis by Groot and Warren⁴⁵ and Español and Warren.⁴⁶ It is closely related to the Brownian Dynamics (BD) method,⁴⁷ in which standard canonical MD is augmented by dissipative and random forces between particles, representing the integrated effects of a coarse-grained fluid medium, in addition to a soft repulsive force which can be deduced from experiments or molecular simulations *via* Flory-Huggins theory. However, unlike in BD, the forces in DPD are always pairwise acting, which guarantees the emergence of true hydrodynamic behaviour in the limit of large system size.⁴⁸ In this extended sense, DPD can be thought of simply as local, hydrodynamics-conserving Langevin-type thermostat for MD (compare with the Berendsen and Nosé-Hoover methods in Section 2.1.2).

The force interactions in DPD can be summarised by the following expressions:

$$\mathbf{F}_{ij}^C = (n\varepsilon_{ij}/\sigma_{ij})\omega^C(\sigma_{ij}/r_{ij})^{n+1}\hat{\mathbf{r}}_{ij} \quad (17)$$

$$\mathbf{F}_{ij}^D = -\kappa\omega^D(\hat{\mathbf{r}}_{ij} \cdot \mathbf{v}_{ij})\hat{\mathbf{r}}_{ij} \quad (18)$$

$$\mathbf{F}_{ij}^R = \lambda\omega^R\theta_{ij}\Delta t^{-1/2}\hat{\mathbf{r}}_{ij} \quad (19)$$

where equation (17) is the conservative soft power law repulsive force, equation (18) is the dissipative force and equation (19) is the random ‘‘Brownian’’ force which act between two spheres i and j whose centres are connected by a vector \mathbf{r}_{ij} and travel with relative velocity \mathbf{v}_{ij} . The parameters ε_{ij} and σ_{ij} are the soft repulsive coefficients of spheres i and j , and θ_{ij} are a set of Gaussian random numbers with zero mean and unit variance. The exponent, n , controls the ‘hardness’ of repulsive interaction.

Español and Warren⁴⁶ showed that if the weight functions ω^D and ω^R are chosen to satisfy a fluctuation-dissipation theorem, then an equilibrium temperature is established in the simulation. The functions include an explicit cut-off distance, r_c , which is typically set to twice the size of the largest particle in the simulation.

$$\omega^C = \begin{cases} 1 & ; r < r_c \\ 0 & ; r \geq r_c \end{cases} \quad (20)$$

$$\omega^D = \begin{cases} (1 - r/r_c)^2 & ; r < r_c \\ 0 & ; r \geq r_c \end{cases} \quad (21)$$

$$\omega^R = \begin{cases} (1 - r/r_c) & ; r < r_c \\ 0 & ; r \geq r_c \end{cases} \quad (22)$$

The friction coefficient κ and the noise amplitude λ are connected by equation (23), where T_{eq} is the desired equilibrium temperature of the simulation. The appearance of the timestep Δt in the expression for the random force, equation (19), is due to the effect of time discretization in the integration algorithm, and the origins of this have been discussed in detail by Groot and Warren.⁴⁵

$$\lambda^2 = 2\kappa k_B T_{\text{eq}} \quad (23)$$

The equations of motion for DPD system are given by expressions (24)–(27). These are based on the combined thermostat and barostat proposed by Hoover¹⁴ and subsequently reformulated by Melchionna,⁴⁹ which is commonly used in standard NpT MD simulations. However, in this system the thermostat has been completely removed, and the barostat has been decoupled from the translational degrees of freedom of the particles. This is necessary in order to ensure that collisions between particles conserve momentum. An equilibrium temperature is then established by the DPD force interactions, given by equations (17)–(19), which act pairwise between each component sphere so that Newton’s third law is obeyed.

$$\dot{\mathbf{r}}_i = \mathbf{v}_i + \eta \mathbf{r}_i \quad (24)$$

$$\dot{\mathbf{p}}_i = \mathbf{F}_i \quad (25)$$

$$\dot{\eta} = \frac{1}{\tau_p^2} \frac{1}{N k_B T_{\text{eq}}} (p(t) - p_{\text{eq}}) \quad (26)$$

$$\dot{V} = 3V\eta \quad (27)$$

The time evolution of the particle coordinates, \mathbf{r}_i , is calculated from equation (24), where η is a fictitious dynamical variable that compensates for any difference between the instantaneous pressure and the desired equilibrium pressure, p_{eq} . The simulation box volume, V , relaxes to its equilibrium value with a characteristic time, τ_p .

It can be shown⁵⁰ that these modified equations of motion sample from a pseudo-Boltzmann constant pressure ensemble which involves the instantaneous temperature of the simulation. Thus, when used in combination with the DPD force interactions, the simulation as a whole relaxes to an equilibrium temperature T_{eq} and equilibrium pressure p_{eq} . By setting the relative levels of T_{eq} and p_{eq} , the assemblies of particles can be mixed or packed densely together as desired.

3.1.2 Discrete element method

The discrete (or distinct) element method (DEM), also known as granular dynamics (GD), is a numerical technique for simulating the dynamics of semi-rigid macroscopic frictional particles with sizes ranging from tens of metres to micrometres, such as pharmaceutical powders, talcs, cement, sand and rocks. In the absence of an interstitial medium, these particles interact with one another *via* short-range contact mechanical forces,^{51,52} which include both elastic and viscoelastic components, along with macroscopic surface friction. The nature of the interactions coupled with the size of the particles (relative to atomic and molecular systems) are such that the dynamics of these systems is rapidly quenched due to the dissipative interactions, unless there is energy input in the form of mechanical excitation. Cundall and Strack⁵³ developed GD, or DEM, to study geophysical systems using simple linear damped spring force models to represent the interactions. Since then the numerical representation of the pairwise particle interactions has evolved to incorporate force-displacement (linear spring, Hertzian) and force-displacement-velocity (spring-dashpot, Hertz-Mindlin, Hertz-Kuwabara-Kono (HKK)) models which account for experimentally measured material properties such as the elastic moduli, Poisson’s ratio and surface friction.⁵⁴ Macroscopic surface friction between particle surfaces is accounted for *via* the use of Coulomb’s yield criteria.⁵⁵

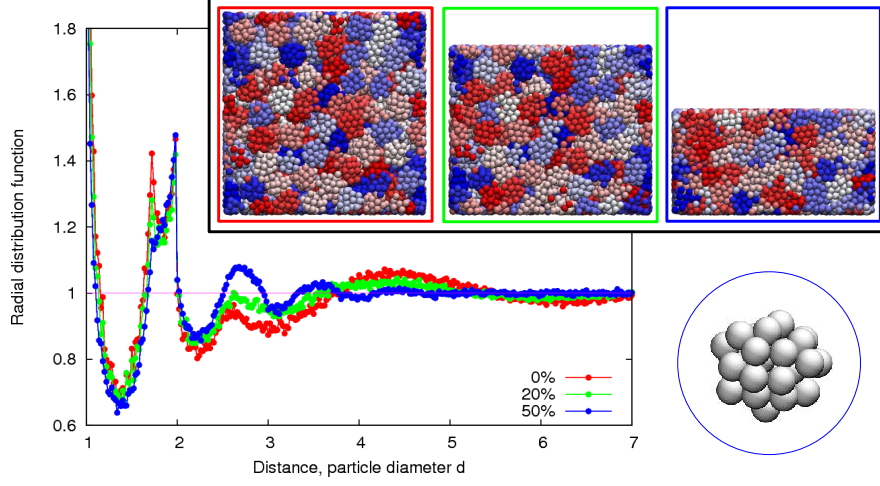


Figure 5. DEM simulation of the uniaxial compaction of 400 cohesive particle agglomerates, each comprised of 33 glued spheres (inset), as function of applied axial strain (0%, 20% and 50%, respectively). The system is three-dimensional and periodic in the plane perpendicular to the smooth, impenetrable compacting surfaces.

The nature of the contact forces, which act parallel and perpendicular to the vector connecting the centres of mass of the interacting particles, require a coupling between translational and rotational degrees of freedom, as the effective tangential force applies a torque on the particle. Figure 5 shows a model for an assembly of cohesive particles, where each agglomerate (identified by its individual colour) is made up of ‘glued’ spheres, fused by normal and tangential bonded potentials with a maximum threshold fracture strength, in order to simulate large scale plastic deformation and fracture. Such models are now providing insights into the constitutive behaviour of cohesive and crushable particle assemblies at the macroscopic scale.

3.2 Mesh-based finite element methods

In the limit of large numbers of particles, mesoscopic simulations should ideally yield the same solutions as the corresponding continuum constitutive equations. This is provably the case for DPD and the Navier-Stokes equations for simple Newtonian liquids, and widely believed to be true also for more complicated assemblies of elasto-rigid or elasto-plastic particles (although the precise constitutive laws are often semi-empirical in nature). The behaviour of large systems acting under such constitutive laws can be solved more efficiently by using standard finite element models (FEM) on a discretized mesh connecting nodal elements. The system of equations can be represented in matrix form by:

$$[K] \{u\} = \{F\} \quad (28)$$

where $[K]$ is the global stiffness matrix (assembled from the individual elements), $\{u\}$ are the nodal displacements, and $\{F\}$ are the nodal forces. By enforcing the appropriate dis-

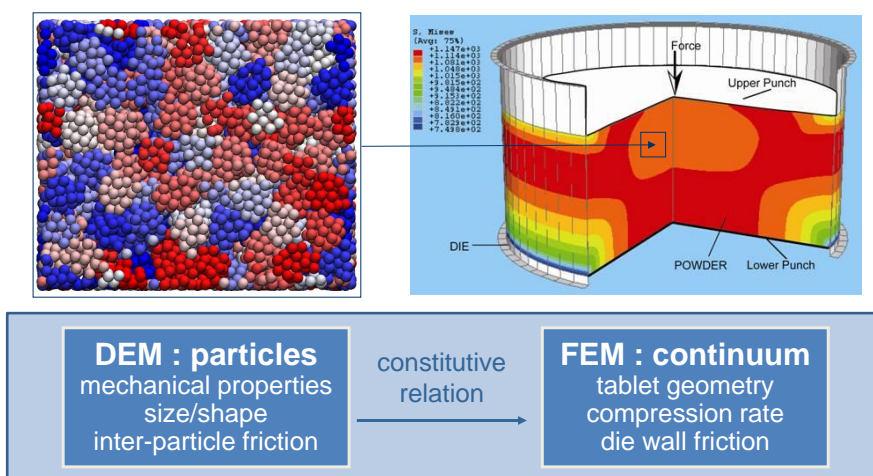


Figure 6. Coupled methods for powder compaction, in which DEM simulation is used to derive material parameters (constitutive model) for larger-scale FEM study of pharmaceutical tablet compaction.

placement and/or stress conditions at the boundaries, the system of equations can be solved to yield the unknown nodal displacements, provided that the stress-strain behaviour of each element is known. For a complex material, such as an organic powder, this usually must be determined experimentally from the properties of bulk material. However, increasingly, multiscale computational simulations are able to yield useful results from first principles.

Figure 6 illustrates the relationship between a coupled DEM simulation of powder compaction and a corresponding FEM simulation of an axisymmetric pharmaceutical tablet. Each element in the FEM simulation effectively contains many thousands of granules, corresponding to many millions of mesoscopic particles. The properties of powder material, such as elastic moduli, Poisson's ratio, particle size distribution, etc. are taken into account by the DEM simulation, whereas the FEM simulation considers the process parameters such as tablet geometry, compaction rate and friction between material and die wall or punch. Such multiscale coupled simulations are now helping to accelerate the formulation of powder blends for compaction at macroscale based on knowledge of the molecular structure and interactions of their constituents.

4 Conclusions and Outlook

In conclusion, we may speculate tentatively on what might be the most fertile areas for the development of multiscale over the next five years. In the author's opinion, the development of a general framework for transforming seamlessly from particle-based to continuum-based representations of materials will enable large-scale simulations of failure in granular and monolithic systems to become almost routine. On-the-fly coarse-graining will permit use of explicit solvent in the vicinity of solute molecules or particles, whilst still allowing the interaction of many thousands or even millions of them, facilitating the study of self-assembly in nanocomposite or biomimetic systems. Furthermore, the inclusion of

unexpected rare events into dynamical simulations will yield new insights into atomistic failure mechanisms in nanocrystalline materials, and open up the possibility of studying very long time scale relaxations in systems containing full molecular detail. However, these developments must be tempered by the need to advance hand-in-hand with theoretical and experimental science, and thus the most likely scenario for the development of multiscale modelling in the near future is its continued incremental growth.

Acknowledgments

The author gratefully acknowledges funding support from UK Engineering and Physics Sciences Research Council (EPSRC) and Pfizer Inc., which supported parts of the work discussed in this article.

References

1. G.E. Moore, *Cramming more components onto integrated circuits*, Electronics, **38**, 114–117, 1965.
2. J. A. Elliott, *Novel approaches to multiscale modelling in materials science*, Int. Mater. Rev., **56**, no. 4, 207–225, 2011.
3. J.A. Elliott and B.C. Hancock, *Pharmaceutical materials science: An active new frontier in materials research*, MRS Bull., **31**, no. 11, 869–873, 2006.
4. J. H. Harding, D. M. Duffy, M. L. Sushko, P. M. Rodger, D. Quigley, and J. A. Elliott, *Computational Techniques at the Organic-Inorganic Interface in Biomineralization*, Chem. Rev., **108**, no. 11, 4823–4854, 2008.
5. J. J. Vilatela, J. A. Elliott, and A. H. Windle, *A Model for the Strength of Yarn-like Carbon Nanotube Fibers*, ACS Nano, **5**, no. 3, 1921–1927, 2011.
6. J. A. Elliott, J. K. W. Sandler, A. H. Windle, R. J. Young, and M. S. P. Shaffer, *Collapse of single-wall carbon nanotubes is diameter dependent*, Phys. Rev. Lett., **92**, no. 9, 095501, 2004.
7. A. R. Leach, *Molecular modelling: principles and applications*, Prentice-Hall, Harlow, second edition, 2001.
8. J. D. Bernal, *The 1962 Bakerian Lecture*, Proc. Roy. Soc., **280**, 299–322, 1964.
9. B. J. Alder and T. E. Wainwright, *Phase transitions for a hard sphere system*, J. Chem. Phys., **27**, 1208–1209, 1957.
10. L. Verlet, *Computer 'experiments' on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules*, Phys. Rev., **159**, 98–103, 1967.
11. D. Frenkel and B. Smit, *Understanding molecular simulation*, Academic Press, San Diego, 2nd edition, 2002.
12. H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. Dinola, and J. R. Haak, *Molecular dynamics with coupling to an external bath*, J. Chem. Phys., **81**, 3684–3690, 1984.
13. S. Nosé, *A Molecular-Dynamics Method for Simulations in the Canonical Ensemble*, Mol. Phys., **52**, no. 2, 255–268, 1984.
14. W. G. Hoover, *Canonical dynamics: Equilibrium phase-space distributions*, Phys. Rev. A, **31**, 1695–1697, 1985.

15. G. J. Martyna, M. L. Klein, and M. Tuckerman, *Nosé-Hoover chains: the canonical ensemble via continuous dynamics*, J. Chem. Phys., **97**, no. 4, 2635–2643, 1992.
16. M. Parinello and A. Rahman, *Crystal structure and pair potentials: a molecular dynamics study*, Phys. Rev. Lett., **45**, 1196–1199, 1980.
17. B. P. Uberuaga and A. E. Voter, *Determining reaction mechanisms*, Springer, Dordrecht, 2005.
18. D. Chandler, *Introduction to Modern Statistical Mechanics*, Oxford University Press, New York, 1987.
19. A. F. Voter, *A method for accelerating the molecular dynamics simulation of infrequent events*, J. Chem. Phys., **106**, no. 11, 4665–4677, 1997.
20. A. F. Voter, F. Montalenti, and T. C. Germann, *Extending the time scale in atomistic simulation of materials*, Ann. Rev. Mater. Res., **32**, 321–346, 2002.
21. A. Laio and M. Parrinello, *Escaping free-energy minima*, P. Natl. Acad. Sci. USA, **99**, no. 20, 12562–12566, 2002.
22. A. Laio and F. L. Gervasio, *Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science*, Rep. Prog. Phys., **71**, no. 12, 126601, 2008.
23. A. F. Voter and M. R. Sorensen, *Accelerating atomistic simulations of defect dynamics: Hyperdynamics, parallel replica dynamics, and temperature-accelerated dynamics*, Mat. Res. Soc. Symp. Proc., **538**, 427–439, 1999.
24. M. R. Sorensen and A. F. Voter, *Temperature-accelerated dynamics for simulation of infrequent events*, J. Chem. Phys., **112**, no. 21, 9599–9606, 2000.
25. D.P. Landau and K. Binder, *A Guide to Monte Carlo Simulations in Statistical Physics*, Cambridge University Press, Cambridge, 3rd edition, 2009.
26. M. E. J. Newman and G. T. Barkema, *Monte Carlo methods in statistical physics*, Clarendon Press, Oxford, 1999.
27. N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, *Equations of state calculations by fast computing machines*, J. Chem. Phys., **21**, no. 1087-1091, 1953.
28. E. Ising, *Beitrag zur Theorie des Ferromagnetismus*, Z. Phys., **31**, 253–258, 1925.
29. G. M. Torrie and J. P. Valleau, *Monte-Carlo Study of a Phase-Separating Liquid-Mixture by Umbrella Sampling*, J. Chem. Phys., **66**, no. 4, 1402–1408, 1977.
30. Fugao 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.
31. B. A. Berg and T. Neuhaus, *Multicanonical Ensemble - A New Approach To Simulate 1st-Order Phase-Transitions*, Phys. Rev. Lett., **68**, no. 1, 9–12, 1992.
32. D. Antypov and J. A. Elliott, *Computer simulation study of a single polymer chain in an attractive solvent*, J. Chem. Phys., **129**, no. 17, 174901, 2008.
33. W. Tschop, 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.
34. W. Tschop, K. Kremer, O. Hahn, J. Batoulis, and T. Burger, *Simulation of polymer melts. II. From coarse-grained models back to atomistic description*, Acta Polym., **49**, no. 2-3, 75–79, 1998.
35. F. Müller-Plathe, *Coarse-graining in polymer simulation: From the atomistic to the mesoscopic scale and back*, ChemPhysChem, **3**, no. 9, 754–769, 2002, Times Cited: 155.

36. G. A. Voth, *Coarse-Graining of Condensed Phase and Biomolecular Systems*, CRC Press, Boca Raton, Florida, 2009.
37. V. Rühle, C. Junghans, A. Lukyanov, K. Kremer, and D. Andrienko, *Versatile Object-Oriented Toolkit for Coarse-Graining Applications*, J. Chem. Theory Comput., **5**, no. 12, 3211–3223, 2009.
38. K. R. Haire, T. J. Carver, and A. H. Windle, *A Monte Carlo lattice model for chain diffusion in dense polymer systems and its interlocking with molecular dynamics simulation*, Comput. Theor. Polym. S., **11**, 17–28, 2001.
39. D. Antypov and J. A. Elliott, *Wang-Landau simulation of polymer-nanoparticle mixtures*, Macromolecules, **41**, no. 19, 7243–7250, 2008.
40. H. P. Zhu, Z. Y. Zhou, R. Y. Yang, and A. B. Yu, *Discrete particle simulation of particulate systems: Theoretical developments*, Chem. Eng. Sci., **62**, no. 13, 3378–3396, 2007.
41. H. P. Zhu, Z. Y. Zhou, R. Y. Yang, and A. B. Yu, *Discrete particle simulation of particulate systems: A review of major applications and findings*, Chem. Eng. Sci., **63**, no. 23, 5728–5770, 2008.
42. N. Ramakrishnan and V. S. Arunachalam, *Finite element methods for materials modelling*, Prog. Mater. Sci., **42**, no. 1-4, 253–261, 1997.
43. P. J. Hoogerbrugge and J. M. V. A. Koelman, *Simulating microscopic hydrodynamic phenomena with dissipative particle dynamics*, Europhys. Lett., **19**, 155–160, 1992.
44. J. M. V. A. Koelman and P. J. Hoogerbrugge, *Dynamic simulations of hard-sphere suspensions under steady shear*, Europhys. Lett., **21**, 363–368, 1993.
45. R. D. Groot and P. B. Warren, *Dissipative particle dynamics: bridging the gap between atomistic and mesoscopic simulation*, J. Chem. Phys., **107**, 4423–4435, 1997.
46. P. Español and P. B. Warren, *Statistical mechanics of dissipative particle dynamics*, Europhys. Lett., **30**, 191–196, 1995.
47. H. J. C. Berendsen, *Simulating the Physical World: Hierarchical Modeling from Quantum Mechanics to Fluid Dynamics*, Cambridge University Press, Cambridge, 2005.
48. P. Español, *Hydrodynamics from Dissipative Particle Dynamics*, Phys. Rev. E, **52**, no. 2, 1734–1742, 1995.
49. S. Melchionna, G. Ciccotti, and B. L. Holian, *Hoover NPT dynamics for systems varying in shape and size*, Mol. Phys., **78**, 533–544, 1993.
50. J. A. Elliott and A. H. Windle, *A dissipative particle dynamics method for modeling the geometrical packing of filler particles in polymer composites*, J. Chem. Phys., **113**, no. 22, 10367–10376, 2000.
51. H.M. Jaeger, Nagel S. R., and R. P. Behringer, *Granular solids, liquids, and gases*, Rev. Mod. Phys., **68**, 1259–1273, 1996.
52. K. L. Johnson, *Contact Mechanics*, Cambridge University Press, United Kingdom, 1st edition, 1985.
53. P. A. Cundall and O. D. L. Strack, *Discrete Numerical-Model for Granular Assemblies*, Geotechnique, **29**, no. 1, 47–65, 1979.
54. J. Schäfer, S. Dippel, and D. E. Wolf, *Force schemes in simulations of granular materials*, J. Phys. I France, **6**, 5–20, 1996.
55. J. A. Åström, H. J. Herrmann, and J. Timonen, *Granular packings and fault zones*, Phys. Rev. Lett., **84**, 638–641, 2000.

Introduction to Modelling, Scalability and Workflows with DL_POLY

**Ilian T. Todorov, Laurence J. Ellison,
Michael A. Seaton, and William Smith**

Scientific Computing Department
STFC Daresbury Laboratory
Warrington WA4 4AD, UK
E-mail: ilian.todorov@stfc.ac.uk

The DL_POLY project is a library of classical molecular dynamics programs that have applications over a wide range of atomic and molecular systems. DL_POLY_4, the CCP5 flagship version of this project, is specifically designed to address very large simulations on massively parallel computers in a scalable manner by stretching its parallel performance from small systems consisting of a few hundred atoms on a few compute cores, up to systems of hundreds of millions of atoms on tens of thousands of compute cores. In this article we briefly describe the structure of the programs, its scalability and possible workflows via UNICORE.

1 Introduction

The DL_POLY initiative was conceived by W. Smith in the early 1990s. Its prime purpose was to provide the UK CCP5¹ community with a classical molecular dynamics (MD) simulation package that was capable of exploiting emergent parallel computers. The project has been under continual development at STFC Daresbury Laboratory with funding streams from EPSRC,² CCP5 and NERC.³ Since its first release to the wider academic community, in 1996, over 12,000 licences have been taken with current uptake per annum of $\sim 2,000$ (as of 2012/2013). The original program, DL_POLY_2,^{4,5} has evolved into two currently available versions, namely DL_POLY_Classic⁸ and DL_POLY_4,^{6,7,9} with still increasing popularity amongst the modelling community world-wide. The main difference between the two versions is their underlying parallelisation strategy – replicated data (RD) for the former and Domain Decomposition (DD) for the latter. Both programs are available as free-of-charge source code to academic researchers world-wide. DL_POLY_4 is under active development and, due to its excellent scalability, is installed on many HPC facilities across the world. In this paper we will concentrate on DL_POLY_4.

2 Software

The DL_POLY_4 program design is based on the principles of portability, maintenance, transparency and user verification. The code architecture adopts Fortran90 modularisation in a C/C++ header style manner, where concepts and functionality are separated in a functional way by modules. The code routines relate to features/actions by their file names, which often relate to module names.

DL_POLY_4 is provided as fully self-contained (with no dependencies), free-formatted Fortran90 source. Additionally, the code also relies upon MPI2 (specifically Fortran90 +

TR15581 + MPI1 + MPI-I/O only) to implement its parallelisation strategies. The source complies strictly with the NAGWare¹⁰ and FORCHECK¹¹ Fortran90 standards with the only exception being the Fortran2003 feature known as TR15581, which is very rarely unavailable in current Fortran95 compilers.

A DL_POLY.4 CUDA port is also available to harness the power offered by NVIDIA®¹² GPUs. However, it includes dependencies on NVIDIA's CUDA libraries and OpenMP.

3 Molecular Structures

The simplest entities recognised by DL_POLY are atoms, which are regarded as point particles interacting with neighbouring particles via a centro-symmetric potential function. Simple atomic ions are also represented in this way. Their dynamics are described by translational motion as in a classical Newtonian treatment. Also possible are rigid molecules, which are point atoms maintained in a fixed geometry. These entities possess both translational (Newtonian) motion and rotational (Eulerian) motion and are useful for describing small molecules such as water. For larger and more flexible structures, such as polymers, point atoms may be connected by rigid bonds allied with some *intra*-molecular interactions, such as bond angle and dihedral angle potentials, which maintain the basic molecular geometry but permit *intra*-molecular conformational changes, which are an essential feature of the dynamics (and chemistry) of chains. Sometimes, completely flexible molecules are required, in which case the rigid bonds are replaced by extensible bond potentials. All of these molecular entities are permitted in any combination by DL_POLY, so a rigid body solvent and a flexible chain polymer may be simulated together, for example.

4 Force Field

The DL_POLY package does not provide any particular set of force field (FF) parameters to describe the interatomic interactions as other packages do such as AMBER,^{14,15} GRO-MACS,^{16,17} NAMD^{18,19} and CHARMM.^{20,21} DL_POLY is designed to cater for molecular systems of any complexity and thus it is impractical to bind the design to a set FF. In order to handle all possible FFs DL_POLY implements an enormous selection of functional forms, both analytic and tabulated, for the interaction potentials arising in many of the FFs commonly used in molecular simulations. It is also easy, due to the structure of the software, for the user to extend the FF potentials set to their liking as well as use as many different kinds of potentials simultaneously. Despite this freedom of unconstrained flexibility in mixing FFs to any complexity the user may wish, this design feature, until recently, was found to be a quite a barrier for many modellers, especially from the biochemical community. However, this has now been addressed in a very elegant way by the satellite program DL_FIELD,¹³ a FF generator for DL_POLY. DL_FIELD facilitates the conversion of a protonated PDB (also some simple ionic solids) input into DL_POLY input with minimal user intervention, using a small set of user specified options for matching the input to a number of FF sets such as AMBER, CHARMM, AMBERs Glycam, OPLS-AA, Dreiding²³ and PCFF.²⁴ The user is also given flexibility to design their own FF by extending and/or overriding the default FF sets.

The total potential energy for DL_POLY can be expressed by following formula:

$$\begin{aligned}
U(\underline{r}_1, \underline{r}_2, \dots, \underline{r}_N) = & \sum_{i_{shell}=1}^{N_{shell}} U_{shell}(i_{shell}, \underline{r}_{core}, \underline{r}_{shell}) \\
& + \sum_{i_{tether}=1}^{N_{tether}} U_{tether}(i_{tether}, \underline{r}_i^{\mathbf{t}=t}, \underline{r}_i^{\mathbf{t}=0}) \\
& + \sum_{i_{bond}=1}^{N_{bond}} U_{bond}(i_{bond}, \underline{r}_a, \underline{r}_b) \\
& + \sum_{i_{angle}=1}^{N_{angle}} U_{angle}(i_{angle}, \underline{r}_a, \underline{r}_b, \underline{r}_c) \\
& + \sum_{i_{dihed}=1}^{N_{dihed}} U_{dihed}(i_{dihed}, \underline{r}_a, \underline{r}_b, \underline{r}_c, \underline{r}_d) \\
& + \sum_{i_{invers}=1}^{N_{invers}} U_{invers}(i_{invers}, \underline{r}_a, \underline{r}_b, \underline{r}_c, \underline{r}_d) \\
& + \sum_{i=1}^{N-1} \sum_{j>i}^N U_{2-body}^{(metal, vdw, electrostatic)}(i, j, |\underline{r}_i - \underline{r}_j|) \quad (1) \\
& + \sum_{i=1}^N \sum_{j \neq i}^N \sum_{k \neq j}^N U_{tersoff}(i, j, k, \underline{r}_i, \underline{r}_j, \underline{r}_k) \\
& + \sum_{i=1}^{N-2} \sum_{j>i}^{N-1} \sum_{k>j}^N U_{3-body}(i, j, k, \underline{r}_i, \underline{r}_j, \underline{r}_k) \\
& + \sum_{i=1}^{N-3} \sum_{j>i}^{N-2} \sum_{k>j}^{N-1} \sum_{n>k}^N U_{4-body}(i, j, k, n, \underline{r}_i, \underline{r}_j, \underline{r}_k, \underline{r}_n) \\
& + \sum_{i=1}^N U_{external}(i, \underline{r}_i, \underline{v}_i) \quad ,
\end{aligned}$$

where U_{shell} acknowledges ion polarisation contributions coming from the extension of the point charge ion model via the shell model of Dick and Overhauser,²⁵ the adiabatic method of Fincham²⁶ or the relaxation model of Lindan.²⁷ The tether potential (U_{tether}) is a simple spring potential intended to keep a particle in the vicinity of its starting position. The rest of the *intra*-molecular interactions in DL_POLY have a wide selection of bond potentials (U_{bond}), angle potentials (U_{angle}), dihedral angle potentials (U_{dihed}) and inversion angle potentials (U_{invers}) which fully covers and exceeds the variety of those available in the custom FFs mentioned above.

The three-body (U_{3-body}) and four-body (U_{4-body}) interactions are non-specific angular potentials (suitable for glasses). Many-body interactions, an increasingly common requirement for modelling complex systems, are available in Tersoff forms ($U_{tersoff}$)^{28,29} for

covalent systems and via two-body decomposed ($U_{2-body}^{(metal)}$) metal potentials for metals and metal alloy systems. The latter include a wide variety of Finnis-Sinclair (FS) forms^{30–34} as well as Embedded Atom Method (EAM) forms.^{35–40} The two-body term also includes (i) all commonly used pair potentials ($U_{2-body}^{(vdw)}$) including Lennard-Jones, Buckingham, 12-6, N-M, Morse, etc. as well as allowing for input in a tabulated form, and (ii) electrostatic interactions ($U_{2-body}^{(electrostatic)}$), available as point charge and polarisable shell models, for which a variety of summation techniques may be selected (see Section 7).

Lastly, DL_POLY permits the user to apply external force fields. This capability is useful for modelling transport (e.g. conduction), or containment (e.g. pores) or mechanical intervention (e.g. shearing).

5 Integration Algorithms

The integration algorithms in DL_POLY handle the dynamics of the system being simulated. From the current positions of the atoms, the forces may be calculated from the first derivatives of the potential functions outlined above and then used to update the atomic velocities and positions. The integration progresses in a sequence of finite steps in time, each time step being of the order 0.001 up to 10 fs depending on the model system potentials and initial conditions of the dynamics simulation. DL_POLY_4 includes an option for self-adjustable timestepping (variable timestep) should the user need to allow for the feature. Although the feature is not needed for systems in equilibrium, it may be useful to determine the most advantageous timestep size for the particular model system.

The integration algorithms in DL_POLY_4 are based on the leapfrog Verlet (LFV)^{41,43} and velocity Verlet (VV)^{42,43} schemes. In addition to providing a numerical solution to the equations of motion, the integration algorithm also defines the thermodynamic ensemble. Although not all integrator implementations define ensembles, notably those of Andersen and Berendsen, we will refer to them as such in the text below! DL_POLY_4 provides access to a variety of ensembles: NVE (constant energy ensemble), NVT (canonical) ensembles of Evans⁴⁴ (a.k.a iso-kinetic, gaussian-constraint kinetic energy), Langevin^{45,46} (a.k.a stochastic dynamics), Andersen,⁴⁷ Berendsen,⁴⁸ Nosé-Hoover⁴⁹ and a gentle-stochastic thermostat.^{50,51} For constant pressure work there are the isothermal-isobaric (NPT) ensembles of Langevin,⁵² Berendsen,⁴⁸ Nosé-Hoover^{53,54} and Martyna-Tuckerman-Klein.⁵⁵ These are complemented by the anisotropic forms (NsT) for simulation of phase transitions in solids. The latter provide for further extensions⁵⁶ to constant normal pressure and surface (NP_nAT) and constant normal pressure and surface tension ($NP_n\gamma T$), which are useful for modelling interfaces.

DL_POLY_4 also accepts molecular structures defined by constraint bonds (CB) and rigid bodies (RB). The types of molecular structures that may be accommodated in a DL_POLY_4 simulation are shown in Figure 1. It is important to note that all such structures may be present in one simulation! CBs adapt easily within the frameworks of LFV and VV through the well-known SHAKE⁵⁷ and RATTLE⁵⁸ algorithms respectively. Similar constructs are used for the potential of mean force (PMF) constraints. It is worth noting that in DL_POLY_4, appropriate CB and PMF solvers are devised for all of the above ensembles.

RBs may be used to represent structures like aromatic hydrocarbons and their derivatives, which arise in all branches of chemistry. In DL_POLY_4 the dynamical treatment of

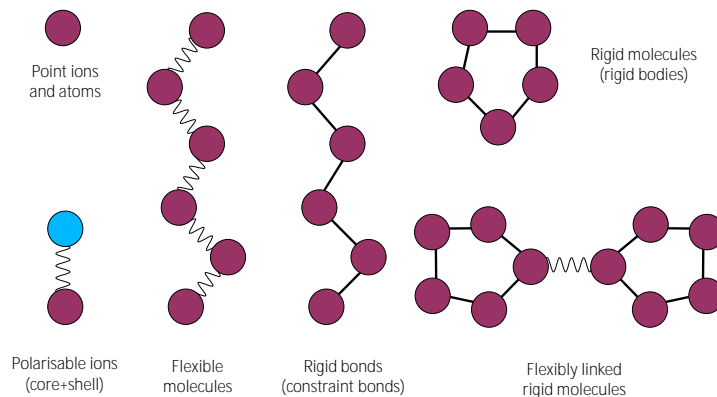


Figure 1. Molecular structures supported by DL_POLY_4. Any or all such structures may be present in a given model at the same time.

such entities is based on Euler’s prescription⁵⁹ treatment of the orientation⁶⁰ augmented by a quaternion. For the LF integration scheme DL_POLY_4 employs the Fincham implicit quaternion algorithm⁶¹ and for the VV scheme the NOSQUISH algorithm of Miller et al.⁶² is used. The latter algorithm has the advantage of being symplectic and therefore stable for long time integrations.⁶²

6 Parallelisation

DL_POLY_4 parallelism relies on equi-spatial domain decomposition (DD), in which the simulation cell is divided spatially into quasi-independent domains which are allocated to individual processor cores (MPI tasks). It follows immediately that in order to have reasonable work load balancing the simulated system must be reasonably isotropic during the simulation. The spatial division naturally does not recognise molecular entities, which are therefore usually divided between processors, creating special communication difficulties. The implementation of DD in DL_POLY_4 is based on Hockney and Eastwoods link cell (LC) algorithm,⁴¹ which was adapted for parallel use by Pinches et al.⁶³ and Rapaport.⁶⁴ A LC approach is not entirely essential for DD, but it provides useful constructs to aid its implementation and yields order N scaling for large numbers of atoms, N . The structural aspects of DD are shown in Figure 2 (a).

The MD cell is most often divided into near-cubic domains, though exception is made for systems with slab geometries to help achieve load balance. Each domain is then subdivided into LCs according to the normal prescription, in which the width of a LC must be greater than the cut-off distance applied to any one inter-atomic interaction. This criterion must also include the distances between particles participating in any *intra*-molecular

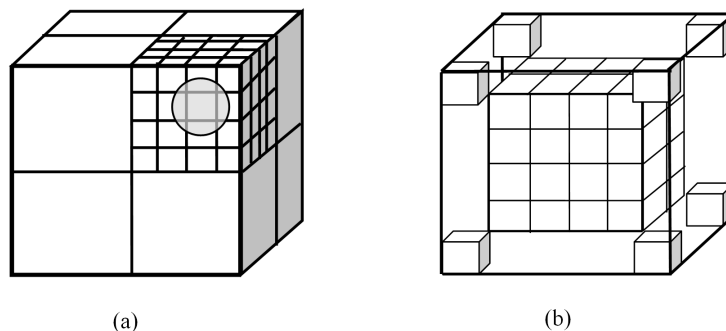


Figure 2. (a) Domain Decomposition (*left*). The MD cell (large cube) is divided into equal-sized domains (medium-sized cubes), each of which is allocated to a specific processor. Each domain is divided into link-cells (small cubes), the width of which must be greater than the radius of interaction cut-off applied to the interatomic force terms. The shaded circle represents the cut-off sphere defining the interaction range. (b) Halo data construction in Domain Decomposition (*right*). The central cube represents a spatial domain that is allocated to a single processor, where it is divided into link-cells (small cubes). It is necessary to add the halo data around the domain, which is one link-cell in width (indicated by the isolated small cube), as it is composed of the coordinates of atoms found in the link-cells at the boundaries of the neighbouring domains. It is apparent from this construction that the smaller the link cells, the more efficient the overall algorithm will be, since less data will need to be transferred.

interaction, such as dihedral angle (e.g. the 1-4 distance) and any *intra*-molecular (like) object such as a core-shell pair, a CB, a RB or a tether. Ideally, these requirements should lead to better than a $3 \times 3 \times 3$ LC partitioning of the domain in the three principal directions. DL_POLY_4 can handle fewer link cells per domain than this, but such scenarios may raise major efficiency issues arising from the construction of the halo data.

The “halo data” represents the construction around each domain of a partial image of all neighbouring domains so that calculation of all the forces relevant to a domain can take place, as illustrated by Figure 2 (b). In DL_POLY_4 this amounts to the transfer of the atomic coordinates of all atoms located in link cells at the boundaries of a domain to the processors managing the neighbouring domains. This is a six-fold transfer operation that moves data in directions *North & South*, *East & West*, and *Up & Down* of each domain. These six transfers do not happen concurrently, although they may happen in pairs as indicated, since some data sorting is necessary to populate the “corners” of the halo data. It is apparent from the nature of the link-cell method that these transfers are sufficient for a complete calculation of the forces on all atoms in any domain. It is also apparent that if the domains have relatively few link cells (or their shape is far from cubic), then the transfer of the halo data represents the transfer of a major proportion of the contents of a domain, which implies a large, possibly prohibitive, communication cost. This may be avoided by running the program on fewer processors. The transfer of halo data is the main communication cost of the basic DD strategy. After the transfer, the atomic forces may be calculated and the equations of motion integrated independently on each processor. Atoms that move sufficiently far may then be reallocated to a new domain.

The computation of *inter*-molecular forces, such as those for van der Waals (VDW), comes straightforwardly from the LC decomposition of the domain and its halo by

constructing a distributed Verlet neighbour list⁴³ (VNL) - enlisting all possible pairs on the domain and its halo within a cut-off distance. Special care is taken that the VNL excludes pairs with both particles lying in the domain halo./

There are particular complications arising from the DD scheme related to the computation of *intra*-molecular forces and the handling of *intra*-molecular objects. There are two aspects to this: firstly, the description of the molecular structures (commonly called the *topology*) is “broken” by the decomposition into domains; and secondly, the evolution of the system demands that the topology be partially reconstructed every time atoms move from one domain to another. In order to accomplish this, the package of data transported with each atom that leaves a domain must contain not only its configurational data (*name, index, position, velocity and force*), but also a topological description of the intra-molecular-like term with the atom.

7 Electrostatics

The treatment of long ranged electrostatic forces represents a particular challenge in molecular simulation. Direct summation of the Coulomb pair interactions is rarely adequate, except for the treatment of atomic clusters, so more sophisticated treatments have evolved. The main method used in DL_POLY_4, the Smoothed Particle Mesh Ewald (SPME),⁶⁶ is based on the Ewald sum.⁶⁵

The Ewald sum casts the sum of Coulomb pair interactions into two separate sums (plus a correction term, which is computationally trivial). The first sum is a screened Coulomb sum, which resembles the Coulomb formula but each term is weighted by a screening function (the complementary error function - *erfc*) which compels the sum to converge in a finite range. The second sum is a sum of structure factors, which are calculated from reciprocal space vectors, and which are again weighted by a screening function (this time a Gaussian) which guarantees a finite sum. The first sum is therefore set in *real*-space, while the second is set in *reciprocal*-space. The convergence of both sums is governed by a single parameter α , which defines the range of both convergence functions and is known as the Ewald convergence parameter.

The calculation of the *real*-space components is managed in the same manner as the VDW terms described above. The *reciprocal*-space terms are derived from a Fourier transform of the system charge density. The method involves the global summation of the structure factors associated with each reciprocal space vector.

In the SPME method the charge density is distributed over a regular 3D grid using Cardinal B-splines.⁶⁶ This permits the use of a 3D Fast Fourier Transform (FFT) to calculate the structure factors, which accelerates the process enormously. DL_POLY_4 uses its own 3D FFT algorithm devised by Bush.⁶⁷ Known as the Daresbury advanced Fourier Transform (DaFT), this FFT employs a domain decomposition of the 3D FFT arrays which maps neatly on to the DD structures. This means that all computations necessary to build the (partial) arrays can take place without inter-processor communication. Furthermore, all communication required by the FFT algorithm is handled internally. While the insertion of communication processes into the heart of the FFT algorithm inevitably affects the efficiency of the FFT calculation, DaFT nevertheless possesses excellent scaling characteristics and the associated economies in data management resulting from its use makes the DL_POLY_4 SPME implementation a highly efficient algorithm.⁶⁸ Overall, due to the

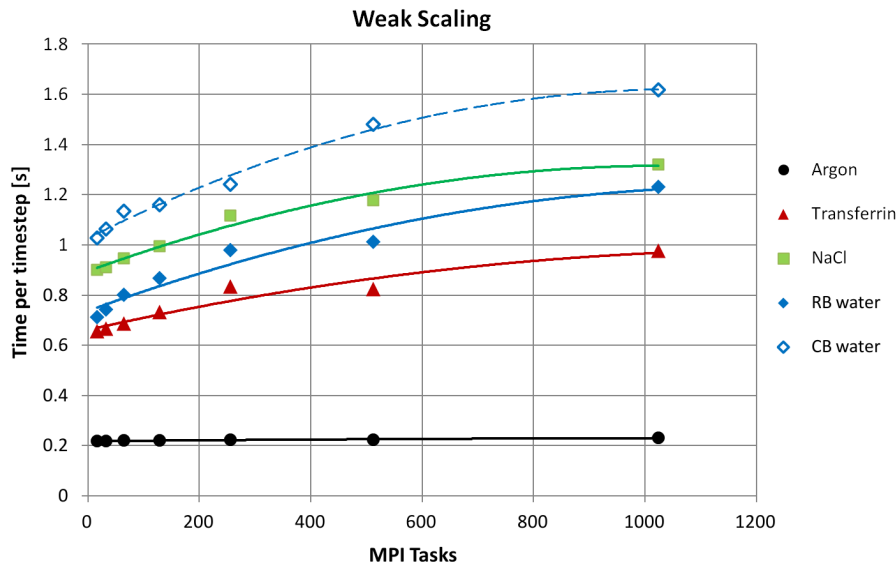


Figure 3. DL.POLY_4 weak scaling performance using five model systems with varied force field complexity as described in the text. Computation times (time-per-timestep) exclude any start-up and close-down timings, and are based on a single, non-reserved run per system for a 100 timesteps of evolution from equilibrium within the velocity Verlet (VV) couched microcanonical ensemble (NVE) on a Cray XE6 platform.⁷⁰

DaFT internal communication and FFT scalability, the electrostatics evaluation is usually one of the most expensive parts of an MD step; it yields order $N \log N$ compute-wise and $(N \log P)/P$ communication-wise, where N is the number of particles in the model system and P the number of domains. Radiation damage simulations of order 10 million atoms (and larger) are regularly performed with DL.POLY_4.⁶⁹

It is worth noting that DL.POLY_4 also offers a number of cheaper alternatives to handle electrostatic interactions than the default SPME. These are the direct Coulomb sum, Coulomb sum with distance dependent dielectric, force-shifted Coulomb sum and reaction field Coulomb sum.⁷² The last two are optionally extended to include screening effects as demonstrated in.⁷¹ The alternative approaches should be used with caution as they are specific short-ranged approximations of the Ewald sum!

8 Scalability and Performance

Figure 3 is an example that best demonstrates the scalability and performance of DL.POLY_4. It represents a comparative weak scaling^a test on a set of model systems with increasing complexity of their force fields (FFs).

The Argon system includes only short-ranged VDW interactions. The NaCl system increases the Argon system's FF complexity by including long-ranged electrostatic inter-

^aIn a weak scaling test, the ratio of the problem size to the compute power (number of cores) is kept constant, i.e. the system size is enlarged by a factor of two every time the core count is increased by a factor 2.

actions - handled with SPME with a convergence factor of $\alpha = 10^{-6}$. The remaining systems used an SPME convergence factor of $\alpha = 10^{-5}$. The transferrin system (representing polymer chains solvated in water) increases further the FF complexity by including a wealth of *intra*-molecular interactions together with constraint bonds - with a relative length convergence factor of $\sigma = 10^{-5}$. Lastly, the CB water and the RB water systems use the same single point charge water model FF SPC-E, but in the former the two O-H and H-H bonds of each water molecule are handled as constraint bonds ($\sigma = 10^{-5}$), whereas in the latter all water molecules are handled as rigid bodies^b.

For purposes of comparison the systems were constructed to nearly the same size, starting at $\approx 250,000$ particles on 16 cores (i.e. domains/MPI tasks), and using the same short-ranged cutoff of 9 Å so that all algorithms related to handling short-ranged interactions (LC and VNL) and minimum necessary communications (halo exchange) were in linear-scaling regimes. Therefore, by performing the weak scaling test on these systems we clearly expose trends related to the impact on communication and computation overheads driven by the complexity of the model system in terms of force field related features.

First, it is clear that the Argon system performs best due to the presence of only short-ranged interactions. Its weak scaling is hardly affected by number of cores. The NaCl system weak scaling deviates from that trend but its relative cost, in terms of time per timestep, is almost the highest. The deviation is due to the non-linear compute and communication performances of the 3D FFT routine (DaFT) needed for the SPME electrostatics (as discussed in Section 7). The transferrin system has the force field that is richest in features and shows similar weak scaling to that of the NaCl system. However, its absolute cost is much lower. This is by and large due to the lesser accuracy of the Ewald sum (i.e. the larger convergence factor) that was used for the simulations. Last but not least the two water systems weak scaling lines reveal that the cost of constraint bond solvers is much larger than that of the rigid body solvers. Thus RB dynamics offers better communication-bound computation than CB dynamics. As discussed in,⁷³ in the case of the CB water system, the communication overheads rise quickly with core count and start dominating almost immediately over computation in strong scaling tests, which is not the case for the RB water system. Thus RB dynamics offer a constant communication-bound computation.

9 I/O Files and Performance

To run an MD simulation using DL_POLY_4 a minimum of three files are required:

- CONFIG - contains configuration information; crystallographic and optionally some dynamic data about the MD cell in data records following a well-documented standard. Briefly, the data includes a type of lattice image condition, lattice parameters, level of information for the particles within, and a list of all particles with name/type ($name_i$), global index ($i = 1, \dots, N$), and coordinates (x_i, y_i, z_i). Depending on the level of information, particles velocities (vx_i, vy_i, vz_i) and forces (fx_i, fy_i, fz_i) may be optionally included too.

^bIt is worth noting that both integration methodologies, RB and CB, solve the same degrees of freedom per water molecule. For CB dynamics this is 3×3 (for the three atoms) $- 3$ (for the three distance constraints O-H1, O-H2, H1-H2) or 6 in total. For RB dynamics this is also 6: 3 for the center of mass translational motion and 3 for the unrestricted rotational motion of the water molecule.

- FIELD - describes the necessary force field information complementing the configuration given in CONFIG. It contains physical, stoichiometrical, optionally *intra*-molecular “like” (topology and interactions), *inter*-molecular (interactions) and external field information. The information must correspond to the contents in CONFIG. Optionally, if so indicated in FIELD, some types of interactions may be provided in a tabulated form in extra files such as TABLE and/or TABEAM.
- CONTROL - contains simulation control directives and conditions, such as timestep size, type of ensemble and target state point, minimisation and runtime options, etc.

After running DL_POLY_4 successfully, at least four output files are produced:

- OUTPUT - contains run time information of the simulation, such as timestep timing data, instantaneous values of measurable quantities such as pressure, temperature, energy components, etc. as well as rolling statistical averages of these and some final information when the simulation comes to an end. Warning and error messages may also be found in this file.
- STATIS - contains all instantaneous data dumped at regular intervals during the simulation.
- REVCON - contains the final configuration data about the MD cell in the same format as CONFIG. It is dumped at regular intervals, as it is necessary as a back-up solution in case of prematurely terminated runs and at the end of a run for restart purposes.
- REVIVE - statistical accumulators holding data necessary to restart and continue a previous simulation from where it last finished or at its last back-up point. Notice that this file is written in binary!

There are a number of other optional files depending on what is specified in the CONTROL file. HISTORY is one which deserves a mention because it is often used as input for visualisation software packages such as VMD⁷⁵ to produce an animation of a simulation run. HISTORY contains instantaneous configuration data (similar to the CONFIG file) about the MD cell dumped at regular intervals during the simulation.

As discussed elsewhere⁷⁴ DL_POLY_4 has an advanced parallel I/O strategy. The consequences of poor I/O parallelisation can be catastrophic in the limit of large system sizes or/and large processor counts. There are a few simple requirements that guided the development:

- To avoid disk contention, the data is gathered and apportioned to a subset of the processes within the job. These, the I/O processors, then perform the reading or writing.
- To avoid disk and communication fragmentation, large sizes of I/O and data transactions are required whilst being small enough to fit in the available memory and bandwidth.
- Keep data contiguous along the particle indices when writing data. The domain decomposition (DD) of the configuration data presents the scrambling data problem: as particles naturally diffuse they move from one domain to another. Therefore, there is no straight mapping of the order of reading to the order of writing via the DD map.

This has implications for the processing of configuration data by many atomistic visualisation software packages.

The practical benefits of the parallel I/O in DL.POLY_4 are that dumping of a configuration frame on disk costs, in terms of computation time, the same order of magnitude as a timestep provided the I/O options supplied by the user in the CONTROL file are tuned for the particular architecture. If no specifications are given, reasonable preset defaults are loaded to match best performance on a Cray XE6 platform.

It is worth mentioning that reading a configuration frame is also very fast, although not as fast as the writing. However, this operation is carried out as a one-off only at the start of a simulation run. The treatment of FIELD can be more problematic as the topological information can only be digested in serial.

10 GridBeans and Workflows

Packages such as DL.POLY_4 undoubtedly serve as versatile and powerful tools to help elucidate the properties and behaviour of a diverse range of materials. However, the scope of an individual application is typically quite narrow in terms of the physical processes it can describe. This is mainly due to the inherent limitations of the physical model that a particular application is based on. For example DL.POLY_4, as a classical molecular dynamics (MD) engine, could never be used to probe phenomena of a quantum nature such as light emission or chemical bonding. The suitability of a particular simulation tool is also constrained by the practical consideration of what is feasible computationally. For example classical MD could, in principle, be used to simulate systems of macroscopic dimensions, but of course the computational resources and time that would be required for such a task make this utterly unrealistic. Nevertheless, the properties of materials are dependent upon a range of processes that collectively span a very broad range of time and length scales. Therefore, in recent years there has been growing interest in finding ways to integrate different models to collectively bridge this gulf.

The Multiscale Modelling of Materials on High Performance Computers (MMM@HPC)⁷⁶ is one of a number of recent projects that seek to meet this challenge. The particular focus of MMM@HPC is to couple together quantum, atomistic and continuum level simulation techniques and calculations to fully model the behaviour of devices such as OLEDs^c, polymer and graphene based electronics and lithium ion batteries. As well as addressing the scientific questions of how to combine the various underlying physical models, the project has also sought to develop and promote suitable methods for implementing the multiscale models on Grid-based compute resources. These methods are built on UNICORE^{d,77} a well established Grid infrastructure that links HPC facilities across Europe. The remainder of this section will present a brief outline of this infrastructure and how it can be used as an efficient and flexible means of running multiscale models in the form of workflows.

The UNICORE infrastructure, as illustrated in Figure 4, consists of three layers – client, server and system. The client is installed on the user’s local workstation, which could be

^cOrganic Light-Emitting Diode

^dUniform Interface to Computing Resources

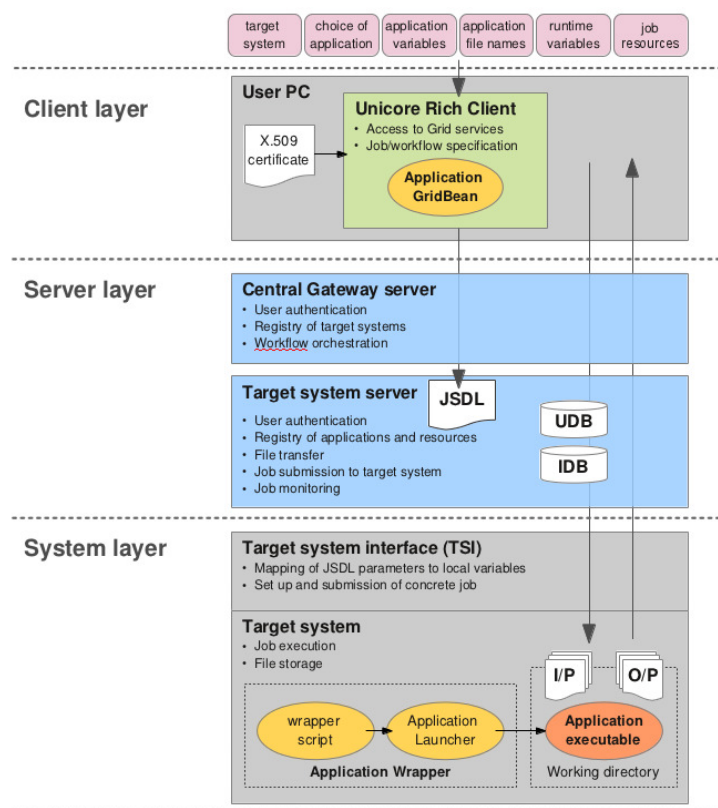


Figure 4. Simplified schematic of the UNICORE Grid infrastructure for job and workflow submission.

virtually any PC or laptop computer, whilst the servers are based at various geographically dispersed HPC facilities. Individual servers are accessed by the client via a central Gateway. A given facility may run any number of servers but each one is associated with a single high performance machine – a so-called target system. All copies of the client and server, which are Java based for portability, are essentially the same. Target systems on the other hand vary greatly in terms of architecture, operating systems, queuing systems and so forth. Therefore, each target system has a program called the Target System Interface (TSI) installed on it that interprets the commands and information that the server passes to it.

In essence, the relationship between these key components of the infrastructure is the following: The user inputs information about the simulation they wish to run – the job – into the client. The client converts the user’s specification into Job Submission Description Language (JSDL), which is written to a single file; this is technically referred to as an abstract job. The JSDL file is sent to the server where the information it contains is unpacked and used to configure instructions for running the job. These instructions are then passed to the target system, via the TSI, along with any files that are required as inputs to the job. The target system is responsible for setting up a working directory for the job and submitting

it to the queuing system. The server monitors the progress of the job and reports its status to the client. When the job is finished, output files can be fetched from the job directory to the users local machine.

The unfolding of this chain of events is contingent on two main conditions: (i) that the user has access to the target system; and (ii) that the application executable they wish to run is installed there. Access to the UNICORE Grid infrastructure as a whole requires that the user is in possession of an X.509 Grid certificate. The certificate is embedded within the client and, when it connects to the central gateway, its authenticity is verified. For access to a particular target system, the user needs a user account on that system. A user database (UDB) on the target system's server contains a register of all the account holders and maps the certificate to the user credentials on the target system. A second database, the Incarnation Database (IDB) contains a register of all applications installed on the target system as well as a list of all the resources it offers, the latter are mostly related to the hardware set-up, e.g. number of cores, cores per CPU, CPUs per node etc.

We will now describe in a little more detail how the client is actually used to submit jobs to a target system. It should be added that there are two versions of the client, the UNICORE Command line Client (UCC) and the UNICORE Rich Client (URC). The former is named simply because it is controlled via command line directives whilst the latter is based on the Eclipse Rich Client,⁷⁸ a generic Java based GUI. Here we confine the discussion to the operation of the URC.

Figure 5 shows a typical view of the URC, in so called workbench mode – the mode of operation in which it is used to submit jobs. It is divided into three main sections:

1. The “Grid browser” is a view of all facilities connected to the UNICORE Grid, here the user selects the particular target system on which to run a job.
2. The job editing pane, where specifications for an individual job are entered.
3. On the lower left are a collection of various tabs, by default these are: the “Navigator” in which the files associated with all the specified jobs can be browsed, the “Keystore” and “Truststore” which contain information about the security related certificates installed in the client and the “Client Log”, which contains a record of information, warnings and error messages issued by the client as it runs.

To create a job, the user begins by selecting a target system in the Grid Browser^e, then, by right clicking on its icon, they can select from a drop-down menu the application they wish to run on it. Only applications listed in the IDB of the target system server will appear in the menu. When an application is selected, various tabs will appear in the job editing area. Generally there are four of these, from left to right: (i) input panels, (ii) files, (iii) variables and (iv) resources, such as number of processors. The variables tab is generic and allows environment variables to be exported to the target system's operating system before the job is run. The options available in the resources tab depend on what is available on the target system selected for the job and correspond to the system information listed in the IDB.

The configuration of the input panels and files tabs on the other hand reflect the identity of the particular application selected for the job. Typically, the input panels are used to enter

^eNote that Grid resources must first be added to the URC by manually specifying a web address or by automatic discovery. Access requires the user to possess a valid Grid certificate that needs to be imported into the URC.

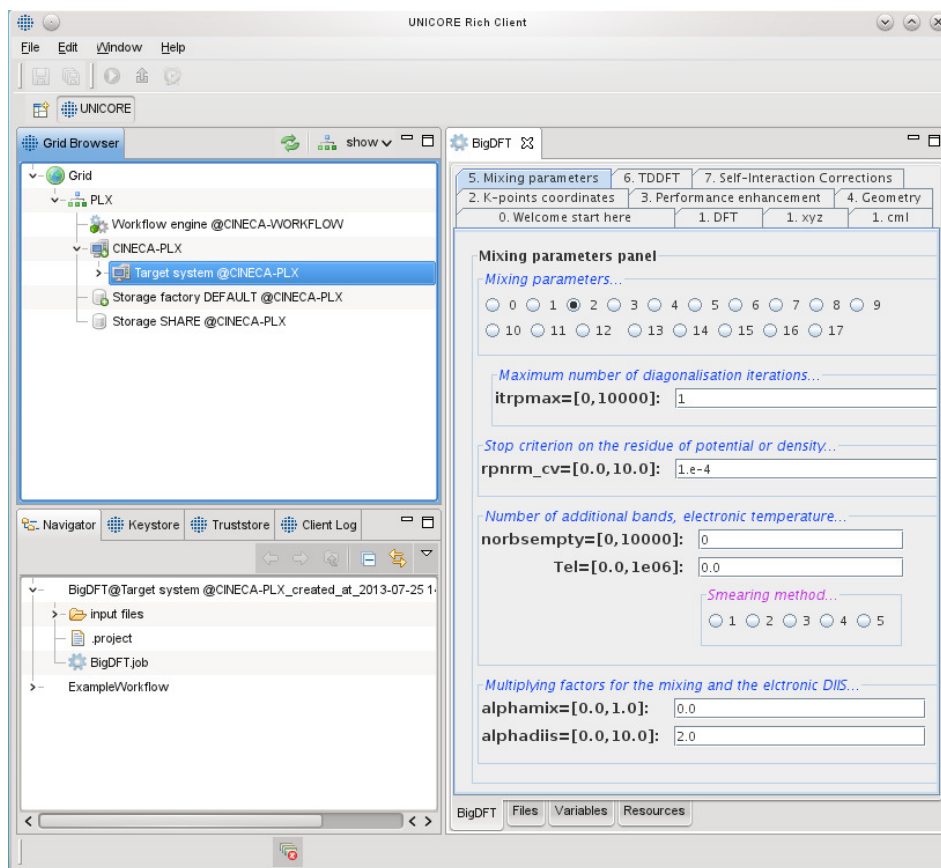


Figure 5. Snapshot of the UNICORE Rich Client. In the job editing pane on the right, the input parameters for a BigDFT job are in the process of being specified.

information that controls how a simulation is run, such as the number of time steps to run for, which algorithm to use and so forth. In the files tab the input and output files for the application are defined, in the case of the input files, paths to their storage locations must be specified. Input files can be stored on the user's local workstation or in a storage area on the target system. The software component embedded in the client that is responsible for the appearance of the input panels and files tabs is called a GridBean. In a sense the GridBean encapsulates the application, in as much as it controls the information that is fed into its executable when it is run on the target system. When the user has entered all the information required to specify his job, they 'save' the job with a single click of an icon on the main menu, this causes the URC to create the JSDL file. Clicking the run icon in the main menu then submits the JSDL file to the target system server. The server subsequently passes instructions to the target system itself to execute the 'concrete' job so to speak, as described above. The user can monitor the status of the job – 'submitted', 'running', or 'finished' and, at any point after submission, inspect the contents of the job's working

directory on the target system in the Grid browser. When the job has finished, the output files can be uploaded to the local machine at the click of an icon.

It should be noted that in order for the job to be executed by submitting via a Grid-Bean, a so called Application Wrapper must be present on the target system. The Application Wrapper, essentially, consists of two components: a bash script, called the wrapper script, and a jar file (compiled Java source code) which we will refer to as the Application Launcher. The location of the script on the target system is listed in the IDB so that it can be accessed when the time comes to run the job. It is usually responsible for three things:

1. Loading any modules or libraries required by the application executable;
2. Gathering any required environment variables present on the operating system; and finally
3. Running the Application Launcher, passing the environmental variables to it as it does so.

The primary role of the Application Launcher is simply to issue the command that will cause the target system to submit the job to the queue. However, it can be augmented to perform a number of ancillary functions such as logging diagnostics, pre-processing (e.g. converting one or more of the input files into another format), and, when the job has run, carrying out post-processing on the output files.

Setting up and running standalone jobs as GridBeans on the URC saves the user considerable effort in respect to routine tasks such as manually uploading files, setting up working directories, managing output data and keeping records of all the jobs they have run. Running jobs in this way also makes it unnecessary to learn about the procedures and protocols of the target system, since the user never interacts with it directly. Also, once a job has been specified in the URC, it can be re-run as many times as desired with the minimum of effort. Furthermore, GridBeans for an increasing number of applications are being developed and committed to a central repository from which they can be downloaded by other users into their own client.

So clearly GridBeans are of great utility when it comes to running standalone jobs. However, where they really come into their own is in the construction of workflows. A workflow, in its simplest form, consists of a series of simulation steps carried out one after the other, with output from one step forming the input to the next step. Thus workflows can be used to implement multiscale models. The task of implementing such schemes manually, even just once, would be an arduous one – a number of standalone jobs would need to be set up (possibly on different target systems), monitored and, when each job has finished, the outputs transferred to the working directory of the next. To have to repeat the process numerous times would be highly impracticable and likely to be prone to human error.

Fortunately, the URC allows the user to easily set up workflows graphically in the form of flowcharts and then submit them in their entirety as easily as a single job. In essence, they are constructed by connecting together individual GridBeans in the desired sequence and then setting the specifications for each job. This is done by first right clicking on the so-called workflow engine in the Grid browser and opening a new workflow project. The workflow engine is a software component installed on the central gateway server mentioned earlier. Its role is essentially to orchestrate the distribution of individual jobs to

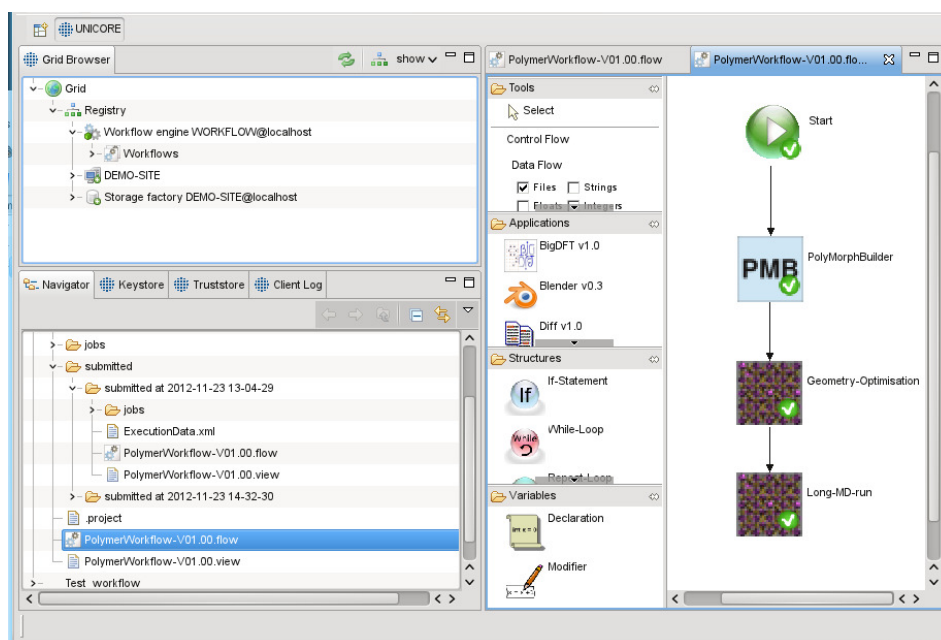


Figure 6. A simple UNICORE workflow consisting of three steps: (i) the application PolyMorphBuilder sets up an initial configuration of polymer molecules; (ii) a short DL_POLY_4 run to perform a geometry optimisation on the initial configuration; and (iii) a long DL_POLY_4 run to generate an accurate morphology for the polymer system at a certain temperature.

particular target systems, thus allowing particular steps in the workflow to be matched to the most appropriate resources. Upon starting a new workflow project, an editing panel into which icons representing the desired application GridBeans are dragged from a side menu. A simple example is shown in Figure 6. More complex workflows can be constructed with the use of logical structures such as conditional statements and loops. Some care has to be taken in specifying the input and output files for each step, but once a particular workflow is set up, the process does not need to be repeated. As with individual jobs, workflows can be duplicated, modified and reused as required. Also, workflow templates can be committed to a central repository and thereby become available to other UNICORE users.

11 Concluding Remarks

We have given a comprehensive overview of the DL_POLY_4 program, outlining the concepts and methodologies that have driven its development and made it into a generic and comprehensive toolbox for many molecular dynamics modellers world-wide. We have also described the generic manner in which one could include the program within a UNICORE server-client framework and get access to workflow automation with the future possibility to include multiscale workflows.

More information about the DL_POLY_4 program and its usage can be found on the project website,⁹ where we provide links to the software, its manual and test-cases, the user forum, and least but not last some training material.

For more detailed technical descriptions of the UNICORE infrastructure components the interested reader can download technical manuals from the UNICORE website.⁷⁷

For more information about GridBeans and workflows as well as the MMM@HPC project in general, the reader is again directed to the MMM@HPC project website,⁷⁶ where the project deliverables can be downloaded.

Acknowledgments

The authors are grateful to all funding bodies and networks, as outlined in Section 1, that have provided continual support to the DL_POLY project. The authors would like to acknowledge the MMM@HPC workflow consortium that was funded by the 7th Framework Programme of the European Commission within the Research Infrastructures with grant agreement number RI-261594.

References

1. CCP5 is the Collaborative Computational Project for computer simulation of condensed phases. <http://www.ccp5.ac.uk/>
2. The Engineering and Physical Sciences Research Council (EPSRC) of Great Britain. <http://www.epsrc.ac.uk/>
3. The Natural Environment Research Council (NERC) of Great Britain. <http://www.nerc.ac.uk/>
4. W. Smith, and T. R. Forester, *DL_POLY 2.0: A general-purpose parallel molecular dynamics simulation package* J. Mol. Graphics **14**, 136, 1996.
5. W. Smith, C. Yong, and P. M. Rodger, *DL_POLY: Application to Molecular Simulation*, Mol. Simulation. **28**(5), 385, 2002.
6. I. T. Todorov, and W. Smith, *DL_POLY_3: the CCP5 national UK code for molecular-dynamics simulations*, Phil. Trans. Roy. Soc. Lond. A **362**, 1835–1852, 2004.
7. I. T. Todorov, W. Smith, K. Trachenko, and M. T. Dove, *DL_POLY_3: new dimensions in molecular dynamics simulations via massive parallelism*, J. Mater. Chem. **16**, 1911–1918, 2006.
8. The DL_POLY_Classic Molecular Simulation Package. http://www.ccp5.ac.uk/DL_POLY_CLASSIC/
9. The DL_POLY Molecular Simulation Package. http://www.ccp5.ac.uk/DL_POLY/
10. NAGWare Home Page. <http://www.nag.co.uk/nagware.asp>
11. FORCHECK Home Page. <http://www.forcheck.nl/>
12. NVIDIA Home Page. <http://www.nvidia.com/>
13. The DL_FIELD Force Field Generator for DL_POLY. http://www.ccp5.ac.uk/DL_FIELD/
14. R. Salomon-Ferrer, D. A. Case, and R. C. Walker, *An overview of the Amber biomolecular simulation package*, WIREs Comput. Mol. Sci. **3**, 198–210, 2013.

15. Amber Home Page. <http://ambermd.org/>
16. E. Lindahl, B. Hess, and D. van dam Spoel, *GROMACS 3.0: a package for molecular simulation and trajectory analysis*, J. Mol. Mod. 7 (8) **306–317**, 2001, .
17. Gromacs Home Page. <http://www.gromacs.org/>
18. J. C. Phillips, R. Rosemary Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kal, and K. Schulten, *Scalable molecular dynamics with NAMD*, J. Comp. Chem. **26 (16)**, 1781–1802, 2005.
19. NAMD Home Page. <http://www.ks.uiuc.edu/Research/namd/>
20. 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. Comput. Chem. **4 (2)**, 187–217, 1983.
21. CHARMM Home Page. <http://www.charmm.org/>
22. W. L. Jorgensen, and J. Tirado-Rives, *The OPLS Force Field for Proteins. Energy Minimizations for Crystals of Cyclic Peptides and Crambin*, J. Am. Chem. Soc. **110 (6)**, 1657–1666, 1988.
23. S. L. Mayo, B. D. Olafson, and W. A. Goddard, *DREIDING: a generic force field for molecular simulations*, J. Phys. Chem. **94 (26)**, 8897–909, 1990.
24. H. Sun, S. J. Mumby, J. R. Maple, and A. T. Hagler, *An ab Initio CFF93 All-Atom Force Field for Polycarbonates*, J. Amer. Chem. Soc. **116**, 2978–2987, 1994.
25. B. G. Dick, and A. W. Overhauser, *Theory of dielectric constants of alkali halide crystals*, Phys. Rev. B **112**, 90, 1958.
26. D. Fincham, and P. J. Mitchell, *Shell model simulations by adiabatic dynamics*, J. Phys. Condens. Matter **5**, 1031, 1993.
27. P. J. D. Lindan, and M. J. Gillan, M.J., *Shell-model molecular-dynamics simulation of superionic conduction in CaF₂*, J. Phys. Condens. Matter **5**, 1019, 1993.
28. J. Tersoff, Phys. Rev. B **39**, 5566, 1989.
29. T. Kumagai, S. Izumi, S. Hara, and S. Sakai, Comput. Mat. Sci. **39**, 457, 2007.
30. M. W. Finnis, and J. E. Sinclair, Philos. Mag. A **50**, 45, 1984.
31. A. P. Sutton, and J. Chen, Philos. Mag. Lett. **61**, 139, 1990.
32. H. Rafii-Tabar, and A. P. Sutton, Philos. Mag. Lett. **63**, 217, 1990.
33. X. D. Dai, Y. Kong, J. H. Li, and B. X. Liu, J. Phys.: Condens. Matter **18**, 45274542, 2006.
34. F. Cleri, and F. Rosato, Phys. Rev. B **48**, 22, 1993.
35. M. S. Daw, and M. I. Baskes, Phys. Rev. B **29**, 6443, 1984.
36. S. M. Foiles, M. I. Baskes, and M. S. Daw, Chem. Phys. Lett. **33**, 7983, 1986.
37. D. J. Hepburn, and G. J. Ackland, Phys. Rev. B **78(16)**, 165115, 2008.
38. T. T. Lau, C. J. Först, X. Lin, J. D. Gale, S. Yip, and K. J. V. Vliet, Phys. Rev. Lett. **98(21)**, 215501, 2007.
39. G. J. Ackland, and S. K. Reed, Phys. Rev. B **67**, 1741081–1741089, 2003.
40. P. Olsson, J. Wallenius, C.; Domain, K. Nordlund, and L. Malerba, Phys. Rev. B **72**, 2141191–2141196, 2005.
41. R. W. Hockney, and J. W. Eastwood, *Computer Simulation Using Particles* (McGraw-Hill, New York 1981).
42. 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*, Journal J. Chem. Phys. **76**, 1982.

43. M. P. Allen, and D. J. Tildesley, *Computer Simulation of Liquids* (Oxford: Clarendon Press 1989).
44. D. J. Evans, and G. P. Morriss, *Computer Physics Reports* **1**, 297, 1984.
45. S. A. Adelman, and J. D. Doll, *J. Chem. Phys.* **64**, 2375, 1976.
46. J. A. Izaguirre, *Langevin stabilisation of multiscale mollified dynamics* (Editors: A. Brandt A., K. Binder, and J. Bernholk) *Multiscale Computational Methods in Chemistry and Physics*, volume 117 of *NATO Science Series: Series III - Computer and System Sciences* pages 34–47. (IOS Press, Amsterdam 2001).
47. H. C. Andersen, *J. Chem. Phys.* **72**, 2384, 1979.
48. H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak, *J. Chem. Phys.* **81**, 3684, 1984.
49. W. G. Hoover, *Phys. Rev. A* **31**, 1695, 1985.
50. B. Leimkuhler, E. Noorizadeh, and F. Theil, *J. Stat. Phys.* **135**, 261–277, 2009.
51. A. Samoletov, M. A. J. Chaplain, and C. P. Dettmann, *J. Stat. Phys.* **128**, 13211336, 2007.
52. D. Quigley, and M. I. J. Probert, *J. Chem Phys.* **120**, 11432, 2004.
53. S. Melchionna, G. Ciccotti, and B. L. Holian, *Molec. Phys.* **78**, 533, 1993.
54. G. M. Martyna, D. J. Tobias, and M. L. Klein, *J. Chem. Phys.* **101**, 4177, 1994.
55. G. M. Martyna, M. E. Tuckerman, D. J. Tobias, and M. L. Klein, *Molec. Phys.* **87**, 1117, 1996.
56. M. Ikeguchi, *J. Comp. Chemi* **25**, 529–541, 2004.
57. J. P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen, *Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes*, *J. Comput. Phys.* **23**, 327, 1977.
58. H. C. Andersen, *Rattle: a velocity version of the SHAKE algorithm for molecular dynamics calculations*, *J. Comput. Phys.* **52**, 24, 1983.
59. H. Goldstein, *Classical Mechanics* (Addison Wesley, 1980).
60. D. J. Evans, *On the representation of orientation space*, *Mol. Phys.* **34**, 317, 1977.
61. D. Fincham, *Leapfrog rotational algorithms*, *Mol. Simul.* **8**, 165, 1992.
62. T. F. Miller, M. Eleftheriou, P. Pattnaik, A. Ndirango, D. Newns, and G. M. Martyna, *Symplectic quaternion scheme for biophysical molecular dynamics*, *J. Chem. Phys.* **116**, 8649, 2002.
63. M. R. S. Pinches, D. Tildesley and W. Smith, *Large scale molecular dynamics on parallel computers using the link cell algorithm*, *Mol. Simul.* **6**, 51, 1991.
64. D. C. Rapaport, *Multi-million particle molecular dynamics. II. Design considerations for distributed processing*, *Comput. Phys. Comm.* **62**, 217, 1991.
65. C. Kittel, *Solid State Physics*, (John Wiley and Sons 1986).
66. U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen, *A smooth particle mesh Ewald method*, *J. Chem. Phys.* **103**, 8577, 1995.
67. I.J. Bush, *The Daresbury Advanced Fourier Transform*, (Daresbury Laboratory 1999).
68. I. J. Bush, I. T. Todorov, and W. Smith, *A DAFT DL POLY distributed memory adaptation of the smoothed particle mesh Ewald method*, *Comput. Phys. Commun.* **175**, 323329, 2006.
69. K. Trachenko, M. T. Dove, E. Artacho, I. T. Todorov, and W. Smith, *Atomistic simulations of resistance to amorphization by radiation damage*, *Phys. Rev. B* **73**, 174207, 2006.

- 70. HECToR: UK National Supercomputing Service.
<http://www.hector.ac.uk/>
- 71. C. J. Fennell, and D. J. Gezelter, J. Chem. Phys. **124**, 234104, 2006.
- 72. M. Neumann, J. Chem. Phys. **82**, 5663, 1985.
- 73. I. T. Todorov, L. J. Ellison and W. Smith, *Rigid Body Molecular Dynamics within the Domain Decomposition framework of DL-POLY-4* (accepted in PaCT-2013 proceedings).
- 74. I. J. Bush, I. T. Todorov, and W. Smith, *Optimisation of the I/O for Distributed Data Molecular Dynamics Applications*, (CUG 2010).
- 75. VMD home page. <http://www.ks.uiuc.edu/Research/vmd/>
- 76. Multiscale Modelling of Materials on High Performance Computers (MMM@HPC) home page. <http://www.multiscale-modelling.eu/>
- 77. UNICORE home page. <http://www.unicore.eu/>
- 78. ECLIPSE home page. <http://www.eclipse.org/>

Atomistic Simulations Using the Approximate DFT Method DFTB+: Applications to Nanomaterials and Bio-Systems

Thomas Frauenheim and Bálint Aradi

Bremen Center for Computational Materials Science,
University of Bremen, 28359 Bremen, Am Fallturm 1
E-mail: {thomas.frauenheim, balint.aradi}@bccms.uni-bremen.de

The density functional tight binding (DFTB) method is based on a second-order expansion of the Kohn-Sham total energy in density-functional theory (DFT) with respect to charge density fluctuations. The zero order approach is equivalent to a common standard non-self-consistent (TB) scheme, while at second order a transparent, parameter-free, and readily calculable expression for generalized Hamiltonian matrix elements can be derived. These are modified by a self-consistent redistribution of Mulliken charges (SCC). SCC-DFTB can be successfully applied to problems, where deficiencies within the non-SCC standard TB approach become obvious. It provides accurate results at a fraction of the cost of a DFT evaluation through parameterization of the integrals. Long-range interactions are described with empirical dispersion corrections and the third order approach handles charged systems accurately. Advanced functions include spin degrees of freedom, time dependent methods for excited state dynamics and multi-scale QM/MM-techniques to treat reactive processes in nanostructures under environmental conditions. Additionally, the combination with non-equilibrium Greens functions allows to address quantum transport in nanostructures and on the molecular scale. An overview about the theoretical background of the DFTB method is presented and a showcase example on bulk amorphous titanium oxide to demonstrate its capabilities.

1 Introduction

In the research area of atomistic simulations there is a multiplicity of methods to cover different time and length scales in simulating the time evolution of a given multi-atom ensemble. These simulation techniques use either a quantum mechanical approach, i.e. density functional theory (DFT) or classical molecular dynamics (MD) and kinetic Monte Carlo (kMC) methods. Being computationally very demanding, the simulations using *ab initio* DFT methods are limited to a small number of atoms and short simulation times in the range of a few picoseconds. If questions should be treated taking place on larger length and time scales predominantly classical molecular dynamics (MD) simulations are employed. The most essential input for such classical MD simulations are classical potentials describing all interatomic interactions of the involved species. For this reason the choice of the right potential is of paramount importance and decides on quality and validity of the simulation. Suitable potentials for a desired combination of materials are derived from adapting a classical potential to the results of *ab initio* simulations, by fitting to experimental data or by semi-empirical procedures.

However, since such classical potentials are adapted to a finite set of equilibrium situations (mostly experimental data and *ab initio* results for equilibrium configurations), they usually apply well to systems that are within the parametrization space, but usually fail far from those. Additionally, they are not transferable to different chemical situations and

more generally to calculations of spectroscopic data relying quantitatively on the detailed knowledge of the electronic structure.

In conclusion, for large-scale applications, a method based on quantum mechanics is highly desirable. It should allow one to follow with confidence the structural dynamics during the time evolution in chemical reactions and bond formation. The equilibrium configurations have to be accurately described as regards details of the geometry, and cohesive and elastic properties, including stability as well as vibrational dynamics. The method has to perform equally well for very different types of materials and inherently should yield electronic structure information to enable comparison of theoretical with spectroscopic data. Furthermore, the predictive quantum mechanical treatment of the complex many-atom structures, taking advantage of reliable approximations, should be implemented efficiently for running on advanced computer architectures.

Therefore, during the last two decades we have put a strong effort into the development of approximate methods, which try to merge the spirit and reliability of DFT with the simplicity and efficiency of TB ansätze. In keeping the computational cost but simultaneously also the number of parameters as small as possible, the method described here and related computer codes offer a high degree of transferability as well as universality for both ground-state and excited-state properties. Thus we claim that the density functional tight binding (DFTB) method operates at the same accuracy and efficiency whether organic molecules or solids, clusters, insulators, semi-conductors and metals or even biomolecular systems are investigated, and, furthermore, independent of the type of atoms which constitute the material.

In the next section we give an overview about the DFTB method, showing all its approximations with respect to *ab initio* DFT. Then on the example of bulk amorphous titanium oxides some of the capabilities of the method for materials simulations are demonstrated.

2 Theory of DFTB

2.1 Expansion of the density

The density functional tight binding method derives from *ab initio* density functional theory. According to the Hohenberg-Kohn theorem,¹ the total energy E of a system of electrons in the external field of atomic nuclei is assumed to be a functional of the electron density $n(\mathbf{r})$

$$E = E[n(\mathbf{r})].$$

Applying the Kohn-Sham one-electron approximation,² the density is assumed to be a sum of densities of independent one-electron wavefunctions $\psi(\mathbf{r})$:

$$n(\mathbf{r}) = \sum_i f_i |\psi_i(\mathbf{r})|^2.$$

The factor f_i specifies the occupation number for the given one-particle wavefunction ψ_i and the summation runs over all one-electron wavefunctions.

Following the derivation of Foulkes and Haydock,³ the true ground state electron density in the system can be written as the sum of an arbitrary reference density $n_0(\mathbf{r})$ and the

deviation from that reference density $\delta n(\mathbf{r})$:

$$n(\mathbf{r}) = n_0(\mathbf{r}) + \delta n(\mathbf{r}).$$

Expanding the total energy up to second order in $\delta n(\mathbf{r})$, one obtains the total energy

$$E = E_{\text{bs}}[n_0] + E_{\text{rep}}[n_0] + E_2[n_0, \delta n^2]$$

as sum of three energy terms, the so called band structure energy E_{bs} , the repulsive energy E_{rep} and the second order energy E_2 .

The band structure energy is calculated by summing up the occupation weighted energies of the electrons in the system (ε_i)

$$E_{\text{bs}}[n_0] = \sum_i f_i \varepsilon_i = \sum_i f_i \left\langle \psi_i \left| -\frac{1}{2} \Delta + v_{\text{eff}}[n_0] \right| \psi_i \right\rangle, \quad (1)$$

with

$$v_{\text{eff}} = V_{\text{ext}} + V_{\text{H}} + V_{\text{xc}} = V_{\text{ext}} + \int d\mathbf{r}' \frac{n_0(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} + V_{\text{xc}}$$

being the Kohn-Sham effective potential. (The equations are given in atomic units with Hartree as energy unit.) The potential V_{ext} is the external potential of the nuclei, while V_{H} and V_{xc} stand for the Coulomb-potential and the exchange-correlation potential, respectively, both arising from the electron-electron interaction.

The repulsive energy is defined as

$$E_{\text{rep}}[n_0] = -\frac{1}{2} \int d\mathbf{r} d\mathbf{r}' \frac{n_0(\mathbf{r}') n_0(\mathbf{r})}{|\mathbf{r} - \mathbf{r}'|} + E_{\text{xc}}[n_0] - \int d\mathbf{r} V_{\text{xc}}[n_0] n_0(\mathbf{r}) + \frac{1}{2} \sum_A \sum_{B \neq A} \frac{Z_A Z_B}{R_{AB}}$$

with E_{xc} being the exchange-correlation energy corresponding to the electron density n_0 . The summation over A and B in the last term (nucleus-nucleus repulsion) is carried out over all nuclei in the system, with Z_A and Z_B being the charge of nuclei A and B and R_{AB} the distance between them.

It is important to note, that both E_{bs} and E_{rep} depend only on the reference density $n_0(\mathbf{r})$. In DFTB this reference density is usually composed as a sum of confined electron densities of neutral atoms

$$n_0(\mathbf{r}) = \sum_A n_{\text{atom}}^{t(A)}(\mathbf{r} - \mathbf{R}_A) \quad (2)$$

(with $n_{\text{atom}}^{t(A)}$ being the neutral atomic density used for the atom type $t(A)$ of atom A), so that the sum of these two terms corresponds to the energy of a system consisting of neutral atoms only. In order to take also the effect of the charge transfers between the atoms into account, at least the second order term

$$E_2[n_0, \delta n] = \frac{1}{2} \int d\mathbf{r} d\mathbf{r}' \left[\frac{1}{|\mathbf{r} - \mathbf{r}'|} + \frac{\delta^2 E_{\text{xc}}[n]}{\delta n(\mathbf{r}) \delta n(\mathbf{r}')} \right]_{n_0} \delta n(\mathbf{r}) \delta n(\mathbf{r}')$$

is needed, which depends on the reference density $n_0(\mathbf{r})$ as well as on the density fluctuation $\delta n(\mathbf{r})$.

2.2 Repulsive energy

Having a reference density $n_0(\mathbf{r})$, theoretically both, E_{bs} and E_{rep} can be calculated. In DFTB, however, only the former is calculated explicitly, while the latter is fitted against calculations with higher level methods or against experimental results in order to achieve high accuracy despite the various approximations in the method (described in the following sections). The repulsive energy is assumed to be composed of atom type specific pairwise interactions, so that

$$E_{\text{rep}} = \frac{1}{2} \sum_A \sum_{B \neq A} E_{\text{rep}}^{t(A) t(B)}(R_{AB}).$$

The pairwise repulsive contributions $E_{\text{rep}}^{t(A) t(B)}$ depend on the atom types $t(A)$ and $t(B)$ of the two interacting atoms A and B and on the distance R_{AB} between them. In order to obtain such distance-dependent atom type specific repulsive functions, higher-level (typically *ab initio*) calculations are carried out for systems containing interacting atoms of the given species at various distances. The repulsive functions are then chosen to minimize the weighted difference between the higher level energies and those obtained in DFTB for the given set of atomic structures:

$$\sum_{\alpha} w_{\alpha} |E_{\text{ab initio}}^{\alpha} - (E_{\text{bs}}^{\alpha} + E_2^{\alpha} + E_{\text{rep}}^{\alpha})| = \min.$$

The weights of the individual structures w_{α} can be chosen according to their importance. Apart of the energy, also other quantities (forces, vibration frequencies, etc.) can be taken into account during the fitting procedure. Further details on it can be found in Ref. 4.

2.3 Band structure energy

In order to calculate the band structure energy E_{bs} as defined in equation (1), one has to obtain the one-electron wave functions ψ_i by solving the according one-electron Schrödinger-equation with the Kohn-Sham effective potential

$$H\psi_i = \left[-\frac{1}{2}\Delta + v_{\text{eff}}[n_0] \right] \psi_i = \varepsilon_i \psi_i, \quad (3)$$

with ε_i being the one-electron energies. In DFTB the one-electron wavefunctions are assumed to be a linear combination of atomic orbitals $\varphi_{\nu}(\mathbf{r})$

$$\psi_i(\mathbf{r}) = \sum_{\nu} c_{i\nu} \varphi(\mathbf{r} - \mathbf{R}_{A(\nu)})$$

with coefficients $c_{i\nu}$ to be determined. The sum over ν runs over all atomic orbitals used as basis functions and $\mathbf{R}_{A(\nu)}$ is the position of atom A containing the orbital ν . This turns equation (3) into the generalized matrix eigenvalue problem

$$\sum_{\nu} c_{\nu i} (H_{\mu\nu} - \varepsilon_i S_{\mu\nu}) = 0 \quad \text{with} \quad H_{\mu\nu} = \langle \varphi_{\mu} | H | \varphi_{\nu} \rangle \quad \text{and} \quad S_{\mu\nu} = \langle \varphi_{\mu} | \varphi_{\nu} \rangle, \quad (4)$$

where $H_{\mu\nu}$ and $S_{\mu\nu}$ represent the Hamiltonian matrix and the overlap matrix, respectively. Latter is needed as atomic orbitals centered around different atoms are usually not orthogonal.

The orbitals $\varphi_\mu(\mathbf{r})$ used as basis functions as well as the densities $n_{\text{atom}}^{t(A)}(\mathbf{r})$ used to build the Kohn-Sham effective potential (see equation (2)) are derived from confined neutral atoms. The confinement is typically done using a power confinement potential, so that the atomic orbitals and the atomic density are the solutions of a modified atomic Schrödinger equation

$$\left[-\frac{1}{2}\Delta + v_{\text{eff}}[n_{\text{atom}}] + \left(\frac{r}{r_0} \right)^n \right] \varphi_\mu(\mathbf{r}) = \epsilon_\mu \varphi_\mu(\mathbf{r}).$$

As the atomic density $n_{\text{atom}}(\mathbf{r})$ depends on the wave functions $\varphi_\mu(\mathbf{r})$, the equation must be solved self-consistently. The power of the compression potential n is typically chosen to be 2 or 4. The confinement radius r_0 can be chosen to be different for the confined density and for the confined basis functions.

The basis functions in DFTB usually only contain the valence orbitals of the atoms, enabling to keep the size of the Hamiltonian and overlap matrices rather small. Additionally, in order to be able to calculate the Hamiltonian matrix elements $H_{\mu\nu}$ as efficient as possible, further approximations are made. First of all, the effective potential $v_{\text{eff}}[n_0]$ is written as a sum of atomic contributions

$$v_{\text{eff}}[n_0(\mathbf{r})] = \sum_A v_{\text{eff}}[n_{\text{atom}}^{t(A)}(\mathbf{r} - \mathbf{R}_A)] = \sum_A v_{\text{eff}}[n_{\text{atom}}^A(\mathbf{r})] = \sum_A v_{\text{eff}}^A$$

yielding the Hamiltonian

$$H_{\mu\nu} = \left\langle \varphi_\mu \left| -\frac{1}{2}\Delta + v_{\text{eff}}^A \right| \varphi_\nu \right\rangle + \sum_{B \neq A} \langle \varphi_\mu | v_{\text{eff}}^B | \varphi_\nu \rangle \quad \text{if } \mu, \nu \in A$$

$$H_{\mu\nu} = \left\langle \varphi_\mu \left| -\frac{1}{2}\Delta + v_{\text{eff}}^A + v_{\text{eff}}^B \right| \varphi_\nu \right\rangle + \sum_{C \neq A \neq B} \langle \varphi_\mu | v_{\text{eff}}^C | \varphi_\nu \rangle \quad \text{if } \mu \in A, \nu \in B \neq A.$$

The notation $\nu \in A$ indicates that the orbital φ_ν is centered around atom A . Both sums above (crystal field terms for $\mu, \nu \in A$ and three-center terms for $\mu \in A, \nu \in B \neq A$) are neglected in the DFTB approach. Additionally, the on-site term is replaced by the corresponding energy level in the free *unconfined* atom

$$H_{\mu\nu} = \epsilon_\mu \delta_{\mu\nu} \quad \mu, \nu \in A$$

to ensure the right energy levels in the dissociation limit. Finally, in order to take the non-linearity of the exchange-correlation potential better into account, the effective potential for the two-center interaction is calculated as the potential of the summed atomic densities rather than the sum of the atomic potentials (so called density superposition):

$$v_{\text{eff}}[n_{\text{atom}}^A] + v_{\text{eff}}[n_{\text{atom}}^B] \longrightarrow v_{\text{eff}}[n_{\text{atom}}^A + n_{\text{atom}}^B].$$

With all the approximations described above, the Hamiltonian matrix to be diagonalized has the form

$$H_{\mu\nu} = \epsilon_\mu \delta_{\mu\nu} \quad \text{if } \mu, \nu \in A$$

$$H_{\mu\nu} = \left\langle \varphi_\mu \left| -\frac{1}{2}\Delta + v_{\text{eff}}[n_{\text{atom}}^A + n_{\text{atom}}^B] \right| \varphi_\nu \right\rangle \quad \text{if } \mu \in A, \nu \in B \neq A. \quad (5)$$

This special two-center form allows a very fast build up of the Hamiltonian matrix during the simulations as the various two-center integrals can be calculated in advance and tabulated as a function of distance between the two atomic orbitals in the integral. During the simulation the Hamilton matrix elements are then instantantly calculated by looking up the tabulated values for the given distances between the atoms and transforming the values with simple geometrical transformation into the actual coordinate system.

2.4 Second order energy

The methodology described so far does not take the charge transfers into account and corresponds to the so called non-SCC-DFTB method⁵ (SCC = self consistent charges). As mentioned above, at least a second order energy term in δn must be additionally taken into account in order to be able to describe the charge transfer between the atoms in the system (the deviation from the reference density). In the so called SCC-DFTB method⁶ this is calculated in the monopole approximation by the Coulomb-like expression

$$E_2 = \frac{1}{2} \sum_A \sum_{B \neq A} \gamma_{AB} \Delta q_A \Delta q_B,$$

where Δq_A and Δq_B indicate the difference in the electron population on atoms A and B with respect of the reference neutral atoms. The Hamiltonian (5) must be corrected accordingly by adding the correction

$$H_{\mu\nu}^2 = \frac{1}{2} S_{\mu\nu} \sum_{C \neq A \neq B} (\gamma_{AC} + \gamma_{BC}) \Delta q_C$$

to it. The electron populations on the individual atoms are calculated by Mulliken analysis.⁷ Since this requires the knowledge of the one-electron wave functions ψ_i (the knowledge of the coefficients $c_{\nu i}$), the eigenvalue problem (4) must be solved in a self consistent manner. Starting with some chosen initial atomic charges $\Delta q_A^{(0)}$ the Hamilton matrix $H_{\mu\nu}^{(0)}$ is built up and diagonalized. Using the resulting eigenvectors one calculates the charges of the atoms $\Delta q_A^{(1)}$ by the Mulliken analysis. This new charges are then used to build up a new Hamiltonian $H_{\mu\nu}^{(1)}$ which will be diagonalized again yielding new eigenvectors and corresponding new atomic charges. The procedure is repeated until self-consistency is reached, so that charges resulting from the eigenvectors of subsequent Hamiltonians do not differ significantly any more.

The coupling term between the net charges on the atoms

$$\gamma_{AB} = \frac{1}{R_{AB}} - s(R_{AB}, U_{t(A)}, U_{t(B)})$$

is composed from the long range Coulomb term $\frac{1}{R_{AB}}$ and a short range term $s(R_{AB}, U_{t(A)}, U_{t(B)})$. Latter incorporates the exchange-correlation effects and ensures the correct limit of γ_{AB} when the distance between the atoms R_{AB} goes to zero. Apart of the distance it also depends on the chemical hardness of the isolated neutral atoms, which can be calculated as the derivative of the energy of the highest occupied orbital ϵ^{hoo} with respect of its occupation f^{hoo} for the given atom:

$$U = \frac{\partial \epsilon^{\text{hoo}}}{\partial f^{\text{hoo}}}.$$

2.5 Further extensions

In order to yield more accurate results and be able to describe a wider range of physical phenomena, the SCC-DFTB scheme, as outlined above, has been extended in various ways in recent years. Those extensions embrace among others the following ones:

- Calculation of magnetic systems with colinear⁸ and non-colinear spin including the effect of spin-orbit coupling.⁹
- Using LDA+U techniques for better description of strongly correlated systems.¹⁰
- Calculating excitations in molecules in the linear response approximation.¹¹
- Calculating electron transport phenomena using non-equilibrium Greens function technique.¹²
- Expanding the total energy up to the third order in the density fluctuation to describe charged systems more accurately.¹³

Detailed descriptions of the theory behind these extensions can be found in the indicated references. All these features have been implemented in the DFTB+ code¹⁴ which is available free of charge¹⁵ for academical, educational and non-profit research use.

3 Example: Bulk Amorphous Oxides

3.1 Computational Details

The structure formation in stoichiometric amorphous TiO₂ thin films has been studied by molecular dynamics (MD) simulations applying the self-consistent-charge density-functional-based tight-binding scheme (DFTB).^{5,6,16} This quantum method provides a good compromise between computational efficiency and chemical accuracy in a wide range of applications. Successful applications include studies on diamond nucleation in amorphous carbon systems,¹⁷ or the discussion of the properties of exo-fluorinated carbon nanotubes.¹⁸ The recently developed `tiorg` set of diatomic Ti-X (X=Ti,H,C,N,O,S) DFTB Slater-Koster integral tables,¹⁹ together with the `mio` set for light elements and their atomic pairs,^{5,6} has been employed (see also www.dftb.org). The `tiorg` set has been shown¹⁹ to provide a reliable description of geometrical, energetic and electronic properties of all titania bulk phases and their low-index surfaces.¹⁹ The DFTB calculations have been performed by using the open source DFTB+ software (version 1.1).^{14,15}

Initial structures for the MD simulations have been prepared containing 216 atoms, spatially and chemically randomly distributed. The atoms are placed in a fixed-volume cubic super-cell arrangement of varying size corresponding to the microscopic mass densities to be studied and the given ideal 1:2 stoichiometry. The model structures with densities of 3.50, 3.80, 4.00, 4.20 and 4.50 g/cm³, respectively have been prepared by using MD simulated annealing (MD-SA). For all models a dynamical quenching path has been followed for relaxation starting from a partly equilibrated liquid state of the model structures at 5000 K progressing a path of exponentially decreasing temperature towards room temperature (300 K). The Newton's equations of motion were solved using the Verlet algorithm²⁰ with

time step length of 1 fs, coupling the MD to a heat bath within the canonical (NVT)-ensemble by using the Anderson thermostat. The total duration of the cooling procedure was 23 ps.

The DFTB method has been validated by performing *ab initio* DFT Car-Parrinello molecular dynamics (CPMD) simulations²¹ under similar conditions. The CPMD simulations have been performed on the basis of norm-conserving pseudo-potentials of the Troullier-Martins type^{22,23} and the Becke88 exchange functional.²⁴ Due to the much higher computational demand the simulation time was shortened to 8.6 ps, discretized in 51000 steps using time steps of 7 atomic time units, to ensure convergence of energy. Here the Nose-Hoover thermostat ensures the conditions of a canonical ensemble, while the volume was kept constant. To address possible larger-scale modeling we validated the classical many-body potential proposed by Matsui and Akaogi (MA)²⁵ against the CPMD and DFTB derived models. The classical potential MD simulations are following again a 23 ps annealing path (identical to the DFTB simulation), using also the Anderson thermostat. In all different method applications the same initial structure has been used. The room temperature CPMD and DFTB structures in Ref. 26 have been subjected to further conjugate gradient relaxation by using the Vienna Ab Initio Simulations Program (VASP) to obtain final zero temperature models.²⁷ An energy cutoff of 400 eV was used and the 3s and 3p semicore states of Ti have been treated as valence states within the PAW potentials throughout this work. The atomic positions have been relaxed using the PBE83 (Perdew-Burke-Ernzerhof) functional, a (2x2x2) Monkhorst-Pack k-point sampling, and a force convergence criterion of 0.01 eV/Å. Further analyzing these structures the electronic structure, band gap and defect localization, as well as frequency-dependent optical data have been calculated.

3.2 Structural and electronic properties of amorphous TiO₂ super-cell models

3.2.1 Structural properties

By using the MD super-cell annealing simulations we have obtained metastable stoichiometric amorphous titanium dioxide models at different mass densities. The amorphous bonding network at mass density of 4.2 g/cm³ are shown in selected images in Figure 1 (upper part). For better visualization of the characteristic TiO_x building units, a polyhedral representation is also given in Figure 1 (lower part). The nanoscale atomistic structure of relaxed TiO₂ models can be characterized by analyzing the mean interatomic distances and coordination numbers, extracted from the radial distribution functions (RDF's). The RDFs and the corresponding structure factors are depicted in Figures 2 and 3, compared with experimentally derived data for sputtered thin films and bulk-powder samples.²⁸ Averaged structural and important electronic information is summarized in Table 1. The models obtained by using the CPMD method and the classical MA potential will be denoted by CP and MA, the models obtained by using the CPMD method and the classical MA-potential by CP and MA, while the models obtained by the DFTB method are indicated by roman numerals (I-V), respectively.

As short-range order fingerprints of the amorphous structure, we list in Table 1 the coordination numbers $k_{\text{Ti-Ti}}$, $k_{\text{Ti-O}}$, $k_{\text{O-Ti}}$, $k_{\text{O-O}}$, which represent average numbers of the nearest neighbors for the corresponding elements in their first coordination shell. Those numbers have been derived from element-specific length statistics, as well as from the

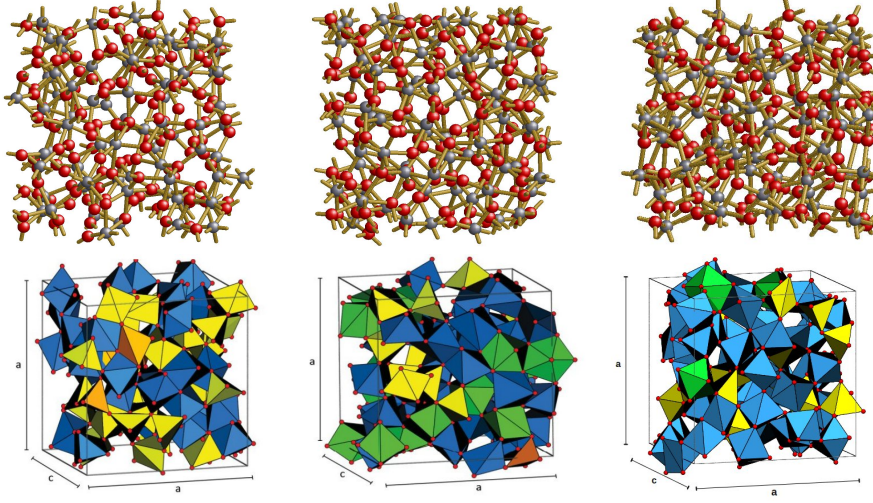


Figure 1. Structural snapshots of amorphous TiO_2 networks at mass density 4.20 g/cm^3 (IV, CP, MA from left to right) – upper part, and in an representation showing TiO_x -building blocks – lower part.

partial radial distribution functions. Additionally, the mean pair distances $R_1^{\text{Ti-Ti}}$, $R_1^{\text{Ti-O}}$, $R_1^{\text{O-O}}$ are given.

Table 1. Structural information of TiO_2 models

Model	$k_{\text{Ti-Ti}}$	$k_{\text{Ti-O}}$	$k_{\text{O-Ti}}$	$k_{\text{O-O}}$	$R_1^{\text{Ti-Ti}}$ (Å)	$R_1^{\text{Ti-O}}$ (Å)	$R_1^{\text{O-O}}$ (Å)	E_{gap} [eV]
a- TiO_2 -(I)	6.14	4.44	2.22	9.88	3.38	1.87	2.91	3.14
a- TiO_2 -(II)	7.00	4.87	2.44	9.81	3.37	1.911	2.84	2.30
a- TiO_2 -(III)	7.75	5.10	2.55	10.50	3.38	1.931	2.83	2.93
a- TiO_2 -(IV)	8.69	5.36	2.68	10.76	3.385	1.94	2.8	2.68
a- TiO_2 -(CP)	8.50	5.83	2.92	10.14	3.38	1.993	2.78	2.74
a- TiO_2 -(MA)	8.97	5.81	2.90	10.80	3.41	1.989	2.80	2.70
a- TiO_2 -(V)	8.53	5.79	2.90	10.42	3.32	1.96	2.72	2.12

The detailed atomistic structure in the amorphous TiO_2 models consists of short staggered chains of TiO_6 octahedrons, like in the crystalline modifications anatase, rutile and brookite. According to the variation of mass density, more or less large number of coordination defects (TiO_5 , TiO_4 units) are identified, which can be found on titanium dioxide surfaces²⁹ or in more complex substoichiometric magneli phases.^{30–33} An increase in the number of TiO_6 octahedral units is clearly mirrored in the Ti-O respectively O-Ti coordination numbers $k_{\text{Ti-O}}$ and $k_{\text{O-Ti}}$, which tend towards the "ideal" crystalline values of $k_{\text{Ti-O}}=6$ and $k_{\text{O-Ti}}=3$. Obviously, those ideal mean values are not reached in the amorphous phase. Coordination numbers determined by experiments as well as reverse Monte Carlo

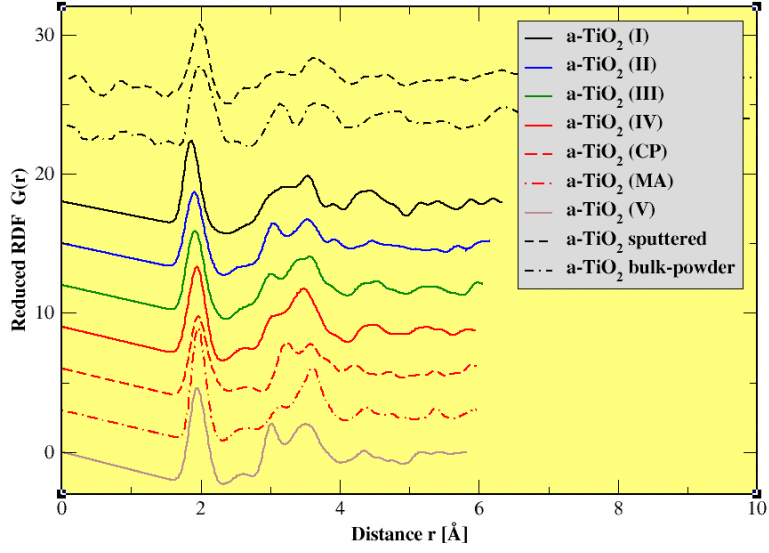


Figure 2. Reduced pair distribution function $G(r)$ of the amorphous TiO_2 models.

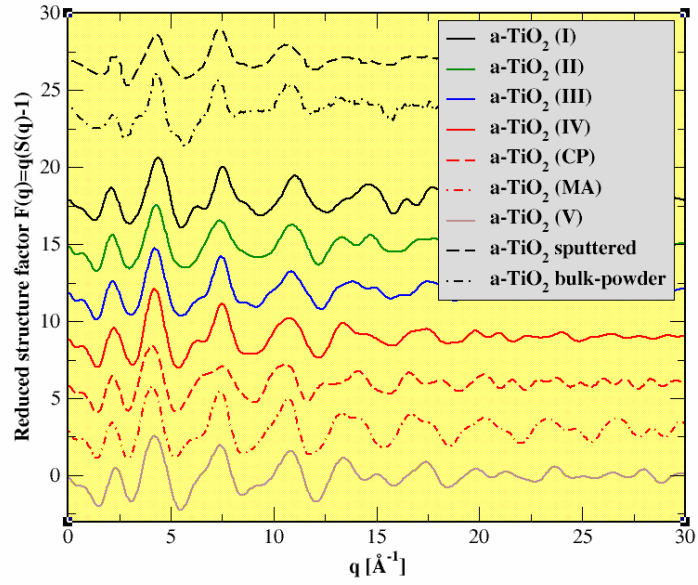


Figure 3. Reduced structure factors $F(q) = q(S(q) - 1)$ for a-TiO_2 models, comparison to experiments.

simulations (RMC)²⁸ show that amorphous TiO₂ systems always stay below the ideal values. For sputtered and sol-gel processed TiO₂ layers Petkov et al.²⁸ report coordination numbers of $k_{\text{Ti-Ti}}=8.8$, $k_{\text{Ti-O}}=5.4$, $k_{\text{O-Ti}}=2.7$, $k_{\text{O-O}}=10.5$ and $k_{\text{Ti-Ti}}=6.5$, $k_{\text{Ti-O}}=4.5$, $k_{\text{O-Ti}}=2.25$, $k_{\text{O-O}}=12.5$, respectively, whereas for bulk-like TiO₂ powders they obtain $k_{\text{Ti-Ti}}=8.7$, $k_{\text{Ti-O}}=5.6$, $k_{\text{O-Ti}}=2.8$, and $k_{\text{O-O}}=10.0$. In analyzing the structure of 4.20 g/cm³, which is close to the rutile density, we see that the DFTB model (IV) nearly perfectly matches the numbers given for sputtered amorphous layers, while the values from CP and MA models are closer to the bulk-like powder samples. Particularly in the case of the classical MA potential, the slightly higher coordination numbers may be understood as resulting from the fitting constraints to reproduce correctly the six-fold coordinated crystalline modifications.³⁴ Contrary, the quantum mechanics approaches cover the under-coordination chemistry more flexibly.

Considering the average values of the pair distances R_1 , and their deviations as characteristic measures for the first neighbor coordination shells, the mean Ti-O bond length for all densities and methods lies approximately between 1.8 - 2.1 Å, which is also found for the crystalline modifications. Experimentally smaller values (1.79-1.93 Å) are seen for layers produced in the sol-gel process. A value of 1.96 Å is given for layers grown by a sputtering process, which was also confirmed by recent RMC-modeling.²⁸ Our values for densities near that of crystalline modifications (see model (IV) at 4.20 g/cm³) again tend more to the numbers of sputtered amorphous titanium-dioxide materials. Here, we monitor a twinned chain of octahedral TiO₆ units percolating through the super-cell and flanked by less coordinated TiO_{4,5} polyhedra, as building blocks of possible under-coordination defects. The averaged bond lengths of 1.98-1.99 Å for the CPMD and MA models at the same density are slightly larger. This is due to their increased coordination number of 5.8, caused by the slightly higher content of ideal octahedron building blocks. As shown in Figures 2 and 3, all trends discussed above are qualitatively reflected in the calculated reduced structure factors $F(q) = q[S(q) - 1]$ and their Fourier transforms, the reduced atomic pair-distribution functions $G(r)$ (see equation (2) in Ref. 26). Comparing those to available experimental data²⁸ on structure factors and RDFs, the rutile-equivalent mass density of 4.20 g/cm³ matches the reported data best and has therefore been chosen for further analysis. Choosing a density same as for one of the crystalline phases also allows the separation of disorder effects from effects of density variations in the electronic structure data.

3.2.2 Electronic properties

One important pronounced feature of all amorphous TiO₂ models is that no electronic defect levels in the band gap appear which could be related to defect Ti-Ti or O-O bonds. All participating elements are "bridged" by their counterpart atomic species. This causes well-defined electronic HOMO-LUMO gap width comparable to the band gap values of crystalline TiO₂. The width of the electronic band gaps, obtained from the DFTB method, is given for all models in Table 1. The total electronic density of states (EDOS) for each model is plotted in Figure 4, showing in Figure 5 a linear decrease with increasing mass density. It is worth noting that the band gap of the three models with the same density of 4.20 g/cm³ but generated by the three different methods (IV, CP, MA), ideally match.

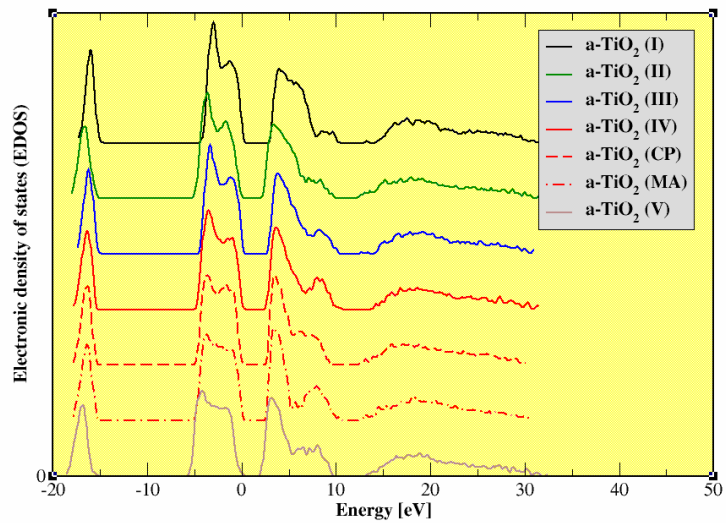


Figure 4. Electronic density of states of a- TiO_2 models. For each case the Fermi energy is shifted to 0 eV.

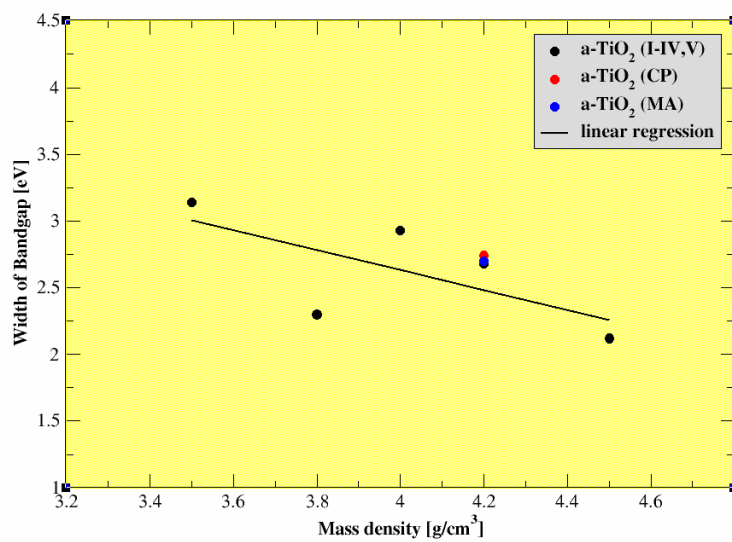


Figure 5. Electronic HOMO-LUMO gap of amorphous TiO_2 models.

4 Summary

We have given an overview about the density functional tight binding method. We have demonstrated, that the DFTB method is a very efficient quantum mechanical simulation tool, often having similar accuracy to *ab initio* DFT calculations while easily outperforming them in time and memory requirements. We have described some results on the investigation of titanium oxide bulk properties as a selected application to demonstrate the capabilities of the DFTB method and the DFTB+ code.

References

1. P. Hohenberg and W. Kohn, Phys. Rev., **136**, 864–871, 1964.
2. W. Kohn and L. J. Sham, Phys. Rev., **140**, A1133–A1138, 1965.
3. W. M. C. Foulkes and R. Haydock, Phys. Rev. B, **39**, 12520–12536, 1989.
4. Z. Bodrog, B. Aradi, and T. Frauenheim, J. Chem. Theory Comput., **7**, 2654–2664, 2011.
5. D. Porezag, T. Frauenheim, T. Köhler, G. Seifert, and R. Kaschner, Phys. Rev. B, **51**, 12947–12957, 1995.
6. M. Elstner, D. Porezag, G. Jungnickel, J. Elsner, M. Haugk, T. Frauenheim, S. Suhai, and G. Seifert, Phys. Rev. B, **58**, 7260–7268, 1998.
7. R. S. Mulliken, J. Chem. Phys., **23**, 1833–1840, 1955.
8. C. Köhler, G. Seifert, and T. Frauenheim, Chem. Phys., **309**, 23, 2005.
9. B. Hourahine, B. Aradi, and T. Frauenheim, J. Phys.: Conf. Ser., **242**, 012005, 2010.
10. B. Hourahine, S. Sanna, B. Aradi, C. Köhler, T. Niehaus, and T. Frauenheim, J. Phys. Chem. A, **111**, 5671–5677, 2007.
11. T. Niehaus, THEOCHEM, **914**, 38–49, 2009.
12. A. Pecchia, L. Salvucci, G. Penazzi, and A. Di Carlo, New Journal of Physics, **10**, 065022, 2008.
13. M. Gaus, Q. Cui, and M. Elstner, J. Chem. Theory Comput., **7**, 931–948, 2011.
14. B. Aradi, B. Hourahine, and T. Frauenheim, J. Phys. Chem. A, **111**, 5678–5684, 2007.
15. The DFTB+ program can be downloaded from <http://www.dftb-plus.info>.
16. M. Elstner, T. Frauenheim, and S. Suhai, J. Mol. Struct. (Theochem), **632**, 2003.
17. Y. Lifshitz, T. Köhler, T. Frauenheim, I. Guzmán, A. Hoffman, R. Q. Zhang, X. T. Zhou, and S. T. Lee, Science, **297**, 1531–1533, 2002.
18. G. Seifert, T. Köhler, and T. Frauenheim, Appl. Phys. Lett., **77**, 1313–1315, 2000.
19. G. Dolgonos, B. Aradi, N. H. Moreira, and T. Frauenheim, J. Chem. Theory Comput., **6**, 266–278, 2010.
20. L. Verlet, Phys. Rev., **159**, 98–103, 1967.
21. R. Car and M. Parrinello, Phys. Rev. Lett., **55**, 2471–2474, 1985.
22. K. M. Glassford, N. Troullier, J. L. Martins, and J. R. Chelikowsky, Solid State Commun., **76**, 635–638, 1990.
23. W. Langel, Surf. Sci., **496**, 141–150, 2002.
24. A. D. Becke, Phys. Rev. A, **38**, 3098–3100, 1988.
25. M. Matsui and M. Akaogi, Mol. Simul., **6**, 239–244, 1991.

26. M. Landmann, T. Köhler, S. Köppen, E. Rauls, T. Frauenheim, and W. G. Schmidt, *Phys. Rev. B*, **86**, 064201–1–20, 2012.
27. G. Kresse and J. Furthmüller, *Comput. Mater. Sci.*, **6**, 15, 1996.
28. V. Petkov, G. Holzrüter, U. Tröge, T. Gerber, and B. Himmel, *J. Non-Cryst. Solids*, **231**, 17–30, 1998.
29. U. Diebold, *Surf. Sci. Rep.*, **48**, 53–229, 2003.
30. S. Anderson, *Acta Chemica Scandinavica*, **14**, 1161–1172, 1960.
31. L. Liborio and N. Harrison, *Phys. Rev. B*, **77**, 104104–1–10, 2008.
32. N. Harrison L. Liborio, G. Mallia, *Phys. Rev. B*, **79**, 245133–1–8, 2009.
33. S. Schreiber, M. Fidler, O. Dulub, M. Schmid, U. Diebold, W. Hou, and A. Selloni U. Aschauer, *Phys. Rev. Lett.*, **109**, 136103, 2012.
34. V. V. Hoang, *Phys. Status Solidi B*, **244**, 1280–1287, 2007.

Wavelets For Electronic Structure Calculations

Thierry Deutsch and Luigi Genovese

Laboratoire de Simulation Atomistique (L.Sim), SP2M/INAC/CEA
17 Av. des Martyrs, 38054 Grenoble, France
E-mail: {*Thierry.D Deutsch, Luigi.Genovese*}@cea.fr

In 2005, the EU FP6-STREP-NEST BigDFT project funded a consortium of four laboratories, with the aim of developing a novel approach for Density Functional Theory (DFT) calculations based on Daubechies wavelets. Rather than simply building a DFT code from scratch, the objective of this three-years project was to test the potential benefit of a new formalism in the context of electronic structure calculations. Daubechies wavelets exhibit a set of properties which make them ideal for a precise and optimized DFT approach. In particular, their systematicity allows to provide a reliable basis set for high-precision results, whereas their locality (both in real and reciprocal space) is highly desired to improve the efficiency and the flexibility of the treatment. In this contribution we will provide a bird's-eye view on the computational methods in DFT, and we then focus on DFT approaches and on the way they are implemented in the BigDFT code, to explain how we can take benefit from the peculiarities of such basis set in the context of electronic structure calculations.

1 Introduction

In the recent years, the development of efficient and reliable methods for studying matter at atomistic level has become an asset for important advancements in the context of material science. Both modern technological evolution and the need for new conception of materials and nanoscaled devices require a deep understanding of the properties of systems of many atoms from a fundamental viewpoint. To this aim, the support of computer simulation can be of great importance. Indeed, via computer simulation scientists try to model systems with many degrees of freedom by giving a set of “rules” of general validity (under some assumptions).

Once these “rules” come from first-principles laws, these simulation have the ambition to model system properties from a fundamental viewpoint. With such a tool, the properties of existing materials can be studied in deep, and new materials and molecules can be conceived, with potentially enormous scientific and technological impact. In this context, the advent of modern supercomputers represent an important resource in view of advancements in this field. In other terms, the physical properties which can be analysed via such methods are tightly connected to the computational power which can be exploited for calculation. A high-performance computing electronic structure program will make the analysis of more complex systems and environments possible, thus opening a path towards new discoveries. It is thus important to provide reliable solutions to benefit from the enhancements of computational power in order to use these tools in more challenging systems.

2 Atomistic Simulations

As an overview, before focusing on more detailed descriptions, we will start this contribution by a brief presentation of the Kohn-Sham formalism of Density Functional Theory. A

number of good references which treat this topic exists. Here we will present some notes, with the aim of defining suitably the problem and fixing notations.

2.1 Born-Oppenheimer Hamiltonian

There is of course no question that a fundamental treatment of a system with many atoms should be performed via the laws of Quantum Mechanics. The properties of the systems are thus governed by its wavefunction, which is related to the Hamiltonian via the Schrödinger equation. It is evident that an immediate solution to this problem does not exist. For a system with N atoms and n electrons, the wavefunction has $3(N + n)$ variables, and the Hamiltonian, in atomic units, has the following form:

$$H = -\frac{1}{2} \sum_{i=1}^n \nabla_{r_i}^2 + \frac{1}{2} \sum_{i \neq j} \frac{1}{|r_i - r_j|} + \\ - \sum_{a=1}^N \sum_{i=1}^n \frac{Z_a}{|R_a - r_i|} + \\ + \sum_{a=1}^N -\frac{1}{2M_a} \nabla_{R_a}^2 + \frac{1}{2} \sum_{a \neq b} \frac{Z_a Z_b}{|R_a - R_b|} . \quad (1)$$

In this equation the Hamiltonian of the electrons (first two terms) is coupled with the one of the ions (last two terms) via the electromagnetic interaction (central term). In atomic units, the action is measured in units of \hbar , the mass in units of the electron mass m_e and the charge in units of the electronic charge $|e|$. For these reasons, the kinetic term which is associated to the nuclei is suppressed by the mass of the ions M_a , which is at least two thousands times heavier than the electrons. It appears thus more than justified to decouple the dynamics of the ions to the one of the electrons. In other terms, the Hamiltonian can be split in two parts:

$$H = -\frac{1}{2} \sum_{i=1}^n \nabla_{r_i}^2 + \frac{1}{2} \sum_{i \neq j} \frac{1}{|r_i - r_j|} + V_{ext}(\{r\}, \{R\}) + H_{ions}[\{R\}] . \quad (2)$$

The Born-Oppenheimer (BO) approximation consists in treating the dynamic of the ions classically. The wavefunction of the system will thus become associated only to the electrons (thus with $3n$ variables), with an external potential $V_{ext}(\{r\}, \{R\})$ which depend of the atomic positions $\{R\}$, which will then appear as external parameters to the quantum problem.

Even though the BO approximation effectively reduces the complexity of the description only to the electronic part, we are still really far from a formalism which is able to treat systems with many electrons. The number of variables of the wavefunction is still much too high to be handled while solving the equation explicitly. One may actually wonder whether we *really* need the complete wavefunction to extract the properties of the system we are interested to. For example, the energy of the system in a given quantum state $|\Psi\rangle$ is

$$E[\Psi] = \langle \Psi | H | \Psi \rangle , \quad (3)$$

which can be interpreted as a functional of the wavefunction $|\Psi\rangle$. A closer inspection reveals that the wavefunction contains too much information for calculating the energy.

Since the Hamiltonian contains two-body operators (the electron-electron interaction), it is easy to show that actually the energy is a functional of the 2-particle reduced density matrix (2-RDM) γ_2 :

$$E = \text{tr}(H\gamma_2) = E[\gamma_2], \quad (4)$$

where

$$\gamma_2(x_1, x_2; x'_1, x'_2) = \binom{n}{2} \int dx_3 \cdots dx_N \Psi(x_1, \dots, x_N) \Psi^*(x'_1, x'_2, x_3, \dots, x_N), \quad (5)$$

is a function of 12 variables. The formulation of the problem seems thus simpler in this way, but the 2-RDM cannot be a generic function. It must be chosen such that it comes from the contraction of a wavefunction as indicated in Eq. (5). Taking into account such a constraint (the so-called n -representability problem) is definitely a far-from-trivial task, and still keeps the formalism difficult to handle.

A big simplification to the problem of finding the ground-state energy of the system have been provided by Hohenberg and Kohn in 1964, via their famous theorem (HK):

Hohenberg – Kohn Theorem. *For a fixed number of electrons n , the charge density of the ground-state of a quantum system determines uniquely – up to an additive constant – the external potential of the electronic Hamiltonian.*

If we take into account that, of course, given *both* n and an external potential, the charge density of the ground state is determined, the HK theorem states that there is a one-to-one correspondence between the charge density, a functional of the 2-RDM^a $\rho(\mathbf{r}) = \frac{2}{n-1} \int d\mathbf{r}_1 \gamma_2(\mathbf{r}, \mathbf{r}_1; \mathbf{r}, \mathbf{r}_1) = \rho[\gamma_2]$ and the external potential which determines the inhomogeneity of the electron gas. This implies that the ground state energy E_0 is a functional of the electronic density ρ . Such a functional reaches its minimum for the true ground state density ρ_0 :

$$E = E[\rho] = \min_{\gamma_2 \text{ s.t. } \rho[\gamma_2] = \rho} \left\{ \text{tr} \left(\left[-\frac{1}{2} \sum_{i=1}^n \nabla_{r_i}^2 + \frac{1}{2} \sum_{i \neq j} \frac{1}{|r_i - r_j|} \right] \gamma_2 \right) \right\} + \int d\mathbf{r} \rho(\mathbf{r}) V_{ext}(\{\mathbf{r}\}, \{R\}), \quad (6)$$

and $E[\rho_0] = E_0$, which is at the basis of the *Density Functional Theory*. We have assumed here that the system has n electrons, i.e. $\int d\mathbf{r} \rho(\mathbf{r}) = n$. Via Eq. (6), we can see that in the functional of the density there is a term which does not depends explicitly of the external potential, which for this reason can be considered as a *universal* functional:

$$\begin{aligned} F[\rho] &= \min_{\gamma_2 \text{ s.t. } \rho[\gamma_2] = \rho} \left\{ \text{tr} \left(\left[-\frac{1}{2} \sum_{i=1}^n \nabla_{r_i}^2 + \frac{1}{2} \sum_{i \neq j} \frac{1}{|r_i - r_j|} \right] \gamma_2 \right) \right\} = \\ &= \min_{\Psi \text{ s.t. } \rho[\Psi] = \rho} \left\{ \left\langle \Psi \left| -\frac{1}{2} \sum_{i=1}^n \nabla_{r_i}^2 + \frac{1}{2} \sum_{i \neq j} \frac{1}{|r_i - r_j|} \right| \Psi \right\rangle \right\}. \quad (7) \end{aligned}$$

^aIn most of the formulations the charge density is seen as a functional of the wavefunction (via e.g. the Levy's constrained search formulation). This allows to bypass the n -representability problem of the 2-RDM. Here we prefer to use this formulation to show that the 2-RDM problem actually *contains* the HK formulation. Indeed, it is easy to see that the minimum of the energy for *all* n -representable γ_2 satisfies the constraint $\rho[\gamma_2] = \rho_0$.

For densities of systems with n electrons, the quantity $E[\rho] = F[\rho] + \int \rho V_{ext}$ reaches its minimum for the ground-state density ρ_0 . It is important to stress that all quantities depend of n , which is supposed fixed.

The demonstration of the HK theorem is independent of the form of the pure electronic Hamiltonian. For an n electron system which has no Coulombic interaction the HK functional (let us call it $T_s[\rho]$) has a pure kinetic term:

$$T_s[\rho] = \min_{\gamma_2: \rho[\gamma_2] = \rho} \left\{ \text{tr} \left(\left[-\frac{1}{2} \sum_{i=1}^n \nabla_{r_i}^2 \right] \gamma_2 \right) \right\} = \min_{\Psi: \rho[\Psi] = \rho} \left\{ \left\langle \Psi \left| -\frac{1}{2} \sum_{i=1}^n \nabla_{r_i}^2 \right| \Psi \right\rangle \right\}. \quad (8)$$

Moreover, the pure Coulombic energy of a system with density ρ is known, and can be seen as (half) the potential energy where the potential is the Hartree potential

$$V_H[\rho](\mathbf{r}) = \int d\mathbf{r}' \frac{\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}:$$

$$E_H[\rho] = \frac{1}{2} \int d\mathbf{r} d\mathbf{r}' \frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} = \frac{1}{2} \int d\mathbf{r} \rho(\mathbf{r}) V_H[\rho](\mathbf{r}). \quad (9)$$

Given these quantities, we can define the Exchange and Correlation functional $E_{xc}[\rho]$, and the associated Exchange and Correlation density *per particle* $\epsilon_{xc}(\mathbf{r})$

$$E_{xc}[\rho] = \int d\mathbf{r} \rho(\mathbf{r}) \epsilon_{xc}[\rho](\mathbf{r}) = F[\rho] - T_s[\rho] - E_H[\rho] \quad (10)$$

The quantity $E[\rho] = T_s[\rho] + E_H[\rho] + E_{xc}[\rho] + \int \rho V_{ext}$ should then be minimal in ρ_0 for all densities which sum up to n . This implies that $E[\rho_0 + \delta\rho] = E[\rho_0]$ for $\int d\mathbf{r} \delta\rho(\mathbf{r}) = 0$. Hence

$$\begin{aligned} 0 &= \int d\mathbf{r} \delta\rho(\mathbf{r}) \frac{\delta E}{\delta\rho(\mathbf{r})} = \\ &= \int d\mathbf{r} \delta\rho(\mathbf{r}) \left\{ \frac{\delta T_s[\rho]}{\delta\rho(\mathbf{r})} + V_H[\rho](\mathbf{r}) + \frac{d}{d\rho} (\rho \epsilon_{xc}[\rho]) (\mathbf{r}) + V_{ext}(\mathbf{r}) \right\}. \end{aligned} \quad (11)$$

It is easy to see that the above equation is the same that one would obtain by searching the ground state of the non-interacting Hamiltonian H_{KS} (so-called Kohn-Sham Hamiltonian):

$$H_{KS}[\rho] = -\frac{1}{2} \sum_{i=1}^n \nabla_{\mathbf{r}_i}^2 + V_H[\rho] + V_{xc}[\rho] + V_{ext}, \quad (12)$$

where we have defined the Exchange and Correlation potential

$$V_{xc}[\rho](\mathbf{r}) = \frac{d}{d\rho} (\rho \epsilon_{xc}[\rho]) (\mathbf{r}) = \frac{\delta}{\delta\rho(\mathbf{r})} E_{xc}[\rho]. \quad (13)$$

Since the Hamiltonian is only made of one-body operators, the energy of such a system can be expressed via the eigenfunctions of H_{KS} and via the one particle reduced density matrix (1-RDM) derived from them:

$$H_{KS}[\rho] |\psi_p\rangle = \varepsilon_p^{KS} |\psi_p\rangle, \quad p \in \mathbb{N} \quad (14)$$

so that the 1-RDM of this system is

$$\gamma_1^{KS} = \sum_p f_p |\psi_p\rangle \langle \psi_p|, \quad (15)$$

where the occupation numbers $0 \leq f_p \leq 1$, $\sum_p f_p = n$ guarantee the n -representability of γ_1^{KS} . The energy of the original system is thus:

$$E[\rho] = \text{tr} (H_{KS}[\rho] \gamma_1^{KS}) - \frac{1}{2} E_H[\rho] + \int d\mathbf{r} \rho(r) (\epsilon_{xc}[\rho](\mathbf{r}) - V_{xc}[\rho](\mathbf{r})) , \quad (16)$$

and, of course, $\rho(\mathbf{r}) = \gamma_1^{KS}(\mathbf{r}; \mathbf{r})$.

We have followed the main steps of the demonstration of the

Kohn – Sham Theorem. *An electronic density which is associated to the ground state of an interacting electron system is also solution of a non-interacting problem submitted to a mean-field potential $V_{xc} + V_H + V_{ext}$.*

The consequences of this theorem are potentially important. If the quantity $\epsilon_{xc}[\rho]$ is known, the energy of the system can be found iteratively: for a given ρ , the eigenvalue problem of the KS Hamiltonian would provide a set of eigenfunctions $|\psi_p\rangle$, and then a new electronic density. Convergence is reached for the density ρ_0 , which minimizes $E[\rho]$. Even though ρ_0 comes from a non-interacting electron system, ρ_0 is the *exact* charge density of the interacting system.

2.2 LDA and GGA exchange correlation approximations

Clearly, the difficulty now resides in finding the correct $\epsilon_{xc}[\rho]$. Surprisingly, if we take as the XC density per electron from a homogeneous electron gas of density n , results of rather good quality can already be obtained. In this way $\epsilon_{xc}[\rho](\mathbf{r}) = \epsilon_{xc}^{\text{hom}}(\rho(\mathbf{r}))$. This is the Local Density Approximation (LDA), which can be parametrized from numerical results. Several other approximations exist, which give good results for the extraction of several properties of real materials.

A particularly used set of XC functionals is implemented as a functional of the density and of its gradient (its modulus for rotational invariance). This is the so-called Generalized Gradient Approximation (GGA):

$$\epsilon_{xc}(\mathbf{r}) = \epsilon_{xc}(\rho(\mathbf{r}), |\nabla\rho|(\mathbf{r})) . \quad (17)$$

In this case, the exchange correlation potential has an additional term:

$$\begin{aligned} V_{xc}(\mathbf{r}) &= \frac{\delta}{\delta\rho(\mathbf{r})} \int \rho(\mathbf{r}') \epsilon_{xc}(\rho(\mathbf{r}'), |\nabla\rho|(\mathbf{r}')) \\ &= \frac{d}{d\rho} (\rho \epsilon_{xc})(\mathbf{r}) + \int \rho(\mathbf{r}') \frac{\partial \epsilon_{xc}}{\partial |\nabla\rho|}(\mathbf{r}') \frac{\delta}{\delta\rho(\mathbf{r})} |\nabla\rho|(\mathbf{r}') d\mathbf{r}' \\ &= \epsilon_{xc}(\mathbf{r}) + \rho(\mathbf{r}) \frac{\partial \epsilon_{xc}}{\partial \rho}(\mathbf{r}) + \int \frac{\rho}{|\nabla\rho|} \frac{\partial \epsilon_{xc}}{\partial |\nabla\rho|}(\mathbf{r}') \sum_{i=x,y,z} \partial_i \rho(\mathbf{r}') \frac{\delta}{\delta\rho(\mathbf{r})} \partial_i \rho(\mathbf{r}') d\mathbf{r}' . \end{aligned} \quad (18)$$

The different components of the gradient of the density $\partial_i \rho(\mathbf{r})$, $i = 1, 2, 3$ can be seen here as a linear functional of the density. For example, for a finite-difference computation on a grid we have

$$\partial_i \rho(\mathbf{r}) = \sum_{\mathbf{r}'} c_{\mathbf{r}, \mathbf{r}'}^i \rho(\mathbf{r}') , \quad (19)$$

such that $\frac{\delta}{\delta\rho(\mathbf{r})} \partial_i \rho(\mathbf{r}') = \sum_{\mathbf{r}''} c_{\mathbf{r}', \mathbf{r}''}^i \delta(\mathbf{r} - \mathbf{r}'')$. This expression can be used to calculate the last term of Eq. (18).

2.3 Hybrid functionals and exact exchange operator

The Kohn-Sham theorem showed us that there exists an antisymmetric wavefunction $|\Phi_0\rangle$ of an n -electron system which satisfies the following properties:

1. The density originated from $|\Phi_0\rangle$ corresponds *exactly* to the density of the original inhomogeneous electron gas:

$$\langle\Phi_0|\mathbf{r}\rangle\langle\mathbf{r}|\Phi_0\rangle = \rho_0(\mathbf{r}) ; \quad (20)$$

2. The wavefunction is the ground state of the *non interacting* Schrödinger equation:

$$H_{KS}[\rho_0]|\Psi_0\rangle = E_0^{KS}[\rho_0]|\Psi_0\rangle , \quad (21)$$

and $E_0[\rho_0] = E_0^{KS}[\rho_0] - \frac{1}{2}E_H[\rho_0] + E_{xc}[\rho_0] - \int \rho_0 V_{xc}[\rho_0]$ is the ground-state energy of the *interacting* system;

3. The density ρ_0 minimizes the value of E_0 , and it is a fixed point for E_0^{KS} .

This wavefunction can be written in the basis of Slater determinants of the eigenfunctions of the one-body Hamiltonian. In this basis, it is easy to show that $E_0^{KS} = \sum_p f_p \varepsilon_p^{KS}$, where f_p is the occupation number defined above. In this context, it is easy to see that for a system for which the Kohn-Sham energies have a gap between ε_n^{KS} and ε_{n+1}^{KS} , the minimum energy is attained when $f_p = \theta(n - p)$ and thus $|\Phi_0\rangle$ is made of only one Slater determinant. Otherwise, multideterminantal configurations are possible.

In this context it is interesting to calculate the contribution of the non-interacting system to the two-body electron-electron interaction. This has a formal equivalence with the Hartree-Fock exchange operator:

$$E_x^{HF} = -\frac{1}{2} \sum_{\sigma=1,2} \sum_{p,q} f_{p,\sigma} f_{q,\sigma} \int d\mathbf{r} d\mathbf{r}' \frac{\psi_{p,\sigma}(\mathbf{r}) \psi_{q,\sigma}^*(\mathbf{r}) \psi_{p,\sigma}^*(\mathbf{r}') \psi_{q,\sigma}(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} , \quad (22)$$

where the spin quantum number σ of the non-interacting electrons has been explicitated. Of course, the system of Kohn-Sham orbitals would now become interacting. This implies that an operator D_x^{HF} should be added to the Kohn-Sham Hamiltonian. The action of this operator onto a wavefunction can be calculated knowing that E_x^{HF} originates from a trace of such an operator over the KS wavefunctions $|\psi_p\rangle$:

$$E_x^{HF} = \sum_{p,\sigma} f_{p,\sigma} \langle \psi_{p,\sigma} | D_x^{HF} | \psi_{p,\sigma} \rangle ;$$

$$\langle \mathbf{r} | D_x^{HF} | \psi_{p,\sigma} \rangle = \frac{1}{f_{p,\sigma}} \frac{\delta E_x^{HF}}{\delta \psi_{p,\sigma}^*(\mathbf{r})} \quad (23)$$

$$= - \sum_q f_{q,\sigma} \int d\mathbf{r}' \frac{\psi_{q,\sigma}^*(\mathbf{r}') \psi_{p,\sigma}(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} \psi_{q,\sigma}(\mathbf{r}) ; \quad (24)$$

As already suggested in the seminal paper of Kohn and Sham, such construction can be used to define an alternative scheme for the Kohn-Sham procedure. By defining a hybrid Kohn-Sham – Hartree-Fock Hamiltonian

$$H_{KSHF}[\rho] = -\frac{1}{2} \sum_{i=1}^n \nabla_{\mathbf{r}_i}^2 + V_H[\rho] + V_{xc}^{KSHF}[\rho] + V_{ext} + \alpha D_x^{HF} \quad (25)$$

and finding its eigenvalues $\varepsilon_p^{\text{KSHF}}$, the energy would become

$$E[\rho] = \sum_p f_p \varepsilon_p^{\text{KSHF}} - \frac{1}{2} E_H[\rho] + \alpha E_x^{\text{HF}} + E_{xc}^{\text{KSHF}}[\rho] - \int \rho V_{xc}^{\text{KSHF}}[\rho],$$

$$E_{xc}^{\text{KSHF}}[\rho] = E_{xc}[\rho] - \alpha E_x^{\text{HF}}, \quad (26)$$

$$V_{xc}^{\text{KSHF}}[\rho] = \frac{\delta E_{xc}^{\text{KSHF}}[\rho]}{\delta \rho}. \quad (27)$$

2.4 Finding the Kohn-Sham wavefunctions: Direct minimization algorithm

We have seen that the electronic density of the system can be constructed via the KS wavefunctions $\psi_p(\mathbf{r})$, which are in turn eigenfunctions of the KS Hamiltonian, which also depends on the density. Thus, a fixed point equation has to be solved. Once the fixed point is reached, the energy of the system can be extracted. The HK theorem guarantees us that the energy $E[\rho]$ is minimal in the ground-state density ρ_0 . The KS construction simplifies things a bit. The problem corresponds to minimize the energy of the KS Hamiltonian as if such Hamiltonian does not evolve. A new Hamiltonian can then be defined. In other terms, in the typical KS procedure the variation is performed over the wavefunctions (supposing that the occupation numbers are integers). The interesting quantity is thus

$$\frac{\delta E[\rho[\{\psi_p\}]]}{\delta \langle \psi_p |} = f_p H_{KS}[\rho] |\psi_p\rangle + \int d\mathbf{r} \frac{\delta \rho(\mathbf{r})}{\delta \langle \psi_p |} \frac{\delta E[\rho]}{\delta \rho(\mathbf{r})}, \quad (28)$$

As already discussed, if $\rho = \sum_p f_p |\psi_p|^2$ the last term of the rhs of this equation is zero. Things goes as if the KS Hamiltonian is fixed. The fixed-point solution ρ_0 thus minimizes both $E[\rho]$ and its KS wavefunctions minimize $E_{KS}[\rho_0]$.

This fact can be derived from the explicit form of KS Hamiltonian:

$$\begin{aligned} \frac{\delta E[\rho]}{\delta \rho(\mathbf{r})} &= \frac{\delta E_{KS}[\rho]}{\delta \rho(\mathbf{r})} - V_H[\rho](\mathbf{r}) + V_{xc}[\rho](\mathbf{r}) - \frac{\delta}{\delta \rho(\mathbf{r})} \int d\mathbf{r}' \rho(\mathbf{r}') V_{xc}[\rho](\mathbf{r}') \\ &= \sum_p f_p \langle \psi_p | \frac{\delta H_{KS}[\rho]}{\delta \rho(\mathbf{r})} | \psi_p \rangle - V_H[\rho](\mathbf{r}) - \rho(\mathbf{r}) \frac{dV_{xc}[\rho]}{d\rho}(\mathbf{r}). \end{aligned}$$

Let us now consider the first term:

$$\begin{aligned} \langle \psi_p | \frac{\delta H_{KS}[\rho]}{\delta \rho(\mathbf{r})} | \psi_p \rangle &= \int d\mathbf{r}' d\mathbf{r}'' \psi_p^*(\mathbf{r}') \psi_p(\mathbf{r}'') \langle \mathbf{r}' | \frac{\delta H_{KS}[\rho]}{\delta \rho(\mathbf{r})} | \mathbf{r}'' \rangle \\ &= \int d\mathbf{r}' d\mathbf{r}'' \psi_p^*(\mathbf{r}') \psi_p(\mathbf{r}'') \left[\langle \mathbf{r}' | \frac{\delta V_H[\rho]}{\delta \rho(\mathbf{r})} | \mathbf{r}'' \rangle + \langle \mathbf{r}' | \frac{\delta V_{xc}[\rho]}{\delta \rho(\mathbf{r})} | \mathbf{r}'' \rangle \right]. \end{aligned} \quad (29)$$

Now the results can be written in term of the Dirac distribution:

$$\langle \mathbf{r}' | \frac{\delta V_H[\rho]}{\delta \rho(\mathbf{r})} | \mathbf{r}'' \rangle = \frac{\delta(\mathbf{r}' - \mathbf{r}'')}{|\mathbf{r}' - \mathbf{r}|}, \quad (30)$$

$$\langle \mathbf{r}' | \frac{\delta V_{xc}[\rho]}{\delta \rho(\mathbf{r})} | \mathbf{r}'' \rangle = \delta(\mathbf{r}' - \mathbf{r}'') \delta(\mathbf{r}' - \mathbf{r}) \frac{dV_{xc}[\rho]}{d\rho}(\mathbf{r}). \quad (31)$$

Hence, since the sum of the squares of the wavefunctions gives the same ρ :

$$\begin{aligned} \sum_p f_p \langle \psi_p | \frac{\delta H_{KS}[\rho]}{\delta \rho(\mathbf{r})} | \psi_p \rangle &= \int d\mathbf{r}' \rho(\mathbf{r}') \left[\frac{1}{|\mathbf{r}' - \mathbf{r}|} + \delta(\mathbf{r}' - \mathbf{r}) \frac{dV_{xc}[\rho]}{d\rho}(\mathbf{r}) \right] \\ &= V_H[\rho](\mathbf{r}) + \rho(\mathbf{r}) \frac{dV_{xc}[\rho]}{d\rho}(\mathbf{r}) , \end{aligned} \quad (32)$$

which implies the KS Lagrangian condition $\frac{\delta E[\rho]}{\delta \rho(\mathbf{r})} = 0$.

While performing the search for the fixed point, the so-called Self Consistent Field (SCF) cycle, the wavefunctions have to be modified between one step and the other, while maintaining orthogonality. The latter can be implemented via a Lagrange multiplier Λ_{pq} , which define the Lagrangian

$$L[\{\psi_p\}] = E[\rho[\{\psi_p\}]] - \sum_{p,q} \Lambda_{pq} (\langle \psi_p | \psi_q \rangle - \delta_{pq}) . \quad (33)$$

Imposing $\frac{\delta L[\{\psi_p\}]}{\delta \langle \psi_p |} = 0$ gives $\Lambda_{pq} = \langle \psi_q | H_{KS}[\rho] | \psi_p \rangle$. Of course, only wavefunctions which are occupied contribute to the energy. The gradient of the KS energy wrt the wavefunction is then

$$|g_p\rangle = H_{KS}[\rho] | \psi_p \rangle - \sum_q \langle \psi_q | H_{KS}[\rho] | \psi_p \rangle | \psi_q \rangle . \quad (34)$$

The vectors $\{|g_p\rangle\}$ provide the direction in which the energy varies the most for a given set of wavefunctions $\{\psi_p\}$. Different algorithms can then be used to find the fixed-point solution. This is the so-called direct minimization algorithm. In Figure 3, the flowchart of the operations is indicated in the case of a plane wave basis set. This flowchart is roughly the same as in the case of other basis sets. The main limitation part for systematic basis sets as the number of atoms increases is the orthonormalization part which scales cubically with the number of atoms if the orbitals are extended over the whole system.

Beware that, since it only involves occupied wavefunctions, such algorithm is correctly defined only if *any* KS Hamiltonian of the SCF cycle exhibits an energy gap, and if the updated wavefunction at each step has components onto all the first n eigenspaces of the new Hamiltonian.

2.5 Finding the Kohn-Sham wavefunctions: Diagonalization of the Hamiltonian

To calculate properly metallic system, the only possibility is to diagonalize the Hamiltonian at each step and populates the Kohn-Sham orbitals in function of the Kohn-Sham eigenvalues.

In Figure 1, the self-consistent equations are shown. At each iteration, the Hamiltonian needs to be diagonalized. We give more details in the plane wave section (see Figure 4) about the different operations. Iterative algorithms are used to diagonalize the Hamiltonian in the case of systematic basis sets because the number of computed orbitals are quite small (by a factor of 100) compared to the number of components. The most used iterative algorithms are conjugate gradient scheme, Davidson,³⁵ Lanczos, RMM-DIIS (Residual Minimum Method – Direct Inversion of the Iterative Subspace used in VASP code³⁶) or LOBPCG methods (Locally Optimal Block Preconditioned Conjugate Gradient³⁷). Except

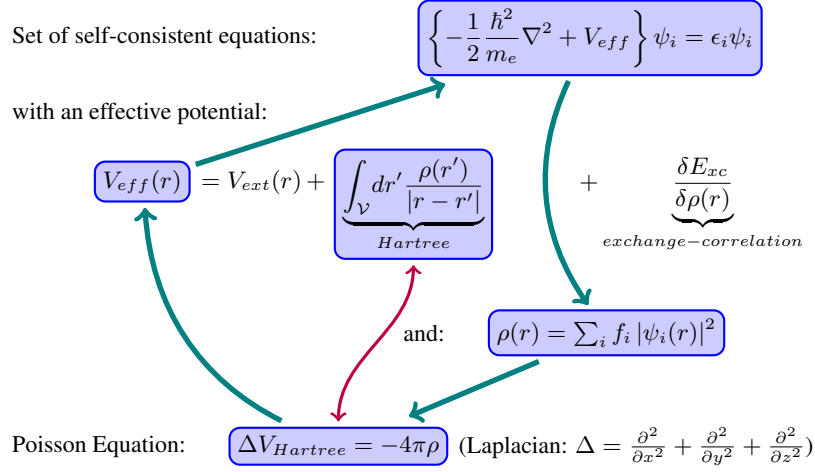


Figure 1. Self-consistent equations used in the diagonalization scheme

the conjugate gradient scheme, this algorithms can be parallelized which is really important to handle systems composed of few hundred of atoms.

3 Pseudopotentials

The KS formalism presents thus a procedure to study the electronic properties of a system with many atoms. However, for such a system the interesting properties are determined by the valence electrons of the atoms involved. Electrons close to the nuclei have a behaviour which can be considered independent of the system under consideration. These electrons contribute to the screening of the atomic charge, but have no significant influence on the behaviour of the peripheric electrons. It may thus appear convenient to consider a system in which *only* the valence electrons appear, where the electron-ion interaction potential is substituted by another object, the *pseudopotential*, (PSP) which mode the effect of the core electron.

From a computational viewpoint, the advantage of using pseudopotential approximation is twofold: on one hand, the overall number of electrons in the system is reduced, which makes lighter the computational treatment. On the other hand, the PSP operator makes the KS wavefunctions close to the position of the nuclei smoother than the ordinary ion-electron potential. This is also important from the implementation viewpoint since a smooth function is always easier to express numerically.

It can be understood easily that the PSP approximation is less severe than the XC approximation. However, the PSP operator should be defined carefully such that several conditions must be respected. Moreover, the influence of the core electrons on the nuclei must be expressed by the insertion of non-local operators, since the screening of the core electrons is different for any of the multipoles of the electron-ion potential.

4 Kohn-Sham DFT with Daubechies Wavelets

In the recent years the KS formalism has been proven to be one of the most efficient and reliable first-principle methods for predicting material properties and processes which undergo a quantum mechanical behavior. The high accuracy of the results together with the relatively simple form of the most common exchange-correlation functionals make this method probably the most powerful tool for *ab-initio* simulations of the properties of matter. The computational machinery of DFT calculations has been widely developed in the last decade, giving rise to a plethora of DFT codes. The usage of DFT calculation has thus become more and more common, and its domain of application comprises solid state physics, chemistry, materials science, biology and geology.

From a computational point of view, one of the most important characteristics of a DFT code is the set of basis functions used for expressing the KS orbitals. The domain of applicability of a code is tightly connected to this choice. For example, a non-localized basis set like plane waves is highly suitable for electronic structure calculations of periodic and/or homogeneous systems like crystals or solids, while it is much less efficient in expanding localized information, which has a wider range of components in the reciprocal space. For these reasons DFT codes based on plane waves are not convenient for simulating inhomogeneous or isolated systems like molecules, due to the high memory requirements for such kind of simulations.

A remarkable difference should be also made between codes which use systematic and non-systematic basis sets. A systematic basis set allows us to calculate the exact solution of the KS equations with arbitrarily high precision as the number of basis functions is increased. In other terms, the numerical precision of the results is related to the number of basis functions used to expand the KS orbitals. With such a basis set it is thus possible to obtain results that are free of errors related to the choice of the basis, eliminating a source of uncertainty. A systematic basis set allows us thus to really calculate the solution of a particular exchange correlation functional. On the other hand, an example of a non-systematic set is provided by Gaussian type basis, for which over-completeness may be achieved before convergence. Such basis sets are more difficult to use, since the basis set must be carefully tuned by hand by the user, which will sometimes require some preliminary knowledge of the system under investigation. This is the most important weakness of this popular basis set.

Another property which has a role in the performances of a DFT code is the orthogonality of the basis set. The use of nonorthogonal basis sets requires the calculation of the overlap matrix of the basis function and performing various operations with this overlap matrix such as inverting the matrix. This makes methods based on non-orthogonal basis functions not only more complicated but also slower.

In Figure 2, we give an overview of the different possibilities to solve the Kohn-Sham equations. The choice of a basis set determines strongly the accuracy of a code and the different operations which need to be computed. The cost of each step in the self-consistent loop is not same and can differ drastically for gaussian or plane wave basis sets.

From the point of view of the developer, some formalisms are easier to program than the other ones. This is the case for plane wave or wavelet basis set in the case, for instance of calculating atomic forces. Another point is the flexibility of the possible boundary conditions (isolated or periodic systems, surfaces or wires). In Section 7, we develop this

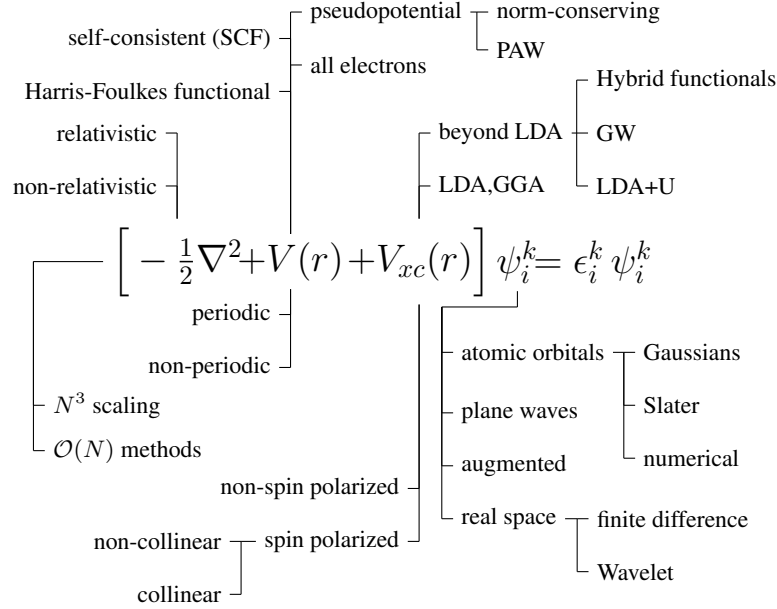


Figure 2. List of options for a DFT code

point applied to the calculation of the Hartree potential *i.e.* the Poisson solver.

We give a short list of codes which is not really exhaustive but give an idea of the diversity of proposed solutions to solve the Kohn-Sham equations:

- Plane Waves

- ABINIT — Louvain-la-Neuve — <http://www.abinit.org>

This code is available under GPL licence and has a strong community of developers and users; The forum discussion are very active and are useful to help the beginners. ABINIT can do electronic structure calculation and calculates many properties based on the linear response as well the many-body perturbation theory (GW method).

- CPMD — Zurich, Lugano — <http://www.cpmd.org>

The code CPMD (Car-Parrinello Molecular Dynamics) is freely distributed and is one the first developed code based on plane waves and massively parallel. It is used to do geometry optimization, molecular dynamics and can be combined with other codes in the multiscale approach of QM/MM (Quantum Mechanics, Molecular Modelling).

- PWSCF — Italy — <http://www.pwscf.org>

The code PWSCF is distributed over the GPL license. It has also a strong community of users and many capabilities specially to calculate electronic properties based on the linear response as ABINIT.

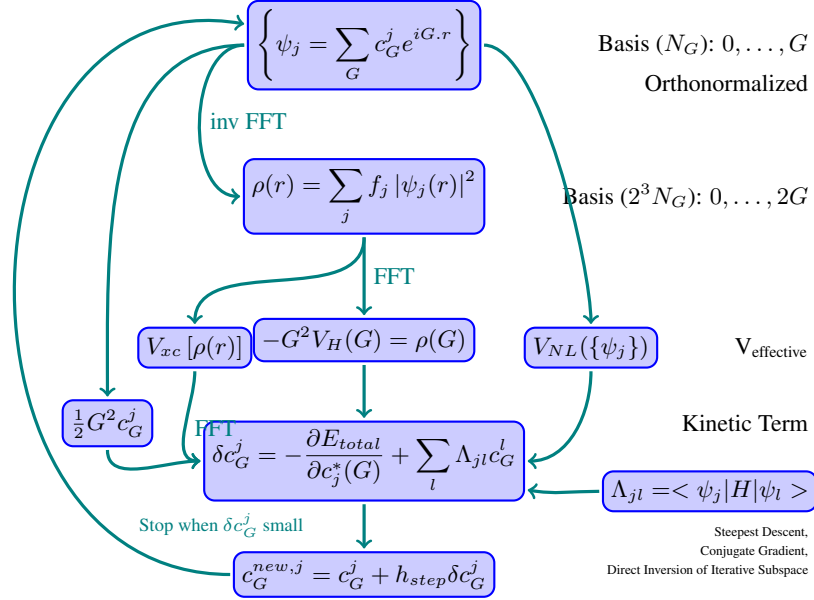


Figure 3. Direct Minimization: Flowchart for a code based on the plane wave basis set

- VASP — Vienna — <http://cms.mpi.univie.ac.at/vasp>
This code has been tuned to be fast and robust. This code is more dedicated to the calculation of structural properties. This code is widely used and has a strong community of users.
- Gaussian
 - CP2K — <http://cp2k.berlios.de>
This code under GPL license combines a Gaussian basis set to describe the wavefunction and plane waves or wavelet to express the electronic density and calculates the Hartree potential.
 - Gaussian — <http://www.gaussian.com>
Gaussian code is a well-known commercial code created by John Pople.
 - DeMon — <http://www.demon-software.com>
DeMon was originally developed in Montreal and is freely available.
- ADF — Amsterdam —
Amsterdam Density Functional code uses Slater orbitals to express the wavefunctions. It is a commercial code with many capabilities.
- Siesta — Madrid — <http://www.uam.es/departamentos/ciencias/fismateriac/siesta>
Siesta uses a numerical basis sets to express the wavefunctions and plane wave to calculate the electronic density and the Hartree potential.

- Wien — Vienna — <http://www.wien2k.at>
This code uses a full-potential linear augmented plane wave (FPLAPW) basis set tuned to represent with few orbitals the wavefunctions in a solid.
- Real space basis set
 - ONETEP — <http://www.onetep.soton.ac.uk>
This code uses sinus cardinal which can represent exactly a plane wave basis set for a given energy cutoff. $O(N)$ approach is already implemented.
 - BigDFT — http://inac.cea.fr/L_Sim/BigDFT
This is the first code based on wavelet using pseudopotential, massively parallel. It is also integrated in the ABINIT package.
 - GPAW — <https://wiki.fysik.dtu.dk/gpaw/>
Under GPL license, GPAW (Grid-bases projector-augmented wave method) uses a finite difference scheme with projected-augmented-wave (PAW) pseudopotentials.

During the last years, developers have tried to share common developments as exchange-correlation library (`libXC`¹³) or input/output libraries (ETSF-IO). The idea is to reuse as much as possible already existing code in order to decrease the cost of development. The main part is the debugging and the maintenance of a code. Using libraries has the advantage to force the modularity of a code and concentrate the effort only to the original part.

Systematic basis sets, such as plane waves or wavelets, have the advantage to permit an easy control over the accuracy of the calculation. We develop first the specificity of plane wave basis sets and then concentrate on wavelet basis sets.

4.1 Plane wave basis sets

Plane waves are widely used as an orthogonal systematic basis set. They are well adapted for periodic systems and based on the Fast Fourier Transform (FFT). The idea from R. Car and M. Parrinello is to express the operators involved in Hamiltonian in the Fourier space for the kinetic operator and in the real space for the local potential. Each time, the operator is diagonal and easy to calculate.

In Figure 3, the flowchart of operations is indicated in the case of the direct minimization. As we mentioned already, the main cost becomes the orthonormalization of wavefunctions which is cubic versus the number of atoms because the number of scalar products grows quadratically and the cost of one scalar product is linear. The cost of the application of the Hamiltonian on one wavefunction is $N \log(N)$ due to the Fast Fourier Transform which is almost linear. So the cost of calculating the Hamiltonian over the whole set of the wavefunctions grows quadratically.

This means that the use of plane wave basis sets for the Kohn-Sham equations is limited to a few hundred of atoms.

We show in Figure 4, the flowchart of the diagonalization scheme applied to the plane wave basis sets. The advantage is that metallic systems or systems with a small gap can be properly calculated. What we need is to have a good family of pseudopotentials and a good density mixing. If the electronic density coming from the new set of wavefunctions is used

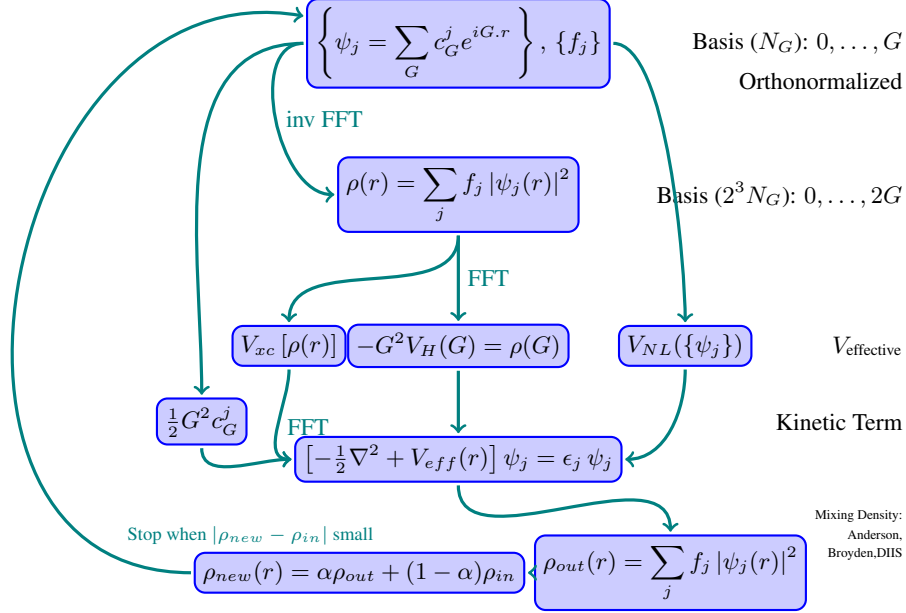


Figure 4. Diagonalization Scheme: Flowchart

directly, the calculation does not converge which is a consequence of the non-linearity of the equations in function of the electronic density. To circumvent this problem, density mixing is used as Anderson, Broyden, DIIS mixing. The robustness of a code is mainly due to the choice of good density mixing.

4.2 Daubechies wavelets family

Daubechies wavelets³ have virtually all the properties that one might desire for a basis set. They form a systematic orthogonal and smooth basis that is localized both in real and Fourier space and that allows for adaptivity. A DFT approach based on such functions will meet both the requirements of precision and localization found in many applications. We will in the following describe in detail a DFT method based on a Daubechies wavelets basis set. This method is implemented in a DFT code, named BigDFT, distributed under GNU-GPL license and integrated in the ABINIT⁴ software package. In the next few paragraphs we will discuss the importance of the properties of Daubechies wavelets in the context of electronic structure calculations.

A wavelet basis consists of a family of functions generated from a mother function and its translations on the points of a uniform grid of spacing h . The number of basis functions is increased by decreasing the value of h . Thanks to the systematicity of the basis, this will make the numerical description more precise. The degree of smoothness determines the speed with which one converges to the exact result as h is decreased. The degree of smoothness increases as one goes to higher order Daubechies wavelets. In our method we

use Daubechies wavelets of order 16. This together with the fact that our method is quasi variational gives a convergence rate of h^{14} . Obtaining such a high convergence rate is essential in the context of electronic structure calculations where one needs highly accurate results for basis sets of acceptable size. The combination of adaptivity and a high order convergence rate is typically not achieved in other electronic structure programs using systematic real space methods.⁶ An adaptive finite element code, using cubic polynomial shape functions,⁷ has a convergence rate of h^6 . Finite difference methods have sometimes low⁸ h^3 or high convergence rates⁹ but are not adaptive.

The most important property of these functions is that they satisfy the so-called refinement equations

$$\begin{aligned}\phi(x) &= \sqrt{2} \sum_{j=1-m}^m h_j \phi(2x - j) \\ \psi(x) &= \sqrt{2} \sum_{j=1-m}^m g_j \phi(2x - j)\end{aligned}\tag{35}$$

which establishes a relation between the scaling functions on a grid with grid spacing h and another one with spacing $h/2$. h_j and $g_j = (-1)^j h_{-j+1}$ are the elements of a filter that characterizes the wavelet family, and m is the order of the scaling function-wavelet family. All the properties of these functions can be obtained from the relations (35). The full basis set can be obtained from all translations by a certain grid spacing h of the mother function centered at the origin. The mother function is localized, with compact support. The maximally symmetric Daubechies scaling function and wavelet of order 16 that are used in this work are shown in Figure 5.

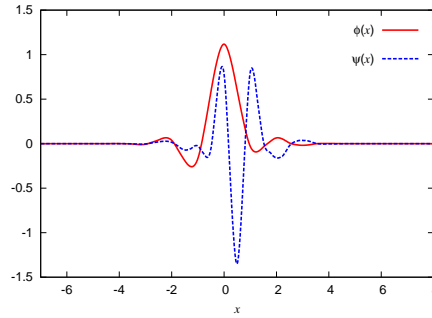


Figure 5. Daubechies scaling function ϕ and wavelet ψ of order 16. Both are different from zero only in the interval from -7 to 8.

For a three-dimensional description, the simplest basis set is obtained by a set of products of equally spaced scaling functions on a grid of grid spacing h'

$$\phi_{i,j,k}(\mathbf{r}) = \phi(x/h' - i) \phi(y/h' - j) \phi(z/h' - k). \tag{36}$$

In other terms, the three-dimensional basis functions are a tensor product of one dimensional basis functions. Note that we are using a cubic grid, where the grid spacing is the

same in all directions, but the following description can be straightforwardly applied to general orthorombic grids.

The basis set of Eq. (36) is equivalent to a mixed basis set of scaling functions on a twice coarser grid of grid spacing $h = 2h'$

$$\phi_{i,j,k}^0(\mathbf{r}) = \phi(x/h - i) \phi(y/h - j) \phi(z/h - k) \quad (37)$$

augmented by a set of 7 wavelets

$$\begin{aligned} \phi_{i,j,k}^1(\mathbf{r}) &= \psi(x/h - i) \phi(y/h - j) \phi(z/h - k) \\ \phi_{i,j,k}^2(\mathbf{r}) &= \phi(x/h - i) \psi(y/h - j) \phi(z/h - k) \\ \phi_{i,j,k}^3(\mathbf{r}) &= \psi(x/h - i) \psi(y/h - j) \phi(z/h - k) \\ \phi_{i,j,k}^4(\mathbf{r}) &= \phi(x/h - i) \phi(y/h - j) \psi(z/h - k) \\ \phi_{i,j,k}^5(\mathbf{r}) &= \psi(x/h - i) \phi(y/h - j) \psi(z/h - k) \\ \phi_{i,j,k}^6(\mathbf{r}) &= \phi(x/h - i) \psi(y/h - j) \psi(z/h - k) \\ \phi_{i,j,k}^7(\mathbf{r}) &= \psi(x/h - i) \psi(y/h - j) \psi(z/h - k) \end{aligned} \quad (38)$$

This equivalence follows from the fact that, from Eq. (56), every scaling function and wavelet on a coarse grid of spacing h can be expressed as a linear combination of scaling functions at the fine grid level h' and vice versa.

The points of the simulation grid fall into 3 different classes. The points which are very far from the atoms will have virtually zero charge density and thus will not carry any basis functions. The remaining grid points are either in the high resolution region which contains the chemical bonds or in the low resolution regions which contains the exponentially decaying tails of the wavefunctions. In the low resolution region one uses only one scaling function per coarse grid point, whereas in the high resolution region one uses both the scaling function and the 7 wavelets. In this region the resolution is thus doubled in each spatial dimension compared to the low resolution region. Figure 6 shows the 2-level adaptive grid around a water molecule.

A wavefunction $\Psi(\mathbf{r})$ can thus be expanded in this basis:

$$\Psi(\mathbf{r}) = \sum_{i_1, i_2, i_3} s_{i_1, i_2, i_3} \phi_{i_1, i_2, i_3}^0(\mathbf{r}) + \sum_{j_1, j_2, j_3} \sum_{\nu=1}^7 d_{j_1, j_2, j_3}^{\nu} \phi_{j_1, j_2, j_3}^{\nu}(\mathbf{r}) \quad (39)$$

The sum over i_1, i_2, i_3 runs over all the grid points contained in the low resolution region and the sum over j_1, j_2, j_3 over all the points contained in the smaller high resolution region.

The decomposition of scaling function into coarser scaling functions and wavelets can be continued recursively to obtain more than 2 resolution levels. We found however that a high degree of adaptivity is not of paramount importance in pseudopotential calculations. In other terms, the pseudopotentials smooth the wavefunctions so that two levels of resolution are enough in most cases to achieve good computational accuracy. In addition, more than two resolution levels lead to more complicated algorithms such as the non-standard operator form¹⁵ that, in turn, lead to larger prefactors.

The transformation from a pure fine scaling function representation (a basis set which contains only scaling functions centered on a finer grid of spacing h') to a mixed coarse

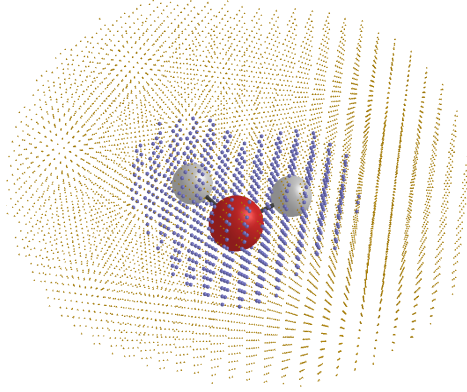


Figure 6. A 2-level adaptive grid around a H₂O molecule. The high resolution grid points carrying both scaling functions and wavelets are shown in blue (larger points), the low resolution grid points carrying only a single scaling function are shown in yellow (smaller points).

scaling function/wavelet representation is done by the fast wavelet transformation¹⁴ which is a convolution and scales linearly with respect to the number of basis functions being transformed.

The wavefunctions are stored in a compressed form where only the nonzero scaling function and wavelets coefficients are stored. The basis set being orthogonal, several operations such as scalar products among different orbitals and between orbitals and the projectors of the non-local pseudopotential can directly be done in this compressed form. In the following sections we will illustrate the main operations which must be performed in the context of a DFT calculation.

5 Overview of the Method

The KS wavefunctions $|\Psi_i\rangle$ are eigenfunctions of the KS Hamiltonian, with pseudopotential V_{psp} :

$$\left(-\frac{1}{2}\nabla^2 + V_{\text{KS}}[\rho]\right)|\Psi_i\rangle = \epsilon_i|\Psi_i\rangle. \quad (40)$$

The KS potential

$$V_{\text{KS}}[\rho] = V_H[\rho] + V_{\text{xc}}[\rho] + V_{\text{ext}}, \quad (41)$$

contains the Hartree potential, solution of the Poisson's equation $\nabla^2 V_H = -4\pi\rho$, the exchange-correlation potential V_{xc} and the external ionic potential V_{ext} acting on the electrons. The method we illustrate in this paper is conceived for isolated systems, namely free boundary conditions.

In our method, we choose the pseudopotential term $V_{\text{ext}} = \sum_{a=1}^N V_{\text{psp}}^{(a)}(\mathbf{r} - \mathbf{R}_a)$ to be of the form of norm-conserving GTH-HGH pseudopotentials,¹⁶⁻¹⁸ which have a local and a nonlocal term, $V_{\text{psp}} = V_{\text{local}} + V_{\text{nonlocal}}$. For each of the ions these potentials have this

form:

$$V_{\text{local}}(\mathbf{r}) = -\frac{Z_{\text{ion}}}{r} \text{erf}\left(\frac{r}{\sqrt{2}r_{\text{loc}}}\right) + \exp\left[-\frac{1}{2}\left(\frac{r}{r_{\text{loc}}}\right)^2\right] \times \\ \times \left[C_1 + C_2\left(\frac{r}{r_{\text{loc}}}\right)^2 + C_3\left(\frac{r}{r_{\text{loc}}}\right)^4 + C_4\left(\frac{r}{r_{\text{loc}}}\right)^6\right] \quad (42)$$

$$V_{\text{nonlocal}} = \sum_{\ell} \sum_{i,j=1}^3 h_{ij}^{(\ell)} |p_i^{(\ell)}\rangle \langle p_j^{(\ell)}| \quad (43)$$

$$\langle \mathbf{r} | p_i^{(\ell)} \rangle = \frac{\sqrt{2}r^{\ell+2(i-1)} \exp\left[-\frac{1}{2}\left(\frac{r}{r_{\ell}}\right)^2\right]}{r_{\ell}^{\ell+(4i-1)/2} \sqrt{\Gamma\left(\ell + \frac{4i-1}{2}\right)}} \sum_{m=-\ell}^{+\ell} Y_{\ell m}(\theta, \phi),$$

where $Y_{\ell m}$ are the spherical harmonics, and r_{loc}, r_{ℓ} are, respectively, the localization radius of the local pseudopotential term and of each projector.

The analytic form of the pseudopotentials together with the fact that their expression in real space can be written in terms of a linear combination of tensor products of one dimensional functions is of great utility in our method.

Each term in the Hamiltonian is implemented differently, and will be illustrated in the following sections. After the application of the Hamiltonian, the KS wavefunctions are updated via a direct minimization scheme,¹⁹ which in its actual implementation is fast and reliable for non-zero gap systems, namely insulators. Since we are using direct minimization algorithm, at present we have concentrated on systems with a gap, however we see no reason why the method can not be extended to metallic systems.

6 Treatment of Kinetic Energy

The matrix elements of the kinetic energy operator among the basis functions of our mixed representation (i.e. scaling functions with scaling functions, scaling function with wavelets and wavelets with wavelets) can be calculated analytically.²⁰ For simplicity, let us illustrate the application of the kinetic energy operator onto a wavefunction Ψ that is only expressed in terms of scaling functions.

$$\Psi(x, y, z) = \sum_{i_1, i_2, i_3} s_{i_1, i_2, i_3} \phi(x/h - i_1) \phi(y/h - i_2) \phi(z/h - i_3)$$

The result of the application of the kinetic energy operator on this wavefunction, projected to the original scaling function space, has the expansion coefficients

$$\hat{s}_{i_1, i_2, i_3} = -\frac{1}{2h^3} \int \phi(x/h - i_1) \phi(y/h - i_2) \phi(z/h - i_3) \times \\ \times \nabla^2 \Psi(x, y, z) dx dy dz.$$

Analytically the coefficients s_{i_1, i_2, i_3} and \hat{s}_{i_1, i_2, i_3} are related by a convolution

$$\hat{s}_{i_1, i_2, i_3} = \frac{1}{2} \sum_{j_1, j_2, j_3} K_{i_1-j_1, i_2-j_2, i_3-j_3} s_{j_1, j_2, j_3} \quad (44)$$

where

$$K_{i_1, i_2, i_3} = T_{i_1} T_{i_2} T_{i_3}, \quad (45)$$

where the coefficients T_i can be calculated analytically via an eigenvalue equation:

$$\begin{aligned} T_i &= \int \phi(x) \frac{\partial^2}{\partial x^2} \phi(x - i) dx \\ &= \sum_{\nu, \mu} 2h_\nu h_\mu \int \phi(2x - \nu) \frac{\partial^2}{\partial x^2} \phi(2x - 2i - \mu) dx \\ &= \sum_{\nu, \mu} 2h_\nu h_\mu 2^{2-1} \int \phi(y - \nu) \frac{\partial^2}{\partial y^2} \phi(y - 2i - \mu) dy \\ &= \sum_{\nu, \mu} h_\nu h_\mu 2^2 \int \phi(y) \frac{\partial^2}{\partial y^2} \phi(y - 2i - \mu + \nu) dy \\ &= \sum_{\nu, \mu} h_\nu h_\mu 2^2 T_{2i - \nu + \mu} \end{aligned}$$

Using the refinement equation (56), the values of the T_i can be calculated analytically, from a suitable eigenvector of a matrix derived from the wavelet filters.²⁰ For this reason the expression of the kinetic energy operator is *exact* in a given Daubechies basis.

Since the 3-dimensional kinetic energy filter K_{i_1, i_2, i_3} is a product of three one-dimensional filters (Eq. (45)) the convolution in Eq. (44) can be evaluated with $3N_1N_2N_3L$ operations for a three-dimensional grid of $N_1N_2N_3$ grid points. L is the length of the one-dimensional filter which is 29 for our Daubechies family. The kinetic energy can thus be evaluated with linear scaling with respect to the number of nonvanishing expansion coefficients of the wavefunction. This statement remains true for a mixed scaling function-wavelet basis where we have both nonvanishing s and d coefficients and for the case where the low and high resolution regions cover only parts of the cube of $N_1N_2N_3$ grid points.

The Daubechies wavefunctions of degree 16 have an approximation error of h^8 , i.e. the difference between the exact wavefunction and its representation in a finite basis set (Eq. (39)) is decreasing as h^8 . The error of the kinetic energy in a variational scheme decreases then as $h^{2 \cdot 8 - 2} = h^{14}$.²¹ As we will see the kinetic energy is limiting the convergence rate in our scheme and the overall convergence rate is thus h^{14} . Figure 7 shows this asymptotic convergence rate.

6.1 Treatment of local potential energy

In spite of the striking advantages of Daubechies wavelets the initial exploration of this basis set²² did not lead to any algorithm that would be useful for practical electronic structure calculations. This was due to the fact that an accurate evaluation of the local potential energy is difficult in a Daubechies wavelet basis.

By definition, the local potential $V(\mathbf{r})$ can be easily known on the nodes of the uniform grid of the simulation box. Approximating a potential energy matrix element $V_{i,j,k;i',j',k'}$

$$V_{i,j,k;i',j',k'} = \int d\mathbf{r} \phi_{i',j',k'}(\mathbf{r}) V(\mathbf{r}) \phi_{i,j,k}(\mathbf{r})$$

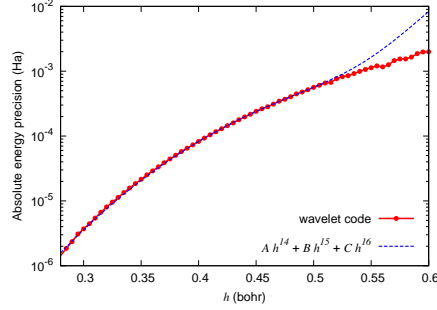


Figure 7. Convergence rate $\mathcal{O}(h^{14})$ of the wavelet code for a test run on a carbon atom. For this run the interpolation parameters are found to be, within 2% accuracy: $A = 344$, $B = -1239$, $C = 1139$. Using lower powers of h for the fit does not give accurate agreement. Other test systems gave comparable convergence rates.

by

$$V_{i,j,k;i',j',k'} \approx \sum_{l,m,n} \phi_{i',j',k'}(\mathbf{r}_{l,m,n}) V(\mathbf{r}_{l,m,n}) \phi_{i,j,k}(\mathbf{r}_{l,m,n})$$

gives an extremely slow convergence rate with respect to the number of grid points used to approximate the integral because a single scaling function is not very smooth, i.e. it has a rather low number of continuous derivatives. A. Neelov and S. Goedecker²⁴ have shown that one should not try to approximate a single matrix element as accurately as possible but that one should try instead to approximate directly the expectation value of the local potential. The reason for this strategy is that the wavefunction expressed in the Daubechies basis is smoother than a single Daubechies basis function. A single Daubechies scaling function of order 16 (i.e. the corresponding wavelet has 8 vanishing moments) has only 2 continuous derivatives. More precisely its index of Hölder continuity is about 2.7 and the Sobolev space regularity with respect to $p = 2$ is about 2.91.²³ A single Daubechies scaling function of order 16 has only 4 continuous derivatives. By suitable linear combinations of Daubechies 16 one can however exactly represent polynomials up to degree 7, i.e functions that have 7 non-vanishing continuous derivatives. The discontinuities get thus canceled by taking suitable linear combinations. Since we use pseudopotentials, our exact wavefunctions are analytic and can locally be represented by a Taylor series. We are thus approximating functions that are approximately polynomials of order 7 and the discontinuities nearly cancel.

Instead of calculating the exact matrix elements we therefore use matrix elements with respect to a smoothed version $\tilde{\phi}$ of the Daubechies scaling functions.

$$V_{i,j,k;i',j',k'} \approx \sum_{l,m,n} \tilde{\phi}_{i',j',k'}(\mathbf{r}_{l,m,n}) V(\mathbf{r}_{l,m,n}) \tilde{\phi}_{i,j,k}(\mathbf{r}_{l,m,n}) = \sum_{l,m,n} \tilde{\phi}_{0,0,0}(\mathbf{r}_{l-i',m-j',n-k'}) V(\mathbf{r}_{l,m,n}) \tilde{\phi}_{0,0,0}(\mathbf{r}_{l-i,m-j,n-k}) \quad (46)$$

where the smoothed wavefunction is defined by

$$\tilde{\phi}_{0,0,0}(\mathbf{r}_{l,m,n}) = \omega_l \omega_m \omega_n$$

and ω_l is the “magic filter”. The relation between the true functional values, i.e. the scaling function, and ω is shown in Figure 8. Even though Eq. (46) is not a particularly good approximation for a single matrix element it gives an excellent approximation for the expectation values of the local potential energy

$$\int dx \int dy \int dz \Psi(x, y, z) V(x, y, z) \Psi(x, y, z)$$

and also for matrix elements between different wavefunctions

$$\int dx \int dy \int dz \Psi_i(x, y, z) V(x, y, z) \Psi_j(x, y, z)$$

in case they are needed. Because of this remarkable achievement of the filter ω we call it the magic filter.

In practice we do not explicitly calculate any matrix elements but we apply only filters to the wavefunction expansion coefficients as will be shown in the following. This is mathematically equivalent but numerically much more efficient.

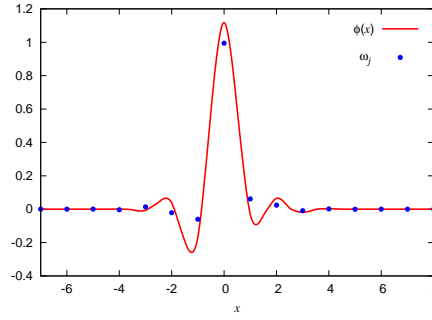


Figure 8. The magic filter ω_i for the least asymmetric Daubechies-16 basis.

Since the operations with the local potential V are performed in the computational box on the double resolution grid with grid spacing $h' = h/2$, we must perform a wavelet transformation before applying the magic filters. These two operations can be combined in one, giving rise to modified magic filters both for scaling functions and wavelets on the original grid of spacing h . These modified magic filters can be obtained from the original ones using the refinement relations and they are shown in Figures 9 and 10. Following the same guidelines as the kinetic energy filters, the smoothed real space values $\tilde{\Psi}_{i,j,k}$ of a wavefunction Ψ are calculated by performing a product of three one-dimensional convolutions with the magic filters along the x , y and z directions. For the scaling function part of the wavefunction the corresponding formula is :

$$\tilde{\Psi}_{i_1, i_2, i_3} = \sum_{j_1, j_2, j_3} s_{j_1, j_2, j_3} v_{i_1-2j_1}^{(1)} v_{i_2-2j_2}^{(1)} v_{i_3-2j_3}^{(1)}$$

where $v_i^{(1)}$ is the filter that maps a scaling function on a double resolution grid. Similar convolutions are needed for the wavelet part. The calculation is thus similar to the treatment of the Laplacian in the kinetic energy.

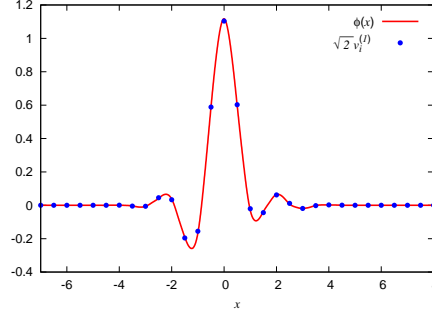


Figure 9. The fine scale magic filter $v_i^{(1)}$ (combination of a wavelet transform and the magic filter in Figure 8) for the least asymmetric Daubechies-16 basis, scaled by $\sqrt{2}$ for comparison with the scaling function. The values of the filter on the graph are almost undistinguishable from the values of the scaling function. However, there is a slight difference which is important for the correct asymptotic convergence at small values of grid spacing h .

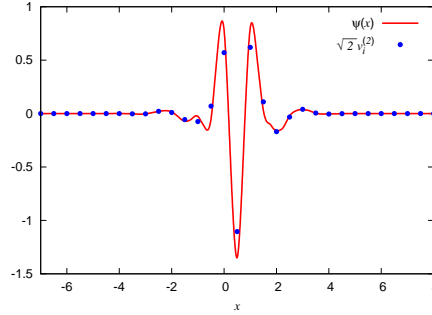


Figure 10. The fine scale magic filter $v_i^{(2)}$ (combination of a wavelet transform and the magic filter in Figure 8) for the least asymmetric Daubechies-16 *wavelet*, scaled by $\sqrt{2}$ for comparison with the wavelet itself.

Once we have calculated $\tilde{\Psi}_{i,j,k}$ the approximate expectation value ϵ_V of the local potential V for a wavefunction Ψ is obtained by simple summation on the double resolution real space grid:

$$\epsilon_V = \sum_{j_1, j_2, j_3} \tilde{\Psi}_{j_1, j_2, j_3} V_{j_1, j_2, j_3} \tilde{\Psi}_{j_1, j_2, j_3}$$

The evaluation of the local potential energy ϵ_V converges with a convergence rate of h^{16} to the exact value where h is the grid spacing. Therefore, the potential energy has a convergence rate two powers of h faster than the rate for the kinetic energy.

6.2 Treatment of the non-local pseudopotential

The energy contributions from the non-local pseudopotential have for each angular momentum l the form

$$\sum_{i,j} \langle \Psi | p_i \rangle h_{ij} \langle p_j | \Psi \rangle$$

where $|p_i\rangle$ is a pseudopotential projector. Once applying the Hamiltonian operator, the application of one projector on the wavefunctions requires the calculation of

$$|\Psi\rangle \rightarrow |\Psi\rangle + \sum_{i,j} |p_i\rangle h_{ij} \langle p_j|\Psi\rangle .$$

If we use for the projectors the representation of Eq. (39) (i.e. the same as for the wavefunctions) both operations are trivial to perform. Because of the orthogonality of the basis set we just have to calculate scalar products among the coefficient vectors and to update the wavefunctions. The scaling function and wavelet expansion coefficients for the projectors are given by¹⁴

$$\int p(\mathbf{r}) \phi_{i_1, i_2, i_3}(\mathbf{r}) d\mathbf{r} , \quad \int p(\mathbf{r}) \psi_{i_1, i_2, i_3}^\nu(\mathbf{r}) d\mathbf{r} . \quad (47)$$

where we used the notation (37),(38).

The GTH-HGH pseudopotentials^{16,17} have projectors which are written in terms of gaussians times polynomials. This form of projectors is particularly convenient to be expanded in the Daubechies basis. In other terms, since the general form of the projector is

$$\langle \mathbf{r}|p\rangle = e^{-cr^2} x^{\ell_x} y^{\ell_y} z^{\ell_z} ,$$

the 3-dimensional integrals can be calculated easily since they can be factorized into a product of 3 one-dimensional integrals.

$$\int \langle \mathbf{r}|p\rangle \phi_{i_1, i_2, i_3}(\mathbf{r}) d\mathbf{r} = W_{i_1}(c, \ell_x) W_{i_2}(c, \ell_y) W_{i_3}(c, \ell_z) , \quad (48)$$

$$W_j(c, \ell) = \int_{-\infty}^{+\infty} e^{-ct^2} t^\ell \phi(t/h - j) dt \quad (49)$$

The one-dimensional integrals are calculated in the following way. We first calculate the scaling function expansion coefficients for scaling functions on a one-dimensional grid that is 16 times denser. The integration on this dense grid is done by the well-known quadrature introduced in,²⁸ that coincides with the magic filter.²⁴ This integration scheme based on the magic filter has a convergence rate of h^{16} and we gain therefore a factor of 16^{16} in accuracy by going to a denser grid. This means that the expansion coefficients are for reasonable grid spacings h accurate to machine precision. After having obtained the expansion coefficients with respect to the fine scaling functions we obtain the expansion coefficients with respect to the scaling functions and wavelets on the required resolution level by one-dimensional fast wavelet transformations. No accuracy is lost in the wavelet transforms and our representation of the projectors is therefore typically accurate to nearly machine precision. In order to treat with the same advantages other pseudopotentials which are not given under the form of gaussians it would be necessary to approximate them by a small number of gaussians.

6.3 XC functionals and implementation of GGA's

To calculate the exchange correlation energy per particle $\epsilon^{xc}[\rho](\mathbf{r})$ and the associated XC potential $V_{xc}(\mathbf{r})$ it is important to have a real-space representation of the density. The

magic filter procedure described in Section 6.1 can be used also to express the real-point values of the charge density.

$$\rho(\mathbf{r}) = \sum_i n_{\text{occ}}^{(i)} |\tilde{\Psi}_i(\mathbf{r})|^2, \quad (50)$$

Evidently, any real-space based implementation of the XC functionals fits well with this density representation. In our program we use the XC functionals as implemented in `libXC`¹³ exchange-correlation library.

A traditional finite difference scheme of fourth order is used on the double resolution grid to calculate the gradient of the charge density

$$\partial_w \rho(\mathbf{r}_{i_1, i_2, i_3}) = \sum_{j_1, j_2, j_3} c_{i_1, i_2, i_3; j_1, j_2, j_3}^{(w)} \rho_{j_1, j_2, j_3}, \quad (51)$$

where $w = x, y, z$. For grid points close to the boundary of the computational volume the above formula requires grid points outside the volume. For free boundary conditions the values of the charge density outside the computational volume in a given direction are taken to be equal to the value at the border of the grid.

As described in Section 2.2, the relation between the gradient and the density must be taken into account when calculating V_{xc} in the standard White-Bird approach,²⁷ where the density gradient is considered as an explicit functional of the density. There the XC potential can be split in two terms:

$$V_{\text{xc}}(\mathbf{r}_{i_1, i_2, i_3}) = V_{\text{xc}}^o(\mathbf{r}) + V_{\text{xc}}^c(\mathbf{r}), \quad (52)$$

where

$$\begin{aligned} V_{\text{xc}}^o(\mathbf{r}_{i_1, i_2, i_3}) &= \epsilon_{\text{xc}}(\mathbf{r}) + \rho(\mathbf{r}) \frac{\partial \epsilon_{\text{xc}}}{\partial \rho}(\mathbf{r}), \\ V_{\text{xc}}^c(\mathbf{r}_{i_1, i_2, i_3}) &= \sum_{j_1, j_2, j_3} \frac{\rho}{|\nabla \rho|} \frac{\partial \epsilon_{\text{xc}}}{\partial |\nabla \rho|}(\mathbf{r}_{j_1, j_2, j_3}) \times \\ &\quad \times \sum_{w=x, y, z} \partial_w \rho(\mathbf{r}_{j_1, j_2, j_3}) c_{j_1, j_2, j_3; i_1, i_2, i_3}^{(w)}, \end{aligned} \quad (53)$$

where the “ordinary” part V_{xc}^o is present in the same form of LDA functionals, while the White-Bird “correction” term V_{xc}^c appears only when the XC energy depends explicitly on $|\nabla \rho|$. The $c^{(w)}$ are the coefficients of the finite difference formula used to calculate the gradient of the charge density (51).

The evaluation of the XC terms and also, when needed, the calculation of the gradient of the charge density, may easily be performed together with the Poisson solver used to evaluate the Hartree potential. This allows us to save computational time.

7 Calculation of Hartree Potential

Electrostatic potentials play a fundamental role in nearly any field of physics and chemistry. Having efficient algorithms to find the electrostatic potential V arising from a charge distribution ρ or, in other words, to solve the Poisson’s equation

$$\nabla^2 V = -4\pi\rho, \quad (54)$$

is therefore essential. The large variety of situations in which this equation can be found lead us to face this problem with different choices of the boundary conditions (BC). The long-range behavior of the inverse Laplacian operator make this problem to be strongly dependent on the BC of the system.

The most immediate approach to the Poisson equation can be achieved for periodic BC, where a traditional reciprocal space treatment is both rapid and simple, since the Laplacian matrix is diagonal in a plane wave representation. If the density ρ is originally given in real space, a first Fast Fourier Transformation (FFT) is used to transform the real space data in reciprocal space. The Poisson equation is then solved in reciprocal space and finally the result is transformed back into real space by a second FFT. Because of the FFT's, the overall computational scaling is $\mathcal{O}(N \log N)$ with respect to the number of grid points N .

The situation is different if one considers the same problem for different BC, like for example free (isolated) BC. In this case the solution of Poisson's equation can formally be obtained from a three-dimensional integral:

$$V(\mathbf{r}) = \int d\mathbf{r}' G(|\mathbf{r} - \mathbf{r}'|) \rho(\mathbf{r}') , \quad (55)$$

where $G(r) = 1/r$ is the Green function of the Laplacian operator in the unconstrained \mathbb{R}^3 space. The long range nature of the kernel operator G does not allow us to mimic free BC with a very large periodic volume. Consequently, the description of non-periodic systems with a periodic formalism always introduces long-range interactions between super-cells that falsify the results. Due to the simplicity of the plane wave methods, various attempts have been made to generalize the reciprocal space approach to free BC.⁴⁰⁻⁴² All of them use a FFT at some point, and have thus a $\mathcal{O}(N \log N)$ scaling. These methods have some restrictions and cannot be used blindly. For example, the method by Füsti-Molnar and Pulay is efficient only for spherical geometries, and the method by Martina and Tuckerman requires artificially large simulation boxes that are expensive numerically. Nonetheless, the usefulness of reciprocal space methods has been demonstrated for a variety of applications, and plane-wave based codes are widely used in the chemical physics community.

Another choice of the BC that is of great interest is for systems that are periodically replicated in two dimensions but with finite extent in the third, namely surface systems. The surface-specific experimental techniques developed in recent years produced important results,⁴³ that can benefit from theoretical prediction and analysis. The development of efficient techniques for systems with such boundary conditions thus became of great importance. Explicit Poisson solvers have been developed in this framework,⁴⁴⁻⁴⁶ with a reciprocal space based treatment. Essentially, these Poisson solvers are built following a suitable generalization for surfaces BC of the same methods that were developed for isolated systems. As for the free BC case, screening functions are present to subtract the artificial interaction between the super-cells in the non-periodic direction. Therefore, they exhibit the same kind of intrinsic limitations, as for example a good accuracy only in the bulk of the computational region, with the consequent need for artificially large simulation boxes which may increase the computational overhead.

Electrostatic potentials can either be calculated by solving the differential Poisson equation or by solving the equivalent integral equation Eq. (55). The methods that solve the differential equation are iterative and they require various tuning. A good representative of these methods is the multigrid approach.⁴⁷ Several different steps such as *smoothing*,

restriction and *prolongation* are needed in this approach. Each of these steps has to be tuned to optimize speed and accuracy. Approaches based on the integral equation are in contrast straightforward and do not require such tuning.

In the following, we will describe two Poisson solvers compatible with free and surfaces boundary conditions respectively. Contrary to Poisson solvers based on reciprocal space treatment, the fundamental operations of these Poisson solver are based on a mixed reciprocal-real space representation of the charge density. This allows us to naturally satisfy the boundary conditions in the different directions. Screening functions or other approximations are thus not needed.

7.1 Interpolating scaling functions

Interpolating scaling functions (ISF)⁴⁹ arise in the framework of wavelet theory.^{3,14} They are one-dimensional functions, and their three main properties are:

- The full basis set can be obtained from all the translations by a certain grid spacing h of the mother function ϕ centered at the origin.
- They satisfy the refinement relation:

$$\phi(x) = \sum_{j=-m}^m h_j \phi(2x - j) \quad (56)$$

where the h_j 's are the elements of a filter that characterizes the wavelet family, and m is the order of the scaling function. Eq. (56) establishes a relation between the scaling functions on a grid with grid spacing h and another one with spacing $h/2$.

- The mother function ϕ is symmetric, with compact support from $-m$ to m . It is equal to one at the origin and to zero at all other integer points (in grid spacing units). The expansion coefficients of any function in this basis are just the values of the function on the grid.
- Given a function in the ISF basis

$$f(x) = \sum_j f_j \phi\left(\frac{x}{h} - j\right) \quad (57)$$

the first m discrete and continuous moments are identical for a m -th order interpolating wavelet family, i.e.

$$h^\ell \sum_j j^\ell f_j = \int dx x^\ell f(x), \quad (58)$$

if $\ell < m$. This follows from the fact, proven in reference⁴⁸ that the first m moments of the scaling function obey the formula:

$$M_l = \int \phi(x) x^l dx = \delta_l, \quad l = 0, \dots, m-1 \quad (59)$$

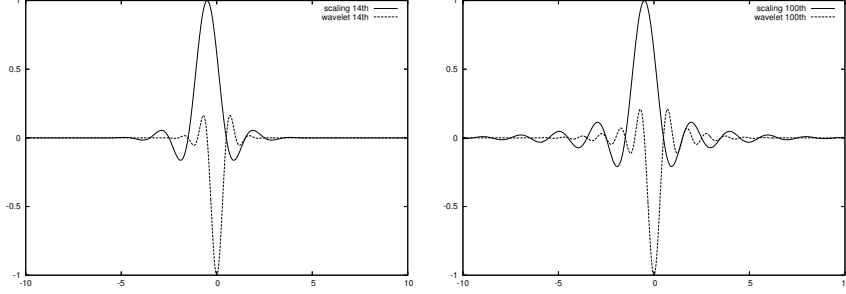


Figure 11. Plots of interpolating scaling functions and wavelets of 14-th and 100-th order.

Shift the integration variable, we have

$$\begin{aligned} \int \phi(x-j)x^l dx &= \int \phi(t)(t+j)^l dt = \\ &= \int \phi(t) \sum_{p=0}^l C_l^p t^p j^{l-p} dt = j^l \end{aligned}$$

Since the various multipoles of the charge distribution determine the major features of the potential the above equalities tell us that a scaling function representation gives the most faithful mapping between a continuous and discretized charge distribution for electrostatic problems. Figure 11 shows an 14-th order and 100-th order interpolating scaling function.

7.2 Poisson solver for Free BC

Continuous charge distributions are represented in numerical work typically by their values $\rho_{i,j,k}$ on a grid. It follows from the above described properties of interpolating scaling functions that the corresponding continuous charge density is given by:

$$\rho(\mathbf{r}) = \sum_{i_1, i_2, i_3} \rho_{i_1, i_2, i_3} \phi(x - i_1) \phi(y - i_2) \phi(z - i_3) \quad (60)$$

Denoting the potential on the grid point: $\mathbf{r}_{j_1, j_2, j_3} = (x_{j_1}, y_{j_2}, z_{j_3})$ by $V_{j_1, j_2, j_3} = V(\mathbf{r}_{j_1, j_2, j_3})$ we have:

$$\begin{aligned} V_{j_1, j_2, j_3} &= \\ &= \sum_{i_1, i_2, i_3} \rho_{i_1, i_2, i_3} \int d\mathbf{r}' \frac{\phi_{i_1}(x') \phi_{i_2}(y') \phi_{i_3}(z')}{|\mathbf{r}_{j_1, j_2, j_3} - \mathbf{r}'|}. \end{aligned} \quad (61)$$

The above integral defines the discrete kernel:

$$\begin{aligned} K(i_1, j_1; i_2, j_2; i_3, j_3) &= \\ &= \int d\mathbf{r}' \phi_{i_1}(x') \phi_{i_2}(y') \phi_{i_3}(z') \frac{1}{|\mathbf{r}_{j_1, j_2, j_3} - \mathbf{r}'|}. \end{aligned} \quad (62)$$

Since the problem is invariant under combined translations of both the source point (i_1, i_2, i_3) and the observation point (j_1, j_2, j_3) the kernel depends only on the difference of the indices:

$$K(i_1, j_1; i_2, j_2; i_3, j_3) = K(i_1 - j_1, i_2 - j_2, i_3 - j_3) \quad (63)$$

and the potential V_{j_1, j_2, j_3} can be obtained from the charge density ρ_{i_1, i_2, i_3} by the following 3-dimensional convolution:

$$V_{j_1, j_2, j_3} = \sum_{i_1, i_2, i_3} K(i_1 - j_1, i_2 - j_2, i_3 - j_3) \rho_{i_1, i_2, i_3} . \quad (64)$$

Once the kernel is available in Fourier space, this convolution can be evaluated with two FFTs at a cost of $O(N \log N)$ operations where $N = n_1 n_2 n_3$ is the number of 3-dimensional grid points. Since all the quantities in the above equation are real, real-to-complex FFT's can be used to reduce the number of operations compared to the case where one would use ordinary complex-complex FFT's. Obtaining the kernel in Fourier space from the kernel $K(j_1, j_2, j_3)$ in real space requires another FFT.

It remains now to calculate the values of all the elements of the kernel $K(k_1, k_2, k_3)$. Solving a 3-dimensional integral for each element would be much too costly and we use therefore a separable approximation of $1/r$ in terms of Gaussians,^{12,50}

$$\frac{1}{r} \simeq \sum_k \omega_k e^{-p_k r^2} . \quad (65)$$

In this way all the complicated 3-dimensional integrals become products of simple 1-dimensional integrals. Using 89 Gaussian functions with the coefficients ω_k and p_k suitably chosen, we can approximate $\frac{1}{r}$ with an error less than 10^{-8} in the interval $[10^{-9}, 1]$. If we are interested in a wider range, e.g. a variable R going from zero to L , we can use $r = \frac{R}{L}$:

$$\frac{L}{R} = \sum_k \omega_k e^{-\frac{p_k}{L^2} R^2} , \quad (66)$$

$$\frac{1}{R} = \frac{1}{L} \sum_k \omega_k e^{-P_k R^2} , \quad (67)$$

$$P_k = \frac{p_k}{L^2} . \quad (68)$$

With this approximation, we have that

$$K_{j_1, j_2, j_3} = \sum_{k=1}^{89} \omega_k K_{j_1}(p_k) K_{j_2}(p_k) K_{j_3}(p_k) , \quad (69)$$

where

$$K_j(p_k) = \int \varphi_j(x) e^{-p_k x^2} dx \quad (70)$$

$$= \int \varphi_0(x) e^{-p_k (x-j)^2} dx . \quad (71)$$

So we only need to evaluate $89 \times \max(\{n_1, n_2, n_3\})$ integrals of the type

$$K_j(p) = \int \varphi_0(x) e^{-p(x-j)^2} dx, \quad (72)$$

for some value of p chosen between $3 \cdot 10^{-5}$ and $3 \cdot 10^{16}$.

The accuracy in calculating the integrals can be further improved by using the refinement relation for interpolating scaling functions (56).

From (72), we can evaluate $K_i(4p)$ as:

$$K_i(4p) = \int \varphi(x) e^{-4p(x-i)^2} dx \quad (73)$$

$$= \frac{1}{2} \int \varphi(x/2) e^{-p(x-2i)^2} dx \quad (74)$$

$$= \frac{1}{2} \sum_j h_j \int \varphi_j(x) e^{-p(x-2i)^2} dx \quad (75)$$

$$= \frac{1}{2} \sum_j h_j K_{2i-j}(p). \quad (76)$$

The best accuracy in evaluating numerically the integral is attained for $p < 1$. For a fixed value of p given by Eq. (65), the relation (76) is iterated $n = \lceil \log_4(p) \rceil$ times starting with $p_0 = \frac{p}{4^n}$. So the numerical calculation of the integrals $K_i(p)$ is performed as follows: for each p , we compute the number n of required recursions levels and calculate the integral $K_i(p_0)$. The value of n is chosen such that $p_0 \simeq 1$ so we have a gaussian function not too sharp. The evaluation of the interpolating scaling functions is fast on a uniform grid of points so we perform a simple summation over all the grid points.

7.3 Poisson solver for Surface Boundary conditions

Consider a three-dimensional domain, periodic (with period L_x and L_y) in x and y directions, and non-periodic in z . Without loss of generality, a function f that lives in such a domain can be expanded as:

$$f(x, y, z) = \sum_{p_x, p_y} e^{-2\pi i(\frac{p_x}{L_x}x + \frac{p_y}{L_y}y)} f_{p_x, p_y}(z). \quad (77)$$

We indicate with $f_{p_x, p_y}(z)$ the one-dimensional function associated to the vector $\vec{p} = (p_x/L_x, p_y/L_y)$ in the reciprocal space of the two dimensional surface. Following these conventions, the Poisson equation (54) becomes a relation between the reciprocal space components of V and ρ :

$$V_{p_x, p_y}(z) = -4\pi \int_{-\infty}^{+\infty} dz' G(2\pi |\vec{p}|; z - z') \rho_{p_x, p_y}(z), \quad (78)$$

where $|\vec{p}|^2 = (p_x/L_x)^2 + (p_y/L_y)^2$, and $G(\mu; z)$ is the Green function of the one-dimensional Helmholtz equation:

$$(\partial_z^2 - \mu^2) G(\mu; z) = \delta(z). \quad (79)$$

The free BC on the z direction fix the form of the Green function:

$$G(\mu; z) = \begin{cases} -\frac{1}{2\mu} e^{-\mu|z|} & \mu > 0, \\ \frac{1}{2}|z| & \mu = 0. \end{cases} \quad (80)$$

In numerical calculations continuous charge distributions are typically represented by their values on a grid. The mixed representation of the charge density given above immediately suggests to use a plane wave expansion in the periodic directions, which may be easily treated with conventional FFT techniques. For the non-periodic direction z we will use interpolating scaling functions representation. The corresponding continuous charge distribution is thus given by:

$$\rho(x, y, z) = \sum_{p_x = -\frac{N_x}{2}}^{\frac{N_x}{2}} \sum_{p_y = -\frac{N_y}{2}}^{\frac{N_y}{2}} \sum_{j_z=0}^{N_z} \rho_{p_x, p_y; j_z} \times \exp \left\{ -2\pi i \left(\frac{p_x}{L_x} x + \frac{p_y}{L_y} y \right) \right\} \phi \left(\frac{z}{h} - j_z \right), \quad (81)$$

where h is the grid spacing in the z direction, and $\phi(j) = \delta_{j,0}$, $j \in \mathbb{Z}$.

Combining Eq. (78) with (81), the discretized Poisson problem thus becomes:

$$V_{p_x, p_y; j_z} = -4\pi h \sum_{j'_z} K(2\pi |\vec{p}|; j_z - j'_z) \rho_{p_x, p_y; j'_z}, \quad (82)$$

where the quantity (kernel):

$$K(\mu; j) = \int_{-\infty}^{+\infty} du G(\mu; h(j-u)) \phi(u) \quad (83)$$

is defined via an integral in the dimensionless variable u . Due to the symmetry of ϕ , the kernel is symmetric in the non-periodic direction $K(\mu; j_z) = K(\mu; -j_z)$. The integral bounds can be restricted from $-m$ to m , thanks to the compact support of ϕ .

Once we have calculated the kernel, which will be described below, our numerical procedure is the following. We perform a two-dimensional FFT on our real space charge density to obtain the Fourier coefficients $\rho_{p_x, p_y; j'_z}$ for all the periodic planes. Then we have to solve Eq. (82). Since this equation is a convolution it can be calculated by zero-padded FFT's. Finally the potential is transformed back from the mixed representation to real space to obtain the potential on the grid by another two-dimensional FFT. Due to the FFT's, the total computational cost is $\mathcal{O}(N \log N)$. Since all quantities are real, the amount of memory and the number of operations for the FFT can be reduced by using real-to-complex FFT's instead of complex-complex FFT's.

It remains now to calculate the values of the kernel function $K(\mu; j)$. The core of the calculation is represented by the function

$$\tilde{K}(\lambda; j) = \begin{cases} \int du e^{-\lambda|u-j|} \phi(u) & \lambda > 0, \\ \int du |u-j| \phi(u) & \lambda = 0. \end{cases} \quad (84)$$

The kernel has the properties $K(\mu; j) = -\tilde{K}(\mu h; j)/(2\mu)$ for $\mu > 0$ and $K(0; j) = \tilde{K}(0; j)/2$. A simple numerical integration with the trapezoidal rule is inefficient since $G(\mu; z)$ is not smooth in $z = 0$ while the scaling function varies significantly

around the integer points. Thanks to the compact support of the scaling function, this problem can be circumvented with a simple and efficient recursive algorithm. We define two functions $\tilde{K}^{(+)}$ and $\tilde{K}^{(-)}$ such that $\tilde{K}(\lambda; j) = \tilde{K}^{(+)}(\lambda; j) + \tilde{K}^{(-)}(\lambda; j)$, where we have, for $\lambda > 0$

$$\tilde{K}^{(+)}(\lambda; j) = \int_{-\infty}^j du e^{\lambda(u-j)} \phi(u) , \quad (85)$$

$$\tilde{K}^{(-)}(\lambda; j) = \int_j^{+\infty} du e^{-\lambda(u-j)} \phi(u) , \quad (86)$$

while with $\lambda = 0$

$$\tilde{K}^{(\pm)}(0; j) = \pm j Z_0^{(\pm)}(j) \mp Z_1^{(\pm)}(j) , \quad (87)$$

$$Z_\ell^{(+)}(j) = \int_{-\infty}^j du u^\ell \phi(u) , \quad (88)$$

$$Z_\ell^{(-)}(j) = \int_j^{+\infty} dx u^\ell \phi(u) , \quad \ell = 0, 1 . \quad (89)$$

These objects satisfy recursion relations:

$$\begin{aligned} \tilde{K}^{(\pm)}(\lambda; j+1) &= e^{\mp\lambda} \left[\tilde{K}^{(\pm)}(\lambda; j) \pm e^{\mp\lambda j} D_\lambda^{(\pm)}(j) \right] , \\ Z_\ell^{(\pm)}(j+1) &= Z_\ell^{(\pm)}(j) \pm C_\ell(j) , \quad \ell = 0, 1 , \end{aligned} \quad (90)$$

where

$$D_\lambda^{(\pm)}(j) = \int_j^{j+1} du e^{\pm\lambda u} \phi(u) , \quad (91)$$

$$C_\ell(j) = \int_j^{j+1} du u^\ell \phi(u) , \quad \ell = 0, 1 . \quad (92)$$

From Eq. (85 – 92), and the properties

$$\begin{aligned} \tilde{K}(\lambda; j) &= \tilde{K}(\lambda; -j) , & \tilde{K}^{(+)}(\lambda; 0) &= \tilde{K}^{(-)}(\lambda; 0) , \\ Z_1^{(+)}(0) &= Z_1^{(-)}(0) , & Z_0^{(+)}(0) &= Z_0^{(-)}(0) = \frac{1}{2} , \end{aligned} \quad (93)$$

the function $\tilde{K}(\lambda; j)$ can be calculated recursively for each $j \in \mathbb{N}$, by knowing $\tilde{K}^{(+)}(\lambda; 0)$ and $Z_1^{(+)}(0)$, then evaluating $D_\lambda^{(\pm)}(j)$ and $C_\ell(j)$ for each value of j . The integrals involved can be calculated with high accuracy with a simple higher-order polynomial quadrature. They are integral of smooth, well-defined quantities, since the interpolating scaling function goes to zero at their bounds. Moreover, for values of j lying outside the support of ϕ we can benefit of a functional relation for calculating the values of the kernel. The support of a m -th order scaling function goes from $-m$ to m , then we have $\forall p > 0$

$$\begin{aligned} K(\mu; m+p) &= e^{-\mu h p} K(\mu; m) , \quad \mu > 0 , \\ K(0; m+p) &= K(0; m) + p Z_0^{(+)}(m) . \end{aligned} \quad (94)$$

To summarize, we have found an efficient method for evaluating equation (83) for $j = 0, \dots, N_z$ and a fixed μ . Instead of calculating $N_z + 1$ integrals of range $2m$, we

can obtain the same result by calculating 2 integrals of range m and $4m$ integrals of range 1, with the help of relation (94). This will also increase accuracy, since the integrands are always smooth functions, which would not be the case with a naive approach.

The accuracy in calculating the integrals can be further improved by using the refinement relation (56) for interpolating scaling functions. For positive λ we have

$$\begin{aligned}
\tilde{K}(2\lambda; i) &= \int du e^{-2\lambda|u-i|} \phi(u) \\
&= \frac{1}{2} \int du e^{-\lambda|u-2i|} \phi(u/2) \\
&= \frac{1}{2} \sum_j h_j \int du e^{-\lambda|u-2i|} \phi(u-j) \\
&= \frac{1}{2} \sum_j h_j \tilde{K}(\lambda; 2i-j) .
\end{aligned} \tag{95}$$

This relation is useful to improve the accuracy in evaluating the kernel for high λ . Since in this case the exponential function is very sharp, it is better to calculate the kernel for lower λ and an enlarged domain and then apply relation (95) as many times as needed. The relation (94) allows us to enlarge the domain with no additional computational cost. With the help of the above described properties the computational time for evaluating the kernel in Fourier space can be considerably optimized, becoming roughly half of the time needed for its application on a real space density.

7.4 Numerical results and comparison with other methods

These Poisson solvers have a convergence rate of h^m , where m is the order of the interpolating scaling functions used to express the Poisson kernel. Since we use interpolating scaling functions of order 16 the convergence rate of the electrostatic potential is faster than the rate for the kinetic energy. All these Poisson Solvers have one thing in common, they perform explicitly the convolution of the density with the Green's functions of the Poisson's equation. The necessary convolutions are done by a traditional zero-padded FFT procedure which leads to an $\mathcal{O}(N \log N)$ operation count with respect to the number of grid points N . The accuracy of the potential is uniform over the whole volume and one can thus use the smallest possible volume compatible with the requirement that the tails of the wavefunctions have decayed to very small values at the surface of this volume. The fraction of the computational time needed for the solution of the Poisson's equation decreases with increasing system size and is roughly 1% for large systems, see Section 12. Moreover, the explicit Green's function treatment of the Poisson's solver allows us to treat isolated systems with a net charge directly without the insertion of compensating charges.

7.4.1 Free BC

For Free BC, we have compared our method with the plane wave methods by Hockney⁴⁰ and Martyna and Tuckerman⁴¹ as implemented in the CPMD electronic structure program.³⁸ As expected Hockney's method does not allow to attain high accuracy. The method by Martyna and Tuckerman has a rapid exponential convergence rate which is

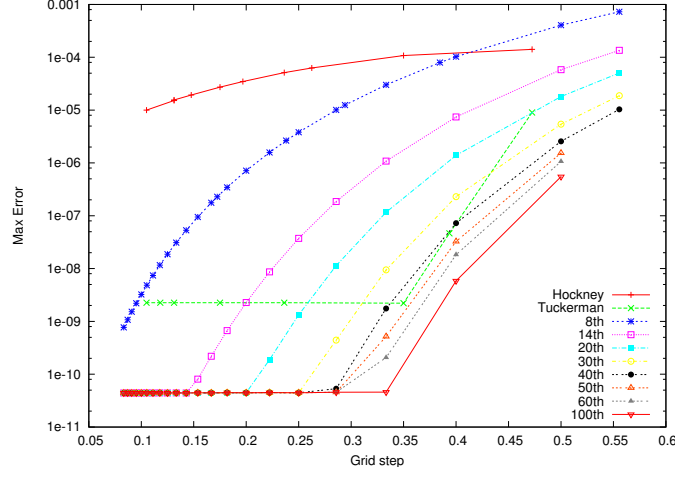


Figure 12. Accuracy comparison between our method with interpolating scaling functions of different orders and the Hockney or Martyna-Tuckerman method as implemented in CPMD. The accuracy of our method is finally limited by the accuracy of the expansion of Eq. (65) with 89 terms.

characteristic for plane wave methods. Our new method has an algebraic convergence rate of h^m with respect to the grid spacing h . By choosing very high order interpolating scaling functions we can get arbitrarily high convergence rates. Since convolutions are performed with FFT techniques the numerical effort does not increase as the order m is increased. The accuracy shown in Figure 12 for the Martyna and Tuckerman method is the accuracy in the central part of the cube that has 1/8 of the total volume of the computational cell. Outside this volume errors blow up. So the main disadvantage of this method is that a very large computational volume is needed in order to obtain accurate results in a sufficiently large target volume. For this reason the less accurate Hockney method is generally preferred in the CPMD program.³⁰

A strictly localized charge distribution, i.e. a charge distribution that is exactly zero outside a finite volume, can not be represented by a finite number of plane waves. This is an inherent contradiction in all the plane wave methods for the solution of Poisson's equation under free boundary conditions. For the test shown in Figure 12 we used a Gaussian charge distribution whose potential can be calculated analytically. The Gaussian was embedded in a computational cell that was so large that the tails of the Gaussian were cut off at an amplitude of less than $1.e-16$. A Gaussian can well be represented by a relatively small number of plane waves and so the above described problem is not important. For other localized charge distributions that are less smooth a finite Fourier representation is worse and leads to a spilling of the charge density out of the original localization volume. This will lead to inaccuracies in the potential.

Table 1 shows the required CPU time for a 128^3 problem as a function of the number of processors on a Cray parallel computer. The parallel version is based on a parallel 3-dimensional FFT.

1	2	4	8	16	32	64
.92	.55	.27	.16	.11	.08	.09

Table 1. The elapsed time in seconds required on a Cray XT3 (based on AMD Opteron processors) to solve Poisson’s equation on a 128^3 grid as a function of the number of processors. Since Poisson’s equation is typically solved many times, the time for setting up the kernel is not included. Including the set up time of the kernel increases the total timing by about 50 percent, since one additional FFT is needed.

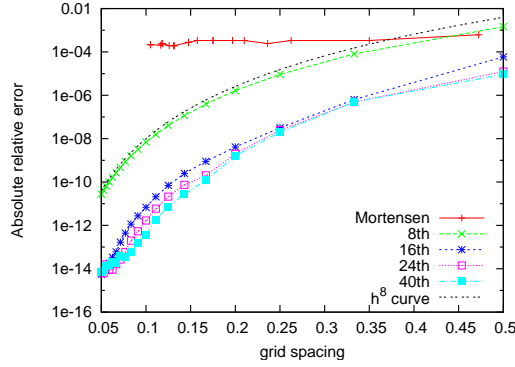


Figure 13. Accuracy comparison (Max. difference between the computed and the analytical result) between our method with scaling functions of different orders and the Mortensen solver for surface systems as implemented in CPMD. The results of the Hockney method are not shown since they are much less precise. The h^8 curve is plotted to show the algebraic decrease of the precision with respect to the grid space h . The accuracy is finally limited by the evaluation of the integral (84), which is computed with nearly machine precision.

7.4.2 Surfaces BC

Our method was compared with the reciprocal space methods by Hockney⁴⁶ and Mortensen⁴⁵ (which is a suitable generalization for 2D slab geometries of the method described in⁴¹) as implemented in the CPMD electronic structure program.³⁸

The accuracy tests shown in Figure 13 are performed with an analytical charge distribution that is the Laplacian of $V(x, y, z) = \exp(\cos(\frac{2\pi}{L_x}x) + \cos(\frac{2\pi}{L_y}y)) \exp(-\frac{z^2}{50L_z^2} - \tan(\frac{\pi}{L_z}z)^2)$. Its behavior along the xy surface is fully periodic, with all the reciprocal space components taken into account. The function $\exp(-\tan(\frac{\pi}{L_z}z)^2)$ guarantees a localized behavior in the non-periodic direction with the potential going explicitly to zero at the borders. This makes also this function suitable for comparison with reciprocal space based approach.

The Gaussian factor is added to suppress high frequency components. Tests with other analytical functions gave comparable accuracies. The reciprocal space Poisson solvers turn out to be much less precise than our approach, which explicitly preserves the BC along each direction. Moreover, the accuracy shown for the Mortensen approach is calculated only for planes that lies in the bulk of the non-periodic direction (30% of the total volume). Outside of this region, errors in the potential blow up.

Table 2 shows the behaviour of the errors in computing the Hartree energy following

the size of the system in the nonperiodic direction. To obtain the same accuracy of our approach with the Mortensen method we need a sytem which is roughly twice larger, which will imply that a very large computational volume is needed to obtain accurate results in a sufficiently large domain of the non-periodic direction.

L_0/L	0.5	0.6	0.7	0.8	0.9	1
$m=14$	$1 \cdot 10^{-12}$	$7 \cdot 10^{-12}$	$4 \cdot 10^{-6}$	$2 \cdot 10^{-3}$	$3 \cdot 10^{-2}$	$2 \cdot 10^{-1}$
Mortensen	0.2	1.3	3.7	6.0	6.8	6.2

Table 2. Evaluation error of the Hartree energy (Ha) for different values of the size L of the nonperiodic direction, for a system with an electrostatic density which is localized in the nonperiodic direction with characteristic length L_0 . The density of this system is identical to the one used for the accuracy tests of Figure 13, with $2L_0 = L_z$ (see text). The accuracy of the Mortensen approach with $L = 2L_0$ is of the same order of the accuracy obtained by our approach with $L = L_0$, which means that to obtain the same precision with Mortensen method the size of the system must be roughly doubled.

To show that our method genuinely preserves the boundary conditions appropriate for surfaces we calculated the electrostatic potential for a plane capacitor. For this system only the zero-th Fourier components in the plane are non-vanishing.

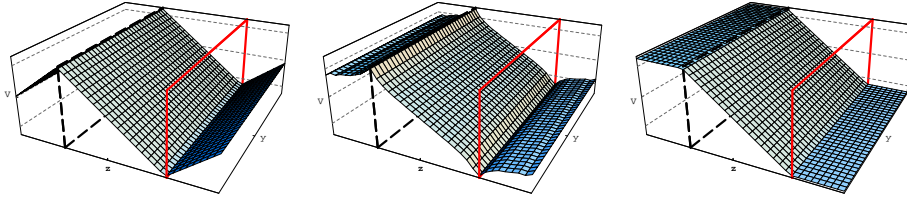


Figure 14. Electrostatic potential V for a system with two periodic planes charged with opposite sign (plane capacitor), oriented along the z direction, calculated by different Poisson solvers. The values of V are taken in the middle of the x (periodic) direction. The position of the positive (black, dashed line) and the negative (red, solid) charged plane is schematically shown in the figure. The represented solutions are, from top to bottom, the results from the Mortensen, the Hockney and our Poisson solver.

Figure 14 shows the results either in the Mortensen/Hockney reciprocal space methods or with our approach. For the plane capacitor, the screening function used in the Mortensen approach vanishes, and the solution is equal to what we would have obtained with a fully periodic boundary conditions. To obtain the good “zero electric field” behavior in the borders that we obtain directly with our method one would have to postprocess the solution obtained from the Mortensen method, by adding to the potential a suitable linear function along the non-periodic direction. This is legitimate since a linear function is annihilated by the Laplace operator and the modified potential is thus also a valid solution of the Poisson equation just with different boundary conditions. The Hockney method presents a better qualitative behavior, though the results are not accurate. Only with our approach we get both accurate and physically sound results.

Table 3 shows the required CPU time for solving the Poisson equation on a grid of 128^3 grid points as a function of the number of processors on a Cray XT3 parallel computer. The

1	2	4	8	16	32	64	128
.43	.26	.16	.10	.07	.05	.04	.03

Table 3. The elapsed time in seconds required on a Cray XT3 (based on AMD Optreron processors) to solve Poisson's equation with surface BC on a 128^3 grid as a function of the number of processors. The time for setting up the kernel (around 50% of the total time) is not included. For a large number of processors, the communication time needed to gather the complete potential to all the processors becomes dominant.

parallel version is based on a parallel 3-dimensional FFT, where the input/output is properly distributed/gathered to all the processors. The FFT's are performed using a modified version of the algorithm described in Ref.51 that gives high performances on a wide range of computers.

To summarize, we have presented a method that allows us to obtain accurate potentials arising from charge distributions on surfaces with a $O(N \log N)$ scaling in a mathematically clean way. This method preserves explicitly the required boundary conditions, and can easily be used for applications inside electronic structure codes where the charge density is either given in reciprocal or in real space. The potential applications of these Poisson solver are of great interest in the electronic strcuture calculations community.

7.5 Exact Exchange operator with ISF Poisson Solver

An example of the applications of the above described Poisson solvers may be found in the calculation of the Exact exchange operator of Eq. (22). One may write this operator in the following way:

$$E_x^{HF} = -\frac{1}{2} \sum_{\sigma=1,2} \sum_{p,q} \int d\mathbf{r} \rho_{p,q,\sigma}(\mathbf{r}) V_{p,q,\sigma}(\mathbf{r}) , \quad (96)$$

$$\rho_{p,q,\sigma}(\mathbf{r}) = \psi_{p,\sigma}(\mathbf{r}) \psi_{q,\sigma}^*(\mathbf{r}) , \quad (97)$$

$$V_{p,q,\sigma}(\mathbf{r}) = \int d\mathbf{r}' \frac{\rho_{q,p,\sigma}(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} , \quad (98)$$

$$D_x^{HF} |\psi_{p\sigma}\rangle = - \sum_q f_{q,\sigma} V_{q,p,\sigma} |\psi_{q,\sigma}\rangle . \quad (99)$$

The features of the Poisson Solver make the calculation of the objects above convenient for several reasons. First of all, the ISF Poisson solver technology implements the correct BC automatically, with optimal efficiency and accuracy. Moreover, the advantage of using a basis set which is independent from the atomic position make simpler the calculation of the atomic forces, since no Pulay terms have to be inserted as corrections. Different parallelization schemes of the Poisson solver application can be implemented, thanks to the flexibility of the real-space implementation of the ISF basis.

8 Calculation of Forces

Atomic forces can be calculated with the same method used for the application of the Hamiltonian onto a wavefunction. Since the scaling function/wavelet basis is not moving

together with atoms, we have no Pulay forces³¹ and atomic forces can be evaluated directly through the Feynman-Hellmann theorem. Except for the force arising from the trivial ion-ion interaction, which for the i -th atom is

$$\mathbf{F}_i^{(\text{ionic})} = \sum_{j \neq i} \frac{Z_i Z_j}{R_{ij}^3} (\mathbf{R}_i - \mathbf{R}_j) , \quad (100)$$

the energy terms which depend explicitly on the atom positions are related to the pseudopotentials. As shown in the previous sections, the GTH-HGH pseudopotentials we are using are based on separable functions,^{16,17} and can be splitted into a local and a non-local contribution.

For an atom i placed at position \mathbf{R}_i , the contribution to the energy that comes from the local part of the pseudopotential is

$$E_{\text{local}}(\mathbf{R}_i) = \int d\mathbf{r} \rho(\mathbf{r}) V_{\text{local}}(|\mathbf{r} - \mathbf{R}_i|) . \quad (101)$$

Where the local pseudopotential can be split into long and a short-ranged terms $V_{\text{local}}(\lambda) = V_L(\lambda) + V_S(\lambda)$, and

$$\begin{aligned} V_L(\lambda) &= -\frac{Z_i}{\lambda} \text{erf}\left(\frac{\lambda}{\sqrt{2}r_\ell}\right) , \\ V_S(\lambda) &= \exp\left(-\frac{\lambda^2}{2r_\ell^2}\right) \left[C_1 + C_2 \left(\frac{\lambda}{r_\ell}\right)^2 + \right. \\ &\quad \left. + C_3 \left(\frac{\lambda}{r_\ell}\right)^4 + C_4 \left(\frac{\lambda}{r_\ell}\right)^6 \right] , \end{aligned} \quad (102)$$

where the C_i and r_ℓ are the pseudopotential parameters, depending on the atom of atomic number Z_i under consideration. The energy contribution $E_{\text{local}}(\mathbf{R}_i)$ can be rewritten in an equivalent form. It is straightforward to verify that

$$E_{\text{local}}(\mathbf{R}_i) = \int d\mathbf{r} \rho_L(|\mathbf{r} - \mathbf{R}_i|) V_H(\mathbf{r}) + \int d\mathbf{r} \rho(\mathbf{r}) V_S(|\mathbf{r} - \mathbf{R}_i|), \quad (103)$$

where V_H is the Hartree potential, and ρ_L is such that

$$\nabla_{\mathbf{r}}^2 V_L(|\mathbf{r} - \mathbf{R}_i|) = -4\pi \rho_L(|\mathbf{r} - \mathbf{R}_i|) .$$

This analytical transformation remains also valid in our procedure for solving the discretized Poisson's equation. From equation (103) we can calculate

$$\rho_L(\lambda) = -\frac{1}{(2\pi)^{3/2}} \frac{Z_i}{r_\ell^3} e^{-\frac{\lambda^2}{2r_\ell^2}} , \quad (104)$$

which is a localized (thus short-ranged) function. The forces coming from the local pseudopotential are thus

$$\mathbf{F}_i^{(\text{local})} = -\frac{\partial E_{\ell}(\mathbf{R}_i)}{\partial \mathbf{R}_i} \quad (105)$$

$$= \frac{1}{r_\ell} \int d\mathbf{r} \frac{\mathbf{r} - \mathbf{R}_i}{|\mathbf{r} - \mathbf{R}_i|} \left[\rho'_L(|\mathbf{r} - \mathbf{R}_i|) V_H(\mathbf{r}) + V'_S(|\mathbf{r} - \mathbf{R}_i|) \rho(\mathbf{r}) \right] , \quad (106)$$

where

$$\begin{aligned}\rho'_L(\lambda) &= \frac{1}{(2\pi)^{3/2}} \frac{Z_{ion}}{r_{loc}^4} \lambda e^{-\frac{\lambda^2}{2r_\ell^2}}, \\ V'_S(\lambda) &= \frac{\lambda}{r_\ell} e^{-\frac{\lambda^2}{2r_\ell^2}} \left[(2C_2 - C_1) + (4C_3 - C_2) \left(\frac{\lambda}{r_\ell}\right)^2 + \right. \\ &\quad \left. + (6C_4 - C_3) \left(\frac{\lambda}{r_\ell}\right)^4 - C_4 \left(\frac{\lambda}{r_\ell}\right)^6 \right].\end{aligned}\quad (107)$$

Within this formulation, the contribution to the forces from the local part of pseudopotential is written in terms of integrals with localized functions (gaussian functions times polynomials) times the charge density and the Hartree potential. This allows us to perform the integrals only in a relatively small region around the atom position and to assign different integrations to different processors. Moreover, the calculation is performed with almost linear ($\mathcal{O}(N \log N)$) scaling.

The contribution to the energy that comes from the nonlocal part of the pseudopotential is, as we saw in Section 6.2,

$$E_{\text{nonlocal}}(\mathbf{R}_i) = \sum_l \sum_{mn} \langle \Psi | p_m^l(\mathbf{R}_i) \rangle h_{mn}^l \langle p_n^l(\mathbf{R}_i) | \Psi \rangle, \quad (108)$$

where we wrote explicitly the dependence of the projector on the atom position \mathbf{R}_i . The contribution of this term to the atomic forces is thus

$$\begin{aligned}\mathbf{F}_i^{(\text{nonlocal})} &= - \sum_l \sum_{m,n} \langle \Psi | \frac{\partial p(\mathbf{R}_i)}{\partial \mathbf{R}_i} \rangle h_{mn} \langle p(\mathbf{R}_i) | \Psi \rangle \\ &\quad - \sum \langle \Psi | p(\mathbf{R}_i) \rangle h_{mn} \langle \frac{\partial p(\mathbf{R}_i)}{\partial \mathbf{R}_i} | \Psi \rangle.\end{aligned}\quad (109)$$

Expressing the derivatives of the projectors in the Daubechies basis, the evaluation of the scalar products is straightforward. The scaling functions - wavelets expansion coefficients of the projector derivatives can be calculated with machine precision accuracy in the same way as the projectors themselves were calculated. This is due to the fact that the derivative of the projectors are like the projectors themselves products of gaussians and polynomials.

9 Preconditioning

As already mentioned, direct minimization of the total energy is used to find the converged wavefunctions. The gradient g_i of the total energy with respect to the i -th wavefunction $|\Psi_i\rangle$ is given by

$$|g_i\rangle = H|\Psi_i\rangle - \sum_j \Lambda_{ij} |\Psi_j\rangle, \quad (110)$$

where $\Lambda_{ij} = \langle \psi_j | H | \psi_i \rangle$ are the Lagrange multipliers enforcing the orthogonality constraints. Convergence is achieved when the average norm of the residue $\langle \overline{g_i | g_i} \rangle^{1/2}$ is below an user-defined numerical tolerance.

Given the gradient direction at each step, several algorithms can be used to improve convergence. In our method we use either preconditioned steepest-descent algorithm or

preconditioned DIIS method.^{29,30} These methods work very well to improve the convergence for non-zero gap systems if a good preconditioner is available.

The preconditioning gradient $|\tilde{g}_i\rangle$ which approximately points in the direction of the minimum is obtained by solving the linear system of equations obtained by discretizing the equation

$$\left(\frac{1}{2}\nabla^2 - \epsilon_i\right)\tilde{g}_i(\mathbf{r}) = g_i(\mathbf{r}) . \quad (111)$$

The values ϵ_i are approximate eigenvalues obtained by a subspace diagonalization in a minimal basis of atomic pseudopotential orbitals during the generation of the input guess. For isolated systems, the values of the ϵ_i for the occupied states are always negative, therefore the operator of Eq. (111) is positive definite.

Eq. (111) is solved by a preconditioned conjugate gradient (CG) method. The preconditioning is done by using the diagonal elements of the matrix representing the operator $\frac{1}{2}\nabla^2 - \epsilon_i$ in a scaling function-wavelet basis. In the initial step we use ℓ resolution levels of wavelets where ℓ is typically 4. To do this we have to enlarge the domain where the scaling function part of the gradient is defined to a grid that is a multiple of 2^ℓ . This means that the preconditioned gradient \tilde{g}_i will also exist in a domain that is larger than the domain of the wavefunction Ψ_i . Nevertheless this approach is useful since it allows us to obtain rapidly a preconditioned gradient that has the correct overall shape. In the following iterations of the conjugate gradient we use only one wavelet level in addition to the scaling functions for preconditioning. In this way we can do the preconditioning exactly in the domain of basis functions that are used to represent the wavefunctions (Eq. (39)). A typical number of CG iterations necessary to obtain a meaningful preconditioned gradient is 5.

10 Orthogonalization

We saw the need of keeping the wavefunctions Ψ_i orthonormal at each step of the minimization loop. This means that the overlap matrix S , with matrix elements

$$S_{ij} = \langle \Psi_j | \Psi_i \rangle \quad (112)$$

must be equal to the identity matrix.

All the orthogonalization algorithms have a cubic complexity causing this part of the program to dominate for large systems, see Figure 18. We therefore optimized this part carefully and found that a pseudo-Gram-Schmidt algorithm that uses a Cholesky factorization of the overlap matrix S is the most efficient method on parallel computers. In the following, we discuss the reasons for this choice by comparing it to two other orthogonalization algorithms: classical Gram-Schmidt and Loewdin orthogonalizations.

10.1 Gram-Schmidt orthogonalization

The classical Gram-Schmidt orthonormalization algorithm generates an orthogonal set of orbitals $\{|\bar{\Psi}_i\rangle\}$ out of a non-orthogonal set $\{|\Psi_i\rangle\}$, by processing separately each orbital. The overlap of the currently processed orbital $|\Psi_i\rangle$ with the set of the already processed

orbitals $\{|\bar{\Psi}_j\rangle\}_{j=1,\dots,i-1}$ is calculated and is removed from $|\Psi_i\rangle$. Thereafter, the transformed orbital $|\bar{\Psi}_i\rangle$ is normalized.

$$|\bar{\Psi}_i\rangle = |\Psi_i\rangle - \sum_{j=1}^{i-1} \langle \bar{\Psi}_j | \Psi_i \rangle |\bar{\Psi}_j\rangle \quad (113)$$

$$|\bar{\Psi}_j\rangle \rightarrow \frac{|\bar{\Psi}_j\rangle}{\sqrt{\langle \bar{\Psi}_j | \bar{\Psi}_j \rangle}} \quad (114)$$

The algorithm consists of the calculation of $n(n+1)/2$ scalar products and wavefunction updates. If the coefficients of each orbital are distributed among several processors $n(n+1)/2$ communication steps are needed to sum up the various contributions from each processor to each scalar product. Such a large number of communication steps leads to a large latency overhead on a parallel computer and therefore to poor performances.

10.2 Loewdin orthogonalization

The Loewdin orthonormalization algorithm is based on the following equation:

$$|\bar{\Psi}_i\rangle = \sum_j S_{ij}^{-\frac{1}{2}} |\Psi_j\rangle, \quad (115)$$

where a new set of orthonormal orbitals $|\bar{\Psi}_i\rangle$ is obtained by multiplying the inverse square-root of the overlap matrix S with the original orbital set.

The implementation of this algorithm requires that the symmetric overlap matrix S is calculated. In contrast to the classical Gram-Schmidt algorithm the matrix elements S_{ij} depend on the original set of orbitals and can be calculated in parallel in the case where each processor holds a certain subset of the coefficients of each wavefunction. At the end of this calculation a single communication step is needed to sum up the entire overlap matrix out of the contributions to each matrix element calculated by the different processors. Since S is an hermitian positive definite matrix, there exist a unitary matrix U which diagonalizes $S = U^* \Lambda U$, where Λ is a diagonal matrix with positive eigenvalues. The inverse square-root of S is then given by $S^{-\frac{1}{2}} = U^\dagger \Lambda^{-\frac{1}{2}} U$. Hence, an eigenvalue problem must be solved in order to find U and Λ .

10.3 Pseudo Gram-Schmidt using Cholesky Factorization

In this scheme a Cholesky factorization of the overlap matrix $S = LL^T$ is calculated. The new orthonormal orbitals are obtained by

$$|\bar{\Psi}_i\rangle = \sum_j (L_{ij}^{-1}) |\Psi_j\rangle, \quad (116)$$

and are equivalent to the orbitals obtained by the classical Gram-Schmidt. The procedure for calculating the overlap matrix out of the contributions calculated by each processor is identical to the Loewdin case. Instead of solving an eigenvalue problem we have however to calculate the decomposition of the overlap matrix. This can be done much faster. This algorithm also requires only one communication step on a parallel computer but has a lower pre-factor than the Loewdin scheme.

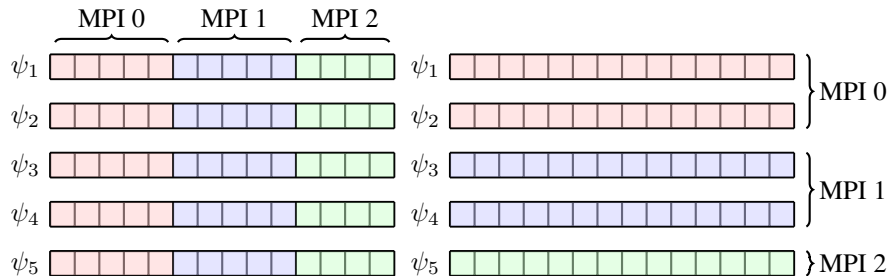


Figure 15. Coefficient distribution scheme on the left and orbital distribution scheme on the right

11 Parallelization

Two data distribution schemes are used in the parallel version of our program. In the orbital distribution scheme, each processor works on one or a few orbitals for which it holds all its scaling function and wavelet coefficients. In the coefficient distribution scheme (see Figure 15) each processor holds a certain subset of the coefficients of all the orbitals. Most of the operations such as applying the Hamiltonian on the orbitals, and the preconditioning is done in the orbital distribution scheme. This has the advantage that we do not have to parallelize these routines and we therefore achieve almost perfect parallel speedup. The calculation of the Lagrange multipliers that enforce the orthogonality constraints onto the gradient as well as the orthogonalization of the orbitals is done in the coefficient distribution scheme (Figure 15). For the orthogonalization we have to calculate the matrix $\langle \Psi_j | \Psi_i \rangle$ and for the Lagrange multipliers the matrix $\langle \Psi_j | H | \Psi_i \rangle$. So each matrix element is a scalar product and each processor is calculating the contribution to this scalar product from the coefficients it is holding. A global reduction sum is then used to sum the contributions to obtain the correct matrix. Such sums can easily be performed with the very well optimized BLAS-LAPACK libraries. Switch back and forth between the orbital distribution scheme and the coefficient distribution scheme is done by the MPI global transposition routine `MPI_ALLTOALL`. For parallel computers where the cross sectional bandwidth³⁴ scales well with the number of processors this global transposition does not require a lot of CPU time. The most time consuming communication is the global reduction sum required to obtain the total charge distribution from the partial charge distribution of the individual orbital.

11.1 OpenMP parallelization

In the parallelization scheme of the BigDFT code another level of parallelization was added via OpenMP directive. In particular, all the convolutions and the linear algebra part can be executed in multi-threaded mode. This adds further flexibility on the parallelization scheme. At present, several strategies are under analysis for systems with different sizes to understand the best repartition of the data between nodes such as to minimize the computational overhead.

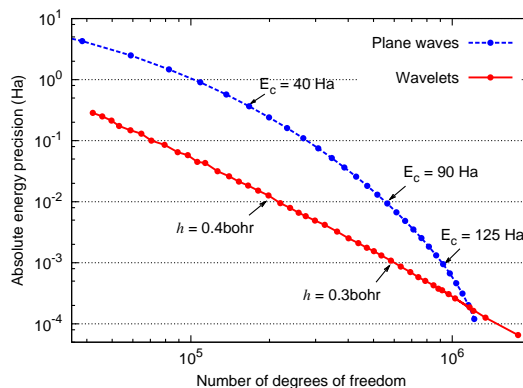


Figure 16. Absolute precision (not precision per atom) as a function of the number of degrees of freedom for a cinchonidine molecule (44 atoms). Our method is compared with a plane wave code. In the case of the plane wave code the plane wave cutoff and the volume of the computational box were chosen such as to obtain the required precision with the smallest number of degrees of freedom. In the case of our wavelet program the grid spacing h and the localization radii were optimized. For very high accuracies the exponential convergence rate of the plane waves beats the algebraic convergence rate of the wavelets. Such high accuracies are however not required in practice. Since convolutions can be executed at very high speed the wavelet code is faster than the plane wave code at any accuracy even if the number of degrees of freedom are similar (see Table 4).

12 Performance Results

We have applied our method on different molecular systems in order to test its performances. As expected, the localization of the basis set allows us to reduce considerably the number of degrees of freedom (i.e. the number of basis functions which must be used) to attain a given absolute precision with respect to a plane wave code. This fact reduces the memory requirements and the number of floating point operations. Figure 16 shows the comparison of the absolute precision in a calculation of a 44 atom molecule as a function of the number of degrees of freedom used for the calculation. In table 4 the comparison of the timings of a single SCF cycle with respect to two other plane wave based codes are shown. Since the system is relatively small the cubic terms do not dominate. For large systems of several hundred atoms the gain in CPU time compared to a plane wave program is proportional to the reduction in the number of degrees of freedom (compare Eq. (117)) and can thus be very significant as one can conclude from Figure 16.

The parallelization scheme of the code has been tested and has given the efficiency detailed in Figure 17. The overall efficiency is always higher than 88%, also for large systems with a big number of processors.

It is also interesting to see which is the computational share of the different sections of the code with respect to the total execution time. Figure 18 shows the percentage of the computational time for the different sections of the code as a function of the number of orbitals while keeping constant the number of orbitals per processor. The different sections considered are the application of the Hamiltonian (kinetic, local plus nonlocal potential), the construction of the density, the Poisson solver for creating the Hartree potential, the preconditioning-DIIS, and the operations needed for the orthogonality constraint as well as the orthogonalization, which are mainly matrix-matrix products or matrix decompositions.

E_c (Ha)	ABINIT (s)	CPMD (s)	Abs. Precision	Wavelets(s)
40	403	173	$3.7 \cdot 10^{-1}$	30
50	570	207	$1.6 \cdot 10^{-1}$	45
75	1123	422	$2.5 \cdot 10^{-2}$	94
90	1659	538	$9.3 \cdot 10^{-3}$	129
145	4109		$2 \cdot 10^{-4}$	474

Table 4. Computational time in seconds for a single minimization iteration for different runs of the cinchonidine molecule used for the plot in Figure 16. The timings for different cutoff energies E_c for the plane waves runs are shown. The input parameters for the wavelet runs are chosen such as to obtain the same absolute precision of the plane wave calculations. The plane wave runs are performed with the ABINIT code, which uses iterative diagonalization and with CPMD code³⁸ in direct minimization. These timings are taken from a serial run on a 2.4GHz AMD Opteron CPU.

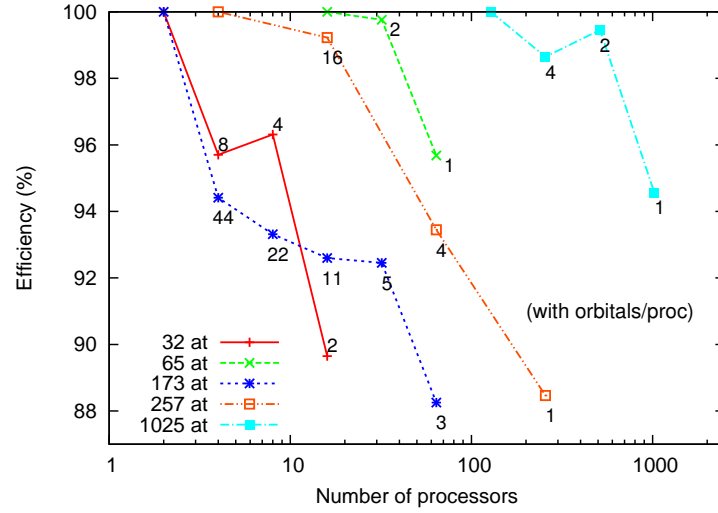


Figure 17. Efficiency of the parallel implementation of the code for several runs with different number of atoms. The number close to each point indicates the number of orbitals treated by each processors, in the orbital distribution scheme.

These operations are all performed by linear algebra subroutines provided by the LAPACK libraries.³⁹ Also, the percentage of the communication time is shown. While for relatively small systems the most time-dominating part of the code is related to the Poisson solver, for large systems the most expensive section is by far the calculation of the linear algebra operations. The operations performed in this section scales cubically with respect to the number of atoms. Apart from the Cholesky factorization, which has a scaling of $\mathcal{O}(n_{\text{orb}}^3)$, where n_{orb} is the number of orbitals, the cubic terms are of the form

$$\mathcal{O}(n \cdot n_{\text{orb}}^2), \quad (117)$$

where n is the number of degrees of freedom, i.e. the number of scaling function and wavelet expansion coefficients. Both the calculation of the overlap matrix in Eq. (112)

and the orthogonality transformation of the orbitals in Eq. (116) lead to this scaling. The number of the coefficients n is typically much larger than the number of orbitals.

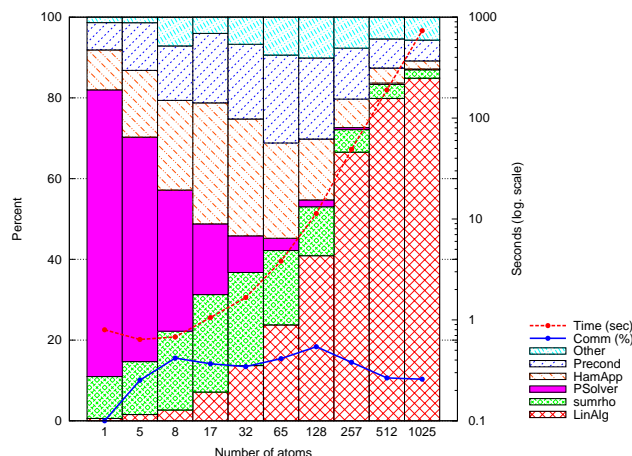


Figure 18. Relative importance of different code sections as a function of the number of atoms of a simple alkane chain, starting from single carbon atom. The calculation is performed in parallel such that each processor holds the same number of orbitals (two in this figure). Also the time in seconds for a single minimization iteration is indicated, showing the asymptotic cubic scaling of present implementation.

13 Conclusions

In this contribution we have shown the principal features of an electronic structure pseudopotential method based on Daubechies wavelets. Their properties make this basis set a powerful and promising tool for electronic structure calculations. The matrix elements, the kinetic energy and nonlocal pseudopotentials operators can be calculated analytically in this basis. The other operations are mainly based on convolutions with short-range filters, which can be highly optimized in order to obtain good computational performances. Our code shows high systematic convergence properties, very good performances and an excellent efficiency for parallel calculations. This code is integrated in the ABINIT software package and is freely available under GNU-GPL license. At present, several developments are in progress concerning mainly a linear scaling version and the possibility to calculate excited states.

References

1. J. Perdew, K. Burke and M. Ernzerhof, Phys. Rev. Lett **77**, 3865 (1996).
2. M. Dion, H. Rydberg, E. Schröder, D. C. Langreth, and B. I. Lundqvist, Phys. Rev. Lett. **92**, 246401 (2004).
3. I. Daubechies, “*Ten Lectures on Wavelets*”, SIAM, Philadelphia (1992).

4. X. Gonze, J.-M. Beuken, R. Caracas, F. Detraux, M. Fuchs, G.-M. Rignanese, L. Sindic, M. Verstraete, G. Zerah, F. Jollet, M. Torrent, A. Roy, M. Mikami, Ph. Ghosez, J.-Y. Raty, D.C. Allan. *Computational Materials Science* **25**, 478-492 (2002).
<http://www.abinit.org>
5. <http://bigdft.org>
6. Thomas L. Beck, *Rev. Mod. Phys.* **72**, 1041 (2000).
7. J. E. Pask, B. M. Klein, C. Y. Fong, and P. A. Sterne *Phys. Rev. B* **59**, 12352 (1999).
8. J. J. Mortensen, L. B. Hansen, and K. W. Jacobsen *Phys. Rev. B* **71**, 035109 (2005).
9. J. R. Chelikowsky, N. Troullier, Y. Saad, *Phys. Rev. Lett.* **72**, 1240 (1994).
10. Stefan Goedecker, *Rev. Mod. Phys.* **71**, 1085 (1999).
11. T. A. Arias, *Rev. Mod. Phys.* **71**, 267 (1999).
12. T. Yanai, G. I. Fann, Z. Gan, R. J. Harrison, and G. Beylkin, *J. Chem. Phys.* **121**, 6680 (2004).
13. M. Marques, Miguel, M. Oliveira, T. Burnus, *Computer Physics Communications* **182**, 2272 (2012).
14. S. Goedecker, "*Wavelets and their application for the solution of partial differential equations*", Presses Polytechniques Universitaires Romandes, Lausanne, Switzerland 1998, (ISBN 2-88074-398-2).
15. G. Beylkin, R. Coifman and V. Rokhlin, *Comm. Pure and Appl. Math.* **44**, 141 (1991).
16. S. Goedecker, M. Teter, J. Hutter, *Phys. Rev. B* **54**, 1703 (1996).
17. C. Hartwigsen, S. Goedecker and J. Hutter, *Phys. Rev. B* **58**, 3641 (1998).
18. M. Krack, *Theor. Chem. Acc.* **114**, 145 (2005).
19. M. Payne, M. Teter, D. Allan, T. Arias and J. Joannopoulos, *Rev. of Mod. Phys.* **64**, 1045 (1992).
20. G. Beylkin, *SIAM J. on Numerical Analysis* **6**, 1716 (1992).
21. J. Strang, G. J. Fix, *An analysis of the Finite Element Method*, Wellesley-Cambridge Press, 1988.
22. C. J. Tymczak and Xiao-Qian Wang, *Phys. Rev. Lett.* **78**, 3654 (1997).
23. Ka-Sing Lau and Qiyu Sun, *Proceedings of the American Mathematical Society*, **128**, 1087 (2000).
24. A. I. Neelov and S. Goedecker, *J. of Comp. Phys.* **217**, 312-339 (2006).
25. L. Genovese, T. Deutsch, A. Neelov, S. Goedecker, G. Beylkin, *J. Chem. Phys.* **125**, 074105 (2006).
26. L. Genovese, T. Deutsch, S. Goedecker, *J. Chem. Phys.* **127**, 054704 (2007).
27. J. A. White and D. M. Bird, *Phys. Rev. B* **50**, 4954 (1994).
28. B. R. Johnson, J. P. Modisette, P. J. Nordlander and J. L. Kinsey, *J. Chem. Phys.* **110**, 8309 (1999).
29. P. Pulay, *Chem. Phys. Lett.*, **73**, 393 (1980).
30. J. Hutter, H.P. Lüthi and M. Parrinello, *Comp. Mat. Sci.* **2** 244 (1994).
31. P. Pulay, in *Modern Theoretical Chemistry*, H. F. Schaefer editor, (Plenum Press, New York) (1977).
32. <http://physics.nist.gov/PhysRefData/DFTdata/Tables/ptable.html>
33. M. M. Morrell, R. G. Parr and M. Levy, *J. Chem Phys* **62**, 549, (1975).

34. S. Goedecker, A. Hoisie, “*Performance Optimization of Numerically Intensive Codes*”, SIAM publishing company, Philadelphia, USA 2001 (ISBN 0-89871-484-2).
35. E. R. Davidson, J. Comp. Phys. **17**, 87 (1975).
36. G. Kresse, J. Furthmüller, Phys. Rev. B **54**, 11169 (1996).
37. F. Bottin, S. Leroux, A. Knyazev, G. Zérah, Computational Materials Science **42**, 329 (2008).
38. CPMD Version 3.8: developed by J. Hutter, A. Alavi, T. Deutsch, M. Bernasconi, S. Goedecker, D. Marx, M. Tuckerman and M. Parrinello, Max-Planck-Institut für Festkörperforschung and IBM Zürich Research Laboratory (1995-1999).
39. E. Anderson *et al.*, “*LAPACK Users’ Guide*”, SIAM publishing company, Philadelphia, USA 1999 (ISBN 0-89871-447-8).
40. R. W. Hockney, The potential calculations and some applications, Methods Comput. Phys. **9** (1970) 135–210.
41. G. J. Martyna, M. E. Tuckerman, A reciprocal space based method for treating long range interactions in *ab initio* and force-field-based calculations in clusters, J. Chemical Physics **110** (6) (1999) 2810–2821.
42. L. Füsti-Molnar, P. Pulay, Accurate molecular integrals and energies using combined plane wave and gaussian basis sets in molecular electronic structure theory, J. Chem. Phys. **116** (18) (2002) 7795–7805.
43. Marie-Catherine Desjonquères, Daniel Spanjaard “*Concepts in Surface Physics*” Springer Series in Surface Sciences (1998).
44. Peter Minar, Mark E. Tuckerman, Katianna A. Pihakari, and Glenn J. Martyna, “*A new reciprocal space based treatment of long range interactions on surfaces*”, J. Chem. Phys. **116**, 5351 (2002).
45. J. J. Mortensen and M. Parrinello, “*A density functional theory study of a silica-supported zirconium monohydride catalyst for depolymerization of polyethylene*”, J. Phys. Chem. B **104**, 2901–2907.
46. R. W. Hockney and J. W. Eastwood, “*Computer Simulation Using Particles*” (McGraw-Hill, New York, 1981).
47. W. Hackbusch and U. Trottenberg, “*A Multigrid Method*”, Springer, Berlin, 1982.
48. N. Saito, G. Beylkin, G., Signal Processing, IEEE Transactions on [see also Acoustics, Speech, and Signal Processing, IEEE Transactions on] Vol. **41** (12), (1993) 3584–3590.
49. G. Deslauriers and S. Dubuc, Constr. Approx. **5**, 49 (1989).
50. G. Beylkin and L. Monzon, Applied and Computational Harmonic Analysis, **19** (2005) 17-48 ;
Algorithms for numerical analysis in high dimensions G. Beylkin and M. J. Mohlenkamp, SIAM J. Sci. Comput., **26** (6) (2005) 2133-2159;
G. Beylkin, M. J. Mohlenkamp, Numerical operator calculus in higher dimensions, in: Proceedings of the National Academy of Sciences, Vol. **99**, 2002, pp. 10246–10251.
51. S. Goedecker, Comput. Phys. Commun. **76**, 294 (1993).

Elmer Finite Element Solver for Multiphysics and Multiscale Problems

Mika Malinen and Peter Raback

CSC – IT Center for Science Ltd.

P.O. Box 405, FI-02101 Espoo, Finland

E-mail: {*Peter.Raback, Mika.Malinen*} @csc.fi

We give an overview of open source finite element software Elmer which has initially been developed by having especially multiphysics simulations in mind. We emphasize the role of the chosen modular design that enables the user to add new computational models and to couple the resulting new applications with the existing models easily. The key features of the general routines which enable performing such tasks as approximation with a wide collection of finite elements, mapping solution data between independent discretization meshes, parallel computations and the flexible utilization of effective linear algebra algorithms are also highlighted. The more general utility of the solver to even handle multiscale couplings is finally exemplified by considering the interaction of a particle-based model and a continuum-scale finite element discretization.

1 Introduction

The finite element method (FEM) has originally been applied to produce discrete solutions to partial differential equation (PDE) models which, in cases of classic applications, describe physical phenomena under the hypothesis of a continuous medium. Traditionally single-physics models have been applied, but nowadays approximate solutions based on such models are often found to be inadequate in providing predictive power desired. Therefore, the need to model the interaction of several physical phenomena simultaneously has greatly impacted the later developments of the finite element method.

Handling multiphysics couplings still brings significant challenges for both software developers and users of finite element software, but even a harder contemporary challenge relates to the need of modelling the interaction of effects which are associated with completely different time or length scales. Treating such multiscale problems has necessitated the development of new computational strategies which do not rely solely on the standard finite element approximation of PDE models. In this context, if suitable methodologies for transferring essential information from small scale to large scale are devised, the finite element method can usually be considered to be a viable method for handling the part of the simulation needed on the large scale, where PDE models based on the continuum description may often be applied.

The aim of this paper is to give an overview of open source finite element software Elmer¹ which has initially been developed by having especially multiphysics simulations in mind. Consequently, Elmer now offers a large selection of physical models which can be combined flexibly to produce computational models describing interactions of multiple physical phenomena. The basic models include equations which are, for example, related to fluid mechanics, structural mechanics, acoustics, electromagnetism, heat transport, and species transport.² In addition to utilizing these ready models, the user may also employ a

wide collection of numerical tools and algorithms contained in Elmer to create completely new models which can then be interfaced easily with the existing models. Although the existing models contained in Elmer are typically based on the continuum description, this flexibility also opens up possibilities of extending the capability of Elmer so that effects which originate from considering smaller-scale phenomena may additionally be taken into account.

It should be noted that the Elmer software package is divided into components in a traditional manner, so that separate software components for preprocessing and postprocessing are available. Here we shall however focus on the part which is customarily referred to as the solver in the finite element context. This part comprises a wide collection of routines for creating computational versions of PDE models and controlling the actual solution procedure, the heart of which is based on the application of efficient linear algebra algorithms contained in Elmer,³ or provided by external libraries. It is also noted that the solver of Elmer can be used independently. The work-flow relating to the usage of Elmer can thus be adapted such that other software may be used for preprocessing and postprocessing tasks, which typically relate to creating discretization meshes and the visualization of results, respectively. Basically the solver program of Elmer can be controlled simply by providing a special text file containing sets of commands.

The representation given in the remainder of the paper can be considered to be divided roughly into two parts. We shall begin by describing the overall modular design of the solver and the functionality of key routines that collectively give the user the power to adapt the solver for handling new coupled problems in a flexible manner. While the development of these routines has mainly been driven by applications relating to multiphysics simulations on the continuum scale, in the latter part of the paper we however consider certain nonstandard applications to exemplify even the more general utility of the solver. To this end, the multiscale interaction of a particle-based model and a continuum-scale finite element discretization is considered as an example.

2 Solving a Coupled Problem with the Solver of Elmer

The concept of multiphysics generally refers to handling problems where the solution consists of more than one component, with each component governed by its own physical law. In addition, an interaction between the constituent components of the whole solution occurs in some way. Typically this may happen as an effect of interaction terms which are defined on the bulk of the body where the constituent model is considered or on a lower dimensional interface, such as the boundary of the body.

Although Elmer is primarily regarded as multiphysics finite element software, this classification may actually be semantically too narrow to reflect the variety of problems to which the solver of Elmer have already been applied. We note that couplings which are characteristic of multiphysics problems occur similarly when treating a variety of other problems. In some applications similar couplings arise when the same physical model is treated via domain decomposition such that different discretization strategies are applied in the separate domains constituting the whole domain. As an example we mention the coupling of finite element and boundary element approximations. On the other hand, the need of reducing the computational cost often motivates the use of alternate strategies where mathematical models of different complicatedness level are used in the different parts of

the whole domain. For example describing the solution in thin boundary layers may necessitate taking into account physical processes that can however be neglected outside the boundary layer, so that a simpler mathematical model can be used there. This example brings us to considering general multiscale couplings where, in addition to having disparate scales, the associated distinctive feature is that more than one formulation is used to describe the same solution component.

That Elmer has been applied successfully to various cases covering all problem categories given above is better attributed to its ability to cope with model couplings than the strict concept of multiphysics ability. Although we therefore see opportunities for further extending Elmer's capabilities in multiscale simulation, we emphasize that the extent of general support for performing such simulations is currently much more limited. Moreover, there are also distinctive aspects between multiscale couplings and classic multiphysics couplings supported readily by Elmer. Performing a multiscale simulation unavoidably necessitates transforming information from small scale to large scale. Standard multiphysics couplings based on treating the bulk or surface coupling cannot thus be reused without addressing how to perform a fine-to-coarse transformation (also referred to as restriction). Similarly, a coarse-to-fine transformation (reconstruction or lifting) must also be handled. Furthermore, separate software may already exist for performing the simulation on the small scale, and a specific routine that enables and controls the interaction between separate software during the simulation has to be created.

After drawing our attention to the more general concept of a coupled problem, it is finally natural to mention here that the actual difficulty of solving a coupled problem depends heavily on the strength of the mutual interaction. Solving a loosely coupled problem (or weakly coupled problem in alternative terms) does not usually pose a major difficulty in terms of finding an effective iteration method for the problem. In this case standard segregation strategies such as applying a version of the Gauss-Seidel procedure can usually be utilized. On the other hand, considerable difficulties in the solver design may occur when a tightly coupled problem (or strongly coupled problem) has to be treated. Standard segregation strategies have then limited applicability and devising alternate solution methods that respect better the strong coupling of the constituent components arises as a typical necessity.

In the following we shall continue by describing some additional basic concepts relating to Elmer simulations along with representing the standard solution procedure which the solver of Elmer uses in order to handle the discrete version of a coupled problem.

2.1 Basic Concepts

The models handled by Elmer may generally be stationary or evolutionary, with nonlinearities possible in both the cases. Starting from a weak formulation of governing field equations, finite element approximation and advancing in time with implicit time integration methods are typically applied in order to obtain the computational version of the model. In the simplest case of single-physics models we are then lead to solving equations

$$F(u) = 0, \tag{1}$$

where u represents either the vector of coefficients in the finite element expansion of the stationary solution or the coefficient vector to describe the evolutionary finite element so-

lution at a given time $t = t_k$. Thus, in the case of evolution, the problems of the type (1) are solved repeatedly when advancing in time.

For linear models the problem (1) reduces to solving a linear system via defining

$$F(u) = b - Ku$$

where the coefficient matrix K is often referred to as the stiffness matrix and b corresponds to the right-hand side vector in the linear system. Otherwise F is a nonlinear mapping and an iteration is needed to handle the solution of the problem (1). In Elmer available nonlinear iteration methods generally depend on the model, as the definition of the linearization strategy is a part of the computational description of each physical model.

We note that many single-physics models offer the possibility of using the Newton iteration where the current nonlinear iterate $u^{(m)}$ to approximate u is updated at each iteration step as

$$\begin{aligned} DF(u^{(m)})[\delta^{(m)}] &= -F(u^{(m)}), \\ u^{(m+1)} &= u^{(m)} + \delta^{(m)}, \end{aligned} \tag{2}$$

where $DF(u^{(m)})$ is the derivative of F at $u^{(m)}$. Thus, performing the nonlinear solution update again entails the solution of the linear system at each iteration step. As an alternate to the Newton method, linearization strategies based on lagged-value approximations are also often available. In addition, relaxation is conventionally offered as a way to enable convergence in cases where the basic nonlinear iteration fails to produce convergence. Given the current nonlinear iterate $u^{(m)}$ and a computed correction $\delta u^{(m)}$ to the approximation, the new nonlinear iterate is then defined by

$$u^{(m+1)} = u^{(m)} + \lambda^{(m)} \delta^{(m)},$$

where $\lambda^{(m)}$ is an adjustable parameter referred to as the relaxation parameter.

2.2 Handling Multimodel Interactions

Having considered the basic concepts in the context of single-physics models, we now proceed to describe how the modularity employed in the design of Elmer allows us to create models which represent interactions of multiple (physical) phenomena. To this end, we assume that the complete model describes an interaction of N constituent models, the computational versions of which are primarily associated with the coefficient vectors u_i , $i = 1, 2, \dots, N$. As before, the coefficients contained in u_i are usually associated with the finite element expansion of either the stationary solution or the evolutionary solution at a time level $t = t_k$.

The fully discrete version of the coupled model leads to handling a problem of the form

$$\begin{aligned} F_1(u_1, u_2, \dots, u_N) &= 0, \\ F_2(u_1, u_2, \dots, u_N) &= 0, \\ &\dots \\ F_N(u_1, u_2, \dots, u_N) &= 0. \end{aligned} \tag{3}$$

If all the constituent models are linear, the problem (3) corresponds to solving a linear system where the coefficient matrix is a $N \times N$ block matrix. Otherwise (3) describes a

nonlinear problem. Although the solution of (3) could in principle be done in the same way as explained in the context of single-physics models in Section 2.1, i.e. by performing either a coupled linear solve or Newton iteration, the coupled problems are usually handled differently in order to enable the reuse of solvers for single-physics models and the easy extendibility of the code to handle new applications.

To this end, the nonlinear Gauss-Seidel iteration is usually applied, so that the coupling of the models is resolved via generating new coupled system iterates $u^{(j)} = (u_1^{(j)}, u_2^{(j)}, \dots, u_N^{(j)})$ as

$$\begin{aligned} F_1(u_1^{(j)}, u_2^{(j-1)}, u_3^{(j-1)}, \dots, u_N^{(j-1)}) &= 0, \\ F_2(u_1^{(j)}, u_2^{(j)}, u_3^{(j-1)}, \dots, u_N^{(j-1)}) &= 0, \\ &\dots \\ F_N(u_1^{(j)}, u_2^{(j)}, \dots, u_N^{(j)}) &= 0. \end{aligned} \tag{4}$$

It is noted that the k th discrete model description in (4) depends implicitly only on the coupled system iterate to its primary variable u_k , while the dependencies on the other constituent model variables are treated explicitly. This brings us to solving a nonlinear single-field problem

$$F(u_k^{(j)}) = F_k(v_1, \dots, v_{k-1}, u_k^{(j)}, v_{k+1}, \dots, v_N) = 0, \text{ with all } v_l \text{ given,} \tag{5}$$

which is handled by using the methods already described in Section 2.1. We also note that if all the constituent models are linear the nonlinear Gauss-Seidel iteration (4) reduces to the block Gauss-Seidel iteration for linear systems. Relaxation may again be applied as an attempt to improve the convergence behaviour of the basic iteration (4).

It is good to pause here to stress that the main advantage of the adopted nonlinear Gauss-Seidel scheme is its support for the modular software design. Also, it brings us to handling coupled problems via solving linear systems which are smaller than those which would result from treating all constraints in (3) simultaneously. Despite these merits, the suitability of the loosely coupled iteration (4) generally is case-dependent as convergence problems may occur in cases where a strong coupling is involved. Such problems are often best handled by methods which treat all the constituent models in (3) simultaneously. Certain physical models available in Elmer indeed employ this alternate tightly coupled solution strategy. However, these models have initially been developed independently, as common high-abstraction Elmer utilities for creating tightly coupled iteration methods in a general manner are less developed.

Sometimes the applicability of the nonlinear Gauss-Seidel scheme may be enhanced by appropriately modifying the set of equations to be used. This presents sometimes an intermediate alternative between the two approaches described above. An example for this case is the method of artificial compressibility that has been used to enable convergence in strongly coupled cases of fluid-structure interaction while still maintaining the benefits of the modular design.⁴

To summarize, the following pseudo-code presentation describes the basic loosely coupled iteration scheme employed by the solver of Elmer. This rough description may be helpful in summarizing what needs to be controlled overall to create a working computational solution procedure for a coupled problem.

```

! The time integration loop
for  $k = 1 : M$ 
  Generate an initial guess  $u^{(0)} = (u_1^{(0)}, u_2^{(0)}, \dots, u_N^{(0)})$  at  $t = t_k$ 
  ! The nonlinear Gauss-Seidel iteration
  for  $j = 1, 2, \dots$ 
    ! Generate the next coupled system iterate  $u^{(j)}$  by performing
    ! single-field updates
    for  $i = 1 : N$ 
      Set  $v_l = u_l^{(j)}$  for  $l = 1, 2, \dots, i - 1$ 
      Set  $v_l = u_l^{(j-1)}$  for  $l = i + 1 : N$ 
      Perform the nonlinear solve of  $F_i(v_1, \dots, v_{i-1}, u_i^{(j)}, v_{i+1}, \dots, v_N) = 0$ 
      Apply a relaxation to set  $u_i^{(j)} := u_i^{(j-1)} + \alpha_i(u_i^{(j)} - u_i^{(j-1)})$ 
    end
  end
end

```

Here the descriptions of the termination criteria for the iterations have been omitted. It is also noted that, obviously, the time integration loop is not needed in the case of a stationary problem. On the other hand, in the case of stationary simulation it is possible to replace the time integration loop by a pseudo-version of time stepping to enable performing multiple solutions for a range of model parameter values.

3 The Key Capabilities of the Solver

In the following, we shall focus on describing the key capabilities of the solver of Elmer which enable the user to create new models and to suit the solver for different purposes.

3.1 Extendibility by Modular Design

A module of the Elmer software which enables the creation of the discrete model description of the type (5) and its solution with respect to the primary variable is generally called a solver. The solvers of Elmer are truly modular in this manner and have a standard interface. Thus, each solver usually contains an implementation of the nonlinear iteration, instructions to assemble the corresponding linear systems from elementwise contributions, and standard subroutine invocations to set constraints and to actually solve the linear systems assembled.

It follows that enabling an interaction with another field, designated by v_l in (5), is simply a matter of solver-level implementation. Therefore, interactions which have not been implemented yet can be enabled by making modifications which are localized to the solvers. In addition, a completely new physical model may be added by introducing a new solver which comprises a separate software module and which can be developed independently with respect to the main program. As a result, applying the loosely coupled solution procedure to a coupled problem based on the new physical model again requires making only solver-level modifications.

3.2 Model-specific Finite Element Approximation and Mesh-to-mesh Mappings

In the most basic setting all constituent model variables u_i , $i = 1, \dots, N$, of a coupled problem are approximated by using the same piecewise polynomial basis functions defined over a single mesh. In addition to this, the solver of Elmer offers a built-in functionality to perform a coupled problem simulation by using solver-specific finite element meshes. The solver description is then augmented by the specification of the independent mesh which the solver uses. To make this functional in connection with the solution of coupled problems, Elmer has the capability of performing the solution data transfer, which is needed between the solvers in the loosely coupled solution procedure, even when the meshes are non-matching. The interpolation between the meshes is implemented using octree-based data structures which scale as $N \log N$ with the size of the problem but may still introduce communication bottle-necks in parallel cases. It must be understood, however, that the loss of high-resolution details is unavoidable when the high-resolution field is represented by using a coarser finite element mesh.

3.3 Approximation by Various Finite Element Formulations

Elmer has traditionally employed the Galerkin finite element approximation of weak formulations. A standard abstraction⁵ of linearized problems which arise from handling (5) can usually be given, so that a typical problem then is to find a finite element solution $U_h \in X_h$ such that

$$B(U_h, V_h) = L(V_h)$$

for any $V_h \in X_h$. In the most typical case the bilinear form $B : X \times X \rightarrow \mathbb{R}$ and the linear functional $L : X \rightarrow \mathbb{R}$ are well-defined when

$$X = H^1(\Omega),$$

where Ω denotes the body where the equation is posed and $H^1(\Omega)$ then contains square-integrable functions over Ω whose all first derivatives also are square-integrable. Traditionally the Lagrange interpolation basis functions defined for various element shapes have been used to obtain the finite dimensional set $X_h \subset X$. In this connection, the piecewise polynomial approximation of degree $1 \leq p \leq 3$ is possible for two-dimensional bodies, while three-dimensional bodies may be discretized by using the elements of degree $1 \leq p \leq 2$. The isoparametric mapping to describe curved element shapes is also supported with these traditional elements.

Discrete models based on more recent versions of the Galerkin finite element approximation are also possible. As an alternate to using the standard Lagrange interpolation basis functions, the Galerkin approximation based on using hierarchic high-degree piecewise polynomials can be employed. In this connection, the degree of polynomial approximation can also be defined elementwise, so that in effect the use of the hp -version of the finite element method is enabled. We note that Elmer provides an in-built mechanism to guarantee the continuity of such solutions. The H^1 -regularity of discrete solutions is thus ensured. However, generic ways to describe curved body surfaces accurately in connection with the high-degree finite elements have not been implemented yet which may limit the current utility of these elements. Anyhow, discretizations to capture localized solution details on

the interior of the body can generally be created without addressing the question of the geometry representation.

The way to define a high-degree approximation is based on the idea that a background mesh for representing the standard lowest-degree continuous finite element expansion is first provided, so that a specific element type definition in relation to elements present in the background mesh may then be given in order to enhance the approximation. The same idea has been adapted to create other alternate finite element formulations. For example, finite element formulations which enhance the approximation defined on the background mesh by a subscale approximation spanned by elementwise bubble basis functions can be obtained in this way. We note that this strategy is widely used in Elmer to stabilize otherwise unstable formulations and has also an interpretation in connection with the variational multiscale method. Another example of the use of the user-supplied element definition relates to creating approximations based on the discontinuous Galerkin method.

As a final example we mention that enhancing the approximation on the background mesh by introducing additional degrees of freedom associated with either faces or edges of elements and then omitting the original nodal degrees of freedom is also possible. This leads to a suitable set of unknowns for creating discretizations based on the face or edge element interpolation. If $L_2(\Omega)$ is used to denote the set of square-integrable scalar functions and $\Omega \subset \mathbb{R}^d$, we are led to bases for approximating vector fields in finite dimensional versions of the spaces

$$X = H(\text{div}, \Omega) = \{v \in L_2(\Omega)^d \mid \text{div } v \in L_2(\Omega)\}$$

or

$$X = H(\text{curl}, \Omega) = \{v \in L_2(\Omega)^d \mid \text{curl } v \in L_2(\Omega)^d\}.$$

A physical motivation for using these spaces is that fields with only either normal or tangential continuity on boundaries can then be approximated in a natural manner.

3.4 Monolithic Discretizations

In the case of a strongly coupled problem the standard segregated solution procedure of Elmer may become ineffective due to the need of using small values of relaxation parameters to avoid the divergence of the iteration. Using monolithic discretizations, i.e. handling all constituent components of the solution simultaneously, may then have a relative merit, at least in terms of obtaining robustness.

Monolithic discretizations may also be created by using Elmer as in principle it does not pose any restriction on how much physics may be included into a solver module definition. Some basic physical models, such as the system of compressible Navier–Stokes equations in flow mechanics, may indeed be thought of as intrinsically multiphysical and handling them in this way may actually appear as the most natural way. It should be noted, however, that high-level abstractions relating to implementing this alternate solution strategy does not necessarily exist and an additional burden is likely required in order to devise an effective solver for the resulting linear systems. We shall return to this issue when discussing linear algebra abilities below.

3.5 Discretizing in Time

For handling evolutionary cases the solver of Elmer offers implementations of many standard time-stepping algorithms applicable to either first-order or second-order problems. The usage of these time integration routines is done in a standardized manner so that only a few modifications must be done into an existing stationary solver to enable evolutionary simulations.

3.6 Linear Algebra Abilities

The ability to solve large linear systems efficiently is a central aspect of the simulation process with Elmer. As already explained, in the basic setting a linear solve is needed to obtain the solution update at each step of the nonlinear iteration. In practice linear solves are usually done iteratively, revealing one unexposed iteration level in relation to the pseudo-code presentation given in the end of Section 2.2.

The solver of Elmer offers a large selection of strategies to construct linear solvers. The majority of them are directly implemented into Elmer software, but interfaces to exploit external linear algebra libraries are also available. Typically the most critical decision in the use of linear solvers relates to identifying an effective preconditioning strategy for the linear system at hand. Traditionally Elmer has employed generic preconditioning strategies based on the fully algebraic approach. Highly efficient alternates to these standard preconditioners may also be obtained by using two-level iterations where the preconditioner is derived from applying multigrid methods.

If the linear system stems from a monolithic discretization of a coupled problem, the solution of the linear system by using the basic options may become a hindrance as the standard preconditioners may not be effective. A coupled linear system of this kind usually has a natural block structure associated and employing standard segregation strategies as preconditioners in the iterative solution then arises as a natural option. The utilities provided by Elmer have been used to generate sophisticated block preconditioned linear solvers for specific models,⁶ but recently attempts to encapsulate common abstraction of certain versions of these solvers have also been taken. We conclude that although the monolithic discretization may ultimately arise as the best way of handling some strongly coupled problems, the question of the efficient solution of the resulting linear systems often needs to be addressed simultaneously at the time of model implementation.

3.7 Parallel Computations

A strength of the solver of Elmer is that it supports the use of parallel computing by employing the message-passing library approach based on the Message Passing Interface (MPI) standard. This opportunity significantly widens the range of problem sizes which can be considered.

After a parallel version of Elmer solver has been made available in the parallel computer used, a principal task in enabling parallel computation is that domain decomposition is applied to partition the body description into the same number of parts as there are actual central processing units (they may be cores in the case of modern multi-core computer architectures) to be used in the simulation. In practice this is done prior to running the Elmer solver by partitioning the mesh files accordingly, so that each computing unit may

primarily work on its own data when the associated piece of computation does not necessitate communication between the units. The Elmer package offers preprocessing tools for performing the mesh partitioning, including the possibility of utilizing the external METIS library. Similarly postprocessing tools for uniting parallel simulation results which are output into separate files are provided.

3.8 Interfaces to Other Software and Libraries

The solver of Elmer employs basic linear algebra libraries, but it has also interfaces in order to utilize linear solvers of the HYPRE and UMFPACK packages. In addition, an option to use MUMPS linear solver exists, and an interface to utilize a set of Trilinos packages in the field of linear algebra has been created recently.

It should be noted that, on a more general level, compatibility with the other software components of the Elmer package which relate to the tasks of preprocessing and postprocessing is naturally provided. We also note that adapting the work-flow such that other software is used for these purposes is also possible. As an example we mention the usage of ParaView application in the visualization of results.

3.9 Obtaining Elmer Software

Elmer is actively developed, mainly by CSC – IT Center for Science Ltd. in Finland, and the newest version of the software maintained under Subversion version control system may be obtained via the project repository site where also Windows binaries are provided.⁷ The user is supplied with accessory configuration management aids and automated tests which help in compiling the software from the source codes and testing the executable programs compiled. The available documentation of the software is best accessed via the software's main site.¹ Additional references such as links to the discussion forum may also be found there.

4 Applying Elmer to Multiscale Problems

In the field of multiscale problems Elmer has not reached the same level of generality as for multiphysics problems. This is largely due to limited effort which has been put to study multiscale problems so far, but also due to the fact that multiscale problems offer a much wider spectrum of possible coupling scenarios as compared with couplings arising in the multiphysics problems. Of course, it makes sense to use Elmer in the solution of multiscale problems only if at least one of the scales is optimally addressed by the finite element method.

If both of the levels may be described by finite element approximations, then the basic features of Elmer may also help to study multiscale problems. For example, one can have nested meshes where the coarser mesh is utilized to obtain boundary conditions for the model employing the finer mesh. Then the results may be mapped automatically between the meshes.

A significantly more interesting scenario for multiscale simulations is a one where two different modelling approaches are combined. Then also the restriction and reconstruction operations become less trivial. Elmer includes a module for tracking particles in a mesh

and this provides a setting for performing related heterogeneous multiscale simulations with Elmer software. The rest of this section describes the particle utilities and how to combine them with the finite element utilities.

4.1 Following Particles in a Finite Element Mesh

The ability to combine the simulation of discrete particles with a finite element description opens the field for many new applications. Unfortunately, this comes with a price tag when we want to follow the particles in the finite element mesh. In the case of a finite difference grid it would be a trivial task to determine in which cell the particle lies in, as for each coordinate direction this requires one division by a grid size parameter followed by a rounding operation to obtain the index. For unstructured finite element meshes it is not as trivial. A generic search algorithm uses hierarchical octree-based data structure to perform the search. This leads to an algorithm with a complexity $N \log N$ and is difficult to parallel efficiently.

In the transient transport of particles the distance travelled by each particle within a timestep is typically more or less comparable to the size of the finite elements. If the timestep chosen is longer, then the approximation for the external field would be sub-optimal. Therefore it is expected that when we need to locate the particles in the mesh, they will be either in the same element as in the previous timestep or in some neighbouring element. This suggests using a search algorithm that utilizes the previous location of the particles.

In Elmer the particles may also be located in the finite element mesh using a marching routine. Then a vector is spanned between the previous and current position of the particle. If the particle lies in the same element as previously, the vector does not cross any element faces. If it does, the potential owner element is taken to be the other parent of the face element. This search is continued until the owner element is found for each particle.

The implementation of the algorithm is quite similar to that found in OpenFOAM. This algorithm is linear in complexity and will therefore outperform any octree-based search for sufficiently small timesteps. The downside of the algorithm is that it is fairly costly to determine the crossing between a vector and a face. As compared with the simple case of a regular finite difference grid the penalty in time is at least two orders of magnitude.

The information about in which element the particles are located may also be used to construct the list of closest neighbours for the particles. The possible distance between particles is then limited by the mesh size parameter and therefore this information is ideally used only for close-range particle-particle interaction.

4.2 Forces Acting on the Particle

We assume that our particles are classical ones, i.e. they follow the Newtonian dynamics. Consider a particle in position \vec{r} , with velocity \vec{v} and mass m . Newton's second law yields

$$m \frac{d\vec{v}}{dt} = \sum_j \vec{f}^j(\vec{r}, \vec{v}, \dots) \quad (6)$$

where a number of different forces \vec{f}^j may be considered. In Elmer we may consider forces due to gravity, electrostatic potential, magnetic field, viscous drag, buoyancy etc.

The particles may also be reflected from walls or scattered from the bulk. Also particle-particle collisions or more complex particle-particle interactions may be considered. Also periodic boundary conditions for the particles are provided.

The basic update sequence for the velocity and the position of the particle is

$$\vec{v}^{(i+1)} = \vec{v}^{(i)} + \frac{dt}{m} \sum_j \vec{f}^j, \quad (7)$$

$$\vec{r}^{(i+1)} = \vec{r}^{(i)} + dt \vec{v}^{(i+1)}. \quad (8)$$

Also higher-order schemes may be applied but the principle is nevertheless the same.

4.3 Reconstruction and Restriction Operators for the Particles

We need to devise reconstruction and restriction operators for the particles. Here the reconstruction operator takes finite element fields and generates the forces acting on the particles in their positions. Thus the reconstruction operators generally depend on interpolation routines used in the finite element method. As an example, consider the force \vec{f}^e which results from a macroscopic electrostatic field expressed in terms of a scalar potential ϕ , so that

$$\vec{f}^e = -q_i \nabla \phi, \quad (9)$$

where q_i is the electric charge of the particle. Then the action of the reconstruction operator basically corresponds to the evaluation of $\nabla \phi$ at the position \vec{r} of the particle via using a finite element approximation

$$\nabla \phi(\vec{r}) \approx \sum_{j=1}^n \phi_j \nabla \psi_j(\vec{r}), \quad (10)$$

where functions ψ_j are finite element shape functions.

Devising a restriction operator is more complicated. Let us consider a case where the particles give the right-hand side for a continuum equation. Such would be the case for an electrostatic equation having fixed charge density ρ as a source term, and also a contribution from the moving particle charges, so that

$$-\nabla \cdot \varepsilon \nabla \phi = \rho + \sum_i q_i \delta(\vec{r} - \vec{r}_i), \quad (11)$$

where ε is the permittivity and δ corresponds to the Dirac delta. The finite element version of (11) yields precisely pointwise contributions in a manner similar to (11) only when the particles are located exactly at the mesh nodes. Otherwise the values of the shape functions are seen to act as weighting factors in the approximation. For example, the contribution which the charge with the position vector \vec{r}_i makes to the j -entry of the right-hand side of the discrete system would be $q_i \psi_j(\vec{r}_i)$, with ψ_j the test function corresponding to the computation of the vector entry j .

The data resulting from a sample of particles always includes some statistical noise whereas the field solved with the finite element method is always deterministic. The noise may require some regularization that could be implemented in the finite element framework by using artificial diffusion. In the above example, the diffuse nature of the electrostatic equation takes care of the noise. However, if we would be modelling some material property, the finite element model would require a sufficient smoothness. Regularization makes

the multiscale coupled system more steady giving a hope for better convergence. Also for post-processing purposes smoothness may be desirable. Assuming a property p_i for each particle, a continuous approximation $p(\vec{r})$ could be solved from

$$-\nabla \cdot (D\nabla p) + \sum_i p(\vec{r}) \delta(\vec{r} - \vec{r}_i) = \sum_i p_i \delta(\vec{r} - \vec{r}_i) \quad (12)$$

where D is a diffusion coefficient used for the regularization of the statistical ensemble.

As an example, the self-consistent Poisson equation would now be solved from equations 6, 9 and 11. Some damping may be needed to keep the solution bounded. Additionally various scattering mechanisms may be added for the particles for a more comprehensive physical model.

5 Concluding Remarks

In this paper, Elmer finite element software for the simulation of multiphysical and multiscale problems has been presented. In multiphysical problems all the fields are usually optimally described by the finite element method. In Elmer multiphysics couplings may be treated in a generic manner by using the loosely coupled solution approach. In the case of multiscale problems multiple computational methods may also be needed. This makes it more challenging to devise a generic framework for handling the multiscale simulations.

The current multiscale implementation is limited to the interaction between continuous fields and discrete particles described by the classical equations of momentum. Even as such it opens the path for many interesting applications, such as carrier transport, microfluidics, and sedimentation. We note that the relative merit of the finite-element based machinery depends largely on the presence of non-trivial shapes. If the shape of the computational domain is a simple rectangular block, economical implementations of other computational methods are often possible. However, we believe that there are many multiscale problems where the current methodology could be quite useful.

References

1. <http://www.csc.fi/elmer>
2. P. R  back, M. Malinen, J. Ruokolainen, A. Pursula, T. Zwinger, Eds., Elmer Models Manual, CSC – IT Center for Science, 2013.
3. J. Ruokolainen, M. Malinen, P. R  back, T. Zwinger, A. Pursula, M. Byckling, Elmer-Solver Manual, CSC – IT Center for Science, 2013.
4. E. J  rvinen, P. R  back, M. Lyly, J.-P. Salenius, A method for partitioned fluid-structure interaction computation of flow in arteries, Med. Eng. & Phys. **30**, 917-923, 2008.
5. D. Braess, *Finite Elements. Theory, Fast Solvers, and Applications in Solid Mechanics*, 3rd Edition, Cambridge University Press, 2007.
6. M. Malinen, *Boundary conditions in the Schur complement preconditioning of dissipative acoustic equations*, SIAM J. Sci. Comput. **29**, 1567–1592, 2007.
7. <http://sourceforge.net/projects/elmerfem>

Modeling Charge Distributions and Dielectric Response Functions of Atomistic and Continuous Media

Martin H. Müser

Institute for Advanced Simulation
Forschungszentrum Jülich, 52425 Jülich, Germany
E-mail: m.mueser@fz-juelich.de

Many physical processes involve a significant redistribution of charge density, be it in a central system of interest or in a polarisable embedding medium providing boundary conditions. Examples range from protein folding in an aqueous solvent to the charge transfer between two unlike solids in relative motion. As modelers, we wish to have at our disposal efficient methods allowing us to describe the relevant changes, for example, to predict in what way charge redistribution affects interatomic forces. At small scales, calculations can be based on density-functional theory, while continuum electrostatics is appropriate for the description at large scales. However, neither of the two methods is well-suited when space is discretised into volume elements of atomic dimensions. At that scale, the most intuitive description is in terms of partial charges plus potentially electrostatic dipoles or higher-order atomic multipoles. Their proper assignment is crucial when dealing with chemically heterogeneous systems, however, it turns out to be non-trivial. Particularly challenging is a description of the charge transfer between atoms. In this chapter, we discuss attempts to describe such charge distribution in the framework of force fields assigning partial charges with so-called charge equilibration methods. This includes their motivation from the bottom-up, i.e., through density functional theory. In the top-down design, we investigate how to construct the microscopic model so that it reproduces the desired macroscopic response to external fields or to an excess charge. Lastly, we present avenues to extend the atom-scale models to non-equilibrium situations allowing one to model contact electrification or the discharge of a Galvanic cell.

1 Introduction

According to density functional theory (DFT),¹ the ground state energy of a system can be determined by minimizing an appropriate energy functional of the electron density for a given external field, which is usually defined by atomic or nuclear center-of-mass positions. As is the case with other field theories, one would like to have a recipe for a systematic coarse-graining so that large systems can be explored. However, coarse graining DFT becomes difficult when the linear size of the mesh (be it implicit or explicit) is no longer small compared to the Bohr radius. Only when the volume elements contain several hundred atoms does it become possible again to describe the charge distributions and the polarisation within a field-theoretical approach, i.e., with electrostatics of continua. Unlike DFT, the theory of electrostatics is based on material-specific parameters rather than on atom-specific parameters or natural constants. Since neither DFT nor continuum electrostatics work well with volume elements of atomic size, there is no seamless transition between them. As a consequence, no concurrent multi-scale methods exist linking DFT and continuum based descriptions of charge densities.

So-called charge equilibration (QE) methods,² also known as chemical potential equalisation methods,³ have been used for more than two decades to bridge the gap between electronic DFT and continuum electrostatics. The main idea of QE is to assign partial

atomic charges, plus potentially dipoles or higher-order multipoles, on the fly. However, QE approaches have traditionally suffered from various deficiencies. They include non-integer charges for molecular fragments⁴ and an excessive overestimation of the polarisability for non-metallic systems.⁵ These and related problems can be overcome by combining the ideas of QE with bond-polarisable models as is done in the so-called split-charge equilibration (SQE) method.⁶ In SQE, non-integer charge transfer between two atoms is penalised as a function of their distance and potentially their local environment. By introducing the concept of oxidation number, the method can also describe non-equilibrium processes by accounting for the history dependence of interatomic forces.⁷ For example, charges of individual atoms of a dissociated NaCl molecule can depend on the polarity of the solvent that was present during bond breaking. They are not merely a function of the instantaneous nuclear configuration.

In this chapter, we introduce central aspects of QE methods. We start by introducing the basic equations from phenomenological considerations in Sect. 2. A more rigorous, DFT-based motivation of the model is presented in Sect. 3. The relation between QE models and continuum electrostatics are derived in Sect. 4. An important aspect of that section is that we learn how to design QE models such that they reproduce macroscopic response functions. In this description, QE methods can be seen as the electrostatic analogue to bead-spring models mimicking the linear elasticity of molecules or solids. Thus, an individual QE degree of freedom is no longer constrained to represent a volume element of atomic size but can extend to much larger volumes. Sect. 5 contains some applications, including the simulation of contact electrification, which constitutes a central part in the modeling of complete Galvanic elements that are also discussed. Finally, we conclude in Sect. 6.

2 General Aspects of Charge-Equilibration Approaches

2.1 Motivation of the SQE model

In this section, we introduce the functional form of QE approaches using phenomenological arguments. A bottom-up and a top-down motivation of the expressions are given in the following two sections. The current presentation closely follows that given by the author in Ref. 8.

Often, polarisation in condensed matter systems is accounted for by placing inducible (point) dipoles onto atoms or (super) atoms.^{9–11} However, in addition to electrostatic polarisation of atoms, there can be charge transfer between them. Although there is no *unique* scheme breaking down the polarisation into intra- and inter-atomic contributions¹² (mainly because atomic charges cannot be defined unambiguously¹³), we present arguments in Sect. 3 why it is still both meaningful and practical to do so. For the moment being, let us simply assume the heuristic working hypothesis that charge transfer between atoms and the polarisation of atoms can be assigned meaningfully:

$$V(\{\mathbf{R}, Q, \boldsymbol{\mu}\}) = V(\{\mathbf{R}, Q_0, \boldsymbol{\mu}_0\}) + \sum_i \left\{ \frac{\partial V}{\partial Q_i} \Delta Q_i + \frac{\partial V}{\partial \mu_{i\alpha}} \Delta \mu_{i\alpha} \right\} \\ + \sum_{i,j} \left\{ \frac{1}{2} \frac{\partial^2 V}{\partial Q_i \partial Q_j} \Delta Q_i \Delta Q_j + \frac{\partial^2 V}{\partial Q_i \partial \mu_{j\alpha}} \Delta Q_i \Delta \mu_{j\alpha} + \frac{1}{2} \frac{\partial^2 V}{\partial \mu_{i\alpha} \partial \mu_{j\beta}} \Delta \mu_{i\alpha} \Delta \mu_{j\beta} \right\} \quad (1)$$

We truncate after second order and after the dipole terms. Here, $\{Q_0\}$ and $\{\mu_0\}$ denote, respectively, a set of reference values for atomic charges and dipoles. In the following, we will assume that these can be set to zero unless mentioned otherwise. Moreover, Roman indices refer to atom numbers while Greek indices enumerate Cartesian coordinates, e.g., $\mu_{i\alpha} \equiv \mu_{i\alpha 0} + \Delta\mu_{i\alpha}$ is the α component of the dipole on (super)atom i . For Cartesian indices, we use the summation convention.

Some terms in the Taylor expansion Eq. (1) are readily interpreted. $V(\{\mathbf{R}, Q_0, \mu_0\})$ represents a fixed-charge, non-polarisable potential – minus the explicitly mentioned electrostatic interactions. It can be a simple two-body or a more sophisticated many-body interaction model, such as a Tersoff or an embedded-atom potential. To interpret the Taylor-expansion related coefficients, it is best to consider isolated atoms: $\partial V/\partial Q_i$ corresponds to the electronegativity χ_i (plus potentially a coupling to an external electrostatic potential), while $\partial^2 V/\partial Q_i^2$ can be associated with the chemical hardness κ_i . They can be parameterised via finite-difference approximations of the ionisation energy I_i and electron affinity A_i . The latter two quantities can be obtained by removing or adding an elementary charge e from atom i ,

$$I_i = \frac{\kappa_i}{2}e^2 + \chi_i e \quad (2)$$

$$A_i = -\frac{\kappa_i}{2}e^2 + \chi_i e \quad (3)$$

and thus $\kappa_i = (I_i - A_i)/e^2$ and $\chi_i = (I_i + A_i)/2e$. (These quantities are commonly stated in units of eV, which means that the underlying unit system uses the elementary charge as the unit of charge.) In principle, κ_i and χ_i should depend on the environment, but within a reasonable approximation, they can be taken from values measured for isolated atoms. In practical applications, i.e., when allowing κ_i and χ_i to be free fit parameters, they turn out within $\mathcal{O}(10\%)$ of their experimentally determined values.^{6,14} Furthermore, it is tempting to associate the mixed derivative $\partial^2 V/\partial Q_i \partial Q_j$ ($i \neq j$) with the Coulomb potential, at least if \mathbf{R}_i and \mathbf{R}_j are sufficiently distant. For nearby atoms, one may want to screen the Coulomb interaction at short distance to account for orbital overlap.

All terms related to the atomic dipoles can be interpreted in a straightforward fashion. The negative of $\partial V/\partial \mu_{i\alpha}$ is the α component of the electrostatic field at \mathbf{R}_i due to external charges. The single-atom terms $\partial^2 V/\partial \mu_{i\alpha} \partial \mu_{i\beta}$, can be associated with the inverse polarisability $1/\gamma_i$ of atom i . Unlike for the charges, practical applications find a large dependence of the polarisability on the chemical environment (in particular for anions),¹⁵ including a direction dependence for directed bonds. The two-atom terms $\partial^2 V/\partial Q_i \partial \mu_{j\alpha}$ and $\partial^2 V/\partial \mu_{i\alpha} \partial \mu_{j\beta}$ correspond to the charge-dipole and dipole-dipole Coulomb interaction, respectively, at least for large distances R_{ij} between atoms i and j .

Unfortunately, it is incorrect to assume that the second-order derivatives $\partial^2 V/\partial Q_i \partial Q_j$ quickly approach the Coulomb interaction as R_{ij} increases beyond typical atomic spacings, which one might conclude from the argument that chemistry is local. This can be seen as follows: we know that isolated fragments (such as atoms or molecules) take integer charges, in many cases zero charge. If we separate two atoms, such as sodium and chlorine to large separation, we would find that the fragments carry a fractional charge

$$Q_{\text{Na,Cl}} = \pm \frac{\chi_{\text{Cl}} - \chi_{\text{Na}}}{\kappa_{\text{Na}} + \kappa_{\text{Cl}} - 1/(4\pi\epsilon_0 R_{\text{NaCl}})}, \quad (4)$$

assuming that $\partial V^2/\partial Q_i \partial Q_j$ quickly approaches the Coulomb potential. Using element-specific numerical values,¹⁶ one obtains partial charges of $\pm 0.4 e$ for a completely dissociated dimer. However, both atoms should be neutral, because $I_{\text{Na}} > A_{\text{Cl}}$. Thus, one needs to modify the model such that non-local, fractional charge transfer cannot occur.

What needs to be done is to penalise the transfer of (fractional) charge over long distances, i.e., when the overlap of orbitals of isolated atoms or ions ceases to be of importance. This can be done as follows. We write the charge of an atom as^{6,17}

$$Q_i = n_i e + \sum_j q_{ij}, \quad (5)$$

where n_i is called the oxidation state of the atom and q_{ij} is the charge donated from atom j to atom i , which is called the split charge. By definition, $q_{ij} = -q_{ji}$. (One may object that such an assignment is meaningless as electrons are indistinguishable. However, this is irrelevant, as one can see in Sect. 3) Next, we do not only penalise built-up of charge on atoms but also the transfer of charge. Thus, the terms in Eq. (1) exclusively related to atomic charges become

$$\begin{aligned} V(\{\mathbf{R}, Q, \dots\}) = & \sum_i \left\{ \frac{\kappa_i}{2} Q_i^2 + (\chi_i + \Phi_i^{\text{ext}}) Q_i \right\} \\ & + \sum_{i,j>i} \left\{ \frac{\kappa_{ij}}{2} q_{ij}^2 + \frac{S_{ij}(R_{ij})}{4\pi\epsilon_0 R_{ij}} Q_i Q_j \right\} + \mathcal{O}(\mu). \end{aligned} \quad (6)$$

Here, we have introduced the split-charge or bond hardness κ_{ij} , which is generally distance-dependent and also environment-dependent, i.e., it diverges as R_{ij} becomes large, prohibiting the transfer of charge over long distances. Moreover, $S_{ij}(R_{ij})$ denotes a screening at small distances with $S_{ij}(R_{ij}) \rightarrow 1$ for $R_{ij} \rightarrow \infty$.

Eq. (6) represents the SQE model. The original QE arises in the limit of vanishing bond hardness term κ_{ij} , while pure bond-polarisable models, such as the atom-atom charge transfer approach (AACT),¹⁸ neglect the atomic-hardness terms κ_i . Partial charges of atoms are deduced by minimizing the energy with respect to the split charges q_{ij} . The total charge of the system automatically adjusts to $Q_{\text{tot}} = \sum_i n_i e$ owing to the $q_{ij} = -q_{ji}$ symmetry. The minimisation of V with respect to the split charges can be done with the usual strategies for finding minima of second-order polynomials, such as steepest descent (good and easy for systems with large band gap, i.e., large values of κ_s , reasonable convergence in two or three iterations), extended Lagrangians (not efficient for systems with zero or small band gap), or conjugate gradient (probably best when dealing with small or zero band gap systems). Direct matrix inversion of the Hessian matrix is strongly advised against due to unfavorable scaling with particle number. Once the partial charges are determined, forces arising due to electrostatic interactions can be computed from $\partial V(\{\mathbf{R}, Q, \dots\})/\partial R_{i\alpha}$.

The numerical overhead of SQE versus QE is minimal, if present at all. As a matter of fact, since QE models all materials as metallic (as we shall see in Section 4), SQE requires much fewer iterations to convergence than QE, at least for systems with a band gap. However, there is a memory overhead within the SQE formulation. For example, assuming 12 neighbors per atom on average, one obtains six split charges per atom, which need to be stored in memory. Despite of this memory overhead in SQE, the number of floating-point operations per SQE minimisation step is not much larger than for QE. The reason is that the bulk of the calculations is related to the evaluation of the Coulomb potential V_C and

the derivatives $\partial V_C / \partial Q_i$. Once the latter are known and stored, the derivatives $\partial V_C / \partial q_{ij}$ can be obtained with little CPU time via

$$\frac{\partial V_C}{\partial q_{ij}} = \frac{\partial V_C}{\partial Q_i} - \frac{\partial V_C}{\partial Q_j}, \quad (7)$$

since $dQ_k / dq_{ij} = \delta_{ik} - \delta_{jk}$.

2.2 QE and redox reactions: Practical aspects

An important aspect of SQE is the possibility to change the (formal) oxidation state n_i of an atom by integer numbers. Since the n_i 's are discrete entities, their change implies a discontinuous alteration of the system. Increasing n_i corresponds to an oxidation, while reducing it reflects a reduction. In general, one would not increase (decrease) the oxidation number on a given atom unless one could decrease (increase) that of another, nearby atom by the same amount. Such a process can be interpreted as a redox reaction, or alternatively, as the transition from one Landau-Zener level to another one. In other words, the set $\{n\}$ indexes the Landau-Zener levels for a given atomic configuration $\{\mathbf{R}\}$. As discussed at the end of this Section and shown in Sect. 5 in more detail, having the option to make the system evolve on different Landau-Zener levels is what allows one to simulate systems in non-equilibrium. This enables one to incorporate history dependence, which is crucial, for example, for the simulation of Galvanic elements.

To illustrate the generic properties of SQE with respect to redox reactions, let us consider a simple model for the dissociation of a NaCl molecule, i.e., a model in which the short-range potential is a simple Lennard Jones (LJ) interaction and where the “free” parameters (LJ coefficients, chemical hardness of the constituents, etc.) are identical for atoms and ions. The SQE model would consider the molecules to be in the state Na^+Cl^- , i.e., $n_{\text{Na}} = -n_{\text{Cl}} = 1$, or in $\text{Na}^{(0)}\text{Cl}^{(0)}$, in which case $n_{\text{Na}} = n_{\text{Cl}} = 0$. For each choice of $\{n\}$, there is well-defined dependence for the energy and the partial charges on the interatomic distance, as depicted in Figure 1.

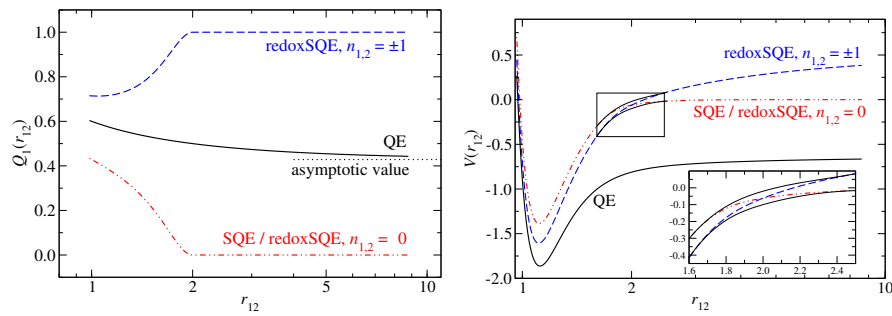


Figure 1. **(Left)** Charge Q_1 of a sodium atom in a simple model for a NaCl molecule and **(right)** total energy as a function of the interatomic distance r_{12} between Na and Cl. Blue lines refer to oxidation states $n_{1,2} = \pm 1$ while red lines indicate $n_{1,2} = 0$. Black lines represent the conventional QE approach. The inset emphasises the crossing of the two SQE energy curves together with the qualitative features of a full quantum-mechanical treatment. The latter would have to obey a non-crossing rule. (From Ref. 7)

The attempt to mimic charge transfer with SQE – be it a redox reaction between two molecules, charge hopping in a semi-conductor, or electron transfer between two metals – invokes three classes of problems, which are under current development: First, we need to parameterise the full model, i.e., all terms arising in Eq. (1), including the first term on the r.h.s., or Eq. (6) not only for atoms, but also for the relevant ions. Second, we need to identify meaningful rules for how we change the oxidation states. For example, the rate at which we attempt a “redox move” (oxidation of one atom and reduction of another nearby atom) should be chosen meaningfully and reflect, for example, that transition rates are small when the two involved atoms are distant but high upon close approach. Moreover, meaningful rules have to be designed for the acceptance criteria of a redox move. For example, a redox move could be subjected to a Metropolis algorithm or only be allowed when energetically favourable. In the latter case, one might want to make the reaction energy-conserving by increasing the relative velocity of the two involved atoms in an appropriate way. Third, we need to identify appropriate compromises between accuracy and efficiency. In principle, each redox move necessitates the re-optimisation of all split charges in the system. However, if we had to solve for each split charge in a large system when attempting locally one redox reaction, the computing time would increase non-linearly with system size N , e.g., with N^3 , depending on the solver for the split charges. In practice, only those split-charges will be affected significantly that lie in the immediate vicinity of the redox center. Thus, if Coulomb interactions are added up in an efficient way, it should be possible to devise approaches scaling linearly with system size N , or with $N \ln N$.

Some intricacies related to the determination of potentials and the identification of rules for the transition between different Landau-Zener levels can be explored in the context of the dissociation of our NaCl molecule in Figure 1. A quantum-chemical calculation can produce reference values for the ground state and the first excited state, even if it proofs difficult in practice to ensure the correct asymptotic, integer-valued charges on the dissociation products. Quantum chemists tend to claim that this is an easy exercise yet may fail to deliver good data even if they have several publications on precisely that topic. Nonetheless, having good reference data at hand, one can design the interaction parameters such that they reproduce the curves except near the transition state, i.e., at the point where the two SQE energy curves cross in violation of the non-crossing rule. In practice, the gap at the transition state tends to be very small, that is, less than the thermal energy at room temperature and thus systematic errors should remain small. However, to keep errors small, it is important that dynamics are designed such that the molecule dissociates correctly, e.g., into two neutral atoms in the case of slow dynamics in a chemically inert, non-polarisable environment. In contrast, when dynamics are fast and/or the dissociation takes place in a sufficiently polar solvent, two charged atoms should result after dissociation. The precise interatomic distance at which the redox reaction takes place is not very likely to affect the products significantly. For example, an NaCl molecule should dissociate into two neutral fragments when the embedding medium is an inert N_2 gas, while it should dissociate into two charged atoms in the presence of water. Once the Na and Cl atoms are sufficiently far apart, they can no longer exchange charge on typical MD time scales, and the transfer of integer charge between the two atoms should no longer be possible. Conventional force fields or even conventional DFT calculations cannot account for such history effects, because forces are unique functions of atomic coordinates and total charge.

3 Bottom-Up Motivation of Charge-Equilibration Models

3.1 Justification of the SQE expansion

A possibility to formally justify polarisable force fields from DFT or other ab-initio methods can be described as follows:^{19–22} In a first-principle calculation, we can constrain the electronic charge density $\rho(\mathbf{r})$ to produce a given set of electric monopoles assigned to individual atoms, $\{Q\}$, and higher-order multipoles $\{\mu, \dots\}$. For any such constraint, one can compute the minimum energy E_0 (be it in the framework of DFT or some other method) and thereby construct a constrained ground state energy function $E_0(\{Q, \mu, \dots\})$. The functional dependence of E_0 on its variables depends itself on how we translate charge density into mono- and multipoles, e.g., on the weight functions, $w_A(\mathbf{r})$, specifying how atoms “own” the three-dimensional space. Once these functions are known, we can compute, for example, the charge on atom A according to

$$Q_A = Z_A e - \int d^3r w_A(\mathbf{r}) \rho(\mathbf{r}), \quad (8)$$

where Z_A is the number of protons in atom A. Defining the weights $w_A(\mathbf{r})$ is an important topic in theoretical chemistry, which we do not want to discuss in more detail here. However, each assignment scheme contains the true ground state as the minimum of $E_0(\{Q, \mu, \dots\})$. For a given assignment scheme, one can then expand $E_0(\{Q, \mu, \dots\})$ around a set $\{Q_0, \mu_0, \dots\}$. The latter can reflect isolated atoms ($Q = \mu = 0$) or those associated with the ground state or a representative state of the system of interest.³

The fundamental philosophy of polarisable force fields is to expand $E_0(\{Q, \mu, \dots\})$ into powers of the leading-order multipoles and to identify accurate approximations for the respective expansion coefficients. This leads to Eq. (1). The underlying assumptions are: (a) the expansion is unique, (b) convergence is fast so that third and higher-order powers can be neglected, (c) only the leading-order multipoles, in most cases the monopoles, need to be considered, and (d) the expansion coefficients can be approximated by simple elementary functions of the atomic coordinates. For example, the “coefficient” $E_0(\{Q_0, \mu_0, \dots\})$, which corresponds to $V(\{Q_0, \mu_0\})$ in Eq. (1), necessitates descriptions of short-range repulsion, covalent effects, etc. Each of the assumptions (a)–(d) deserves particular attention. (a) Since there is no unique way to assign electron density to partial charges, the function $E_0(\{Q_0, \dots\})$ is not unique. However, it should be unique once we have decided on how to divide up electronic density to individual atoms. One could argue that good weight-function schemes lead to fast convergence. (b) Truncation after second order is not always appropriate. For example, solid hydrogen becomes infrared active under high pressure before the H_2 molecules dissociate. Such a spontaneous symmetry breaking can only be cast into a higher-order expansion. (c) Higher-order multipoles contain a lot of information on the hybridisation of atoms and thus on their bonding. One might hope that many-body potentials can reflect the pertinent effects if one does not solve explicitly for quadrupoles, octupoles, etc. (d) Instead of fitting free parameters of elementary functions, one can also envision machine-learning strategies in order to minimise human bias in the construction of polarisable force fields.

The main difficulty arising in the parameterisation of Eq. (1) is that coefficients (in addition to those related to long-range Coulomb interactions) do not disappear sufficiently quickly, i.e., much more slowly than the overlap of two atoms. We abstain from providing a

proof but simply point out the observation that isolated molecular fragments carry integer charge. Much of the non-locality is due to the kinetic energy. Problems arising due to the non-locality are alleviated in Lieb’s formulation of DFT²³ in which the Hohenberg-Kohn functional is replaced by the Legendre transform of the energy. Before elucidating this point further, we first demonstrate that SQE is in fact non-local in the charges despite being local in the split charges.

3.2 Locality of SQE formulations and their relation to DFT

In this section, we rewrite the SQE model following a recent work by Verstraelen *et al.*²² The only term of interest here is the one containing the bond hardnesses, i.e., the term $V_{\text{SQ}} = \sum_{i,j>i} \kappa_{ij} q_{ij}^2 / 2$ where we assume that κ_{ij} is local or at most semi local, i.e., it disappears if the orbitals of i and j do not overlap. Formally, one can write the charge on atom Q_i as $Q_i = \sum_j T_{ij} q_{ij}$, where T_{ij} is the connectivity matrix. As atoms may be bonded to more than one neighbor, the connectivity matrix is not a square matrix and can therefore not be inverted. Let us assume, however, that a so-called Moore-Penrose pseudoinverse, T_{MPP}^{-1} exists even if we do not know how to construct it. This matrix can be and generally will be non-local since only its inverse is local. We can then write

$$V_{\text{SQ}} = \sum_{i,j>i} \frac{K_{ij}}{2} Q_i Q_j \quad (9)$$

with

$$K_{ij} = \sum_{i'j'} [T_{\text{MPP}}^{-1}]_{ii'} \kappa_{i'j'} [T_{\text{MPP}}^{-1}]_{j'j}. \quad (10)$$

Then K_{ij} is not generally a sparse matrix, and Eq. (9) will be expensive to evaluate numerically. However, by taking the Legendre transform and introducing Lagrange multipliers μ_i for the charges, one can express V_{SQ} as

$$V_{\text{SQ}} = \max_{\{\mu\}} \sum_{i,j} Q_i \mu_j \delta_{ij} - \frac{1}{2} \mu_i S_{ij} \mu_j \quad (11)$$

where the coefficients of the softness matrix²⁴ S is given by

$$S_{ij} = \sum_{i'j'} T_{\text{MPP}ii'} \frac{1}{\kappa_{i'j'}} T_{\text{MPP}j'j} \quad (12)$$

resulting in a sparse matrix given that $T_{\text{MPP}ij}$ and κ_{ij} are both sparse.

Interestingly, the same functional form for the ground state energy is obtained if Lieb’s formulation of the kinetic energy in DFT is “condensed” to partial charges. In his case, the Lagrange multipliers correspond to the external potential producing the same (constrained ground-state) density $\rho(\mathbf{r})$ for non-interacting electrons as in the full problem. We refer to the original literature for more details²² and content ourselves by saying that (after making some approximations) the approach allows one to deduce the softness matrix directly from first-principle calculations. The advantage is that this allows one to deduce split charge stiffnesses directly from DFT calculations.²² However, given that current and popular DFT functionals do not describe the polarizability of large molecules correctly, it remains to be seen how useful these formal insights are in practice.

4 Top-Down Approach to Charge-Equilibration Models

In this section, we explore what dielectric response functions the SQE model produces. The analysis includes the limiting cases of pure atom-polarisable approaches, i.e., the original QE, or pure bond-polarisable models, such as the AACT model).¹⁸ From such calculations we can learn how each term in the microscopic model affects the macroscopic response. This, in turn, can guide the development and the parameterisation of the microscopic model.

This section is divided into the analysis of the response to an external field and to an excess charge that is added to the system. In the current treatment, we neglect any atomic dipoles for reasons of simplicity. Including them changes the numerical values, for say, the wave vector dependence of the dielectric permittivity or the work function of a solid, but it does not affect the leading-order scaling. A comparison between modeling dielectric media in terms of a simple SQE approach [no atomic dipoles, (super)-atoms placed onto a simple cubic lattice] or in terms of a pure dipole model (as often assumed in text books for the derivation of the Clausius-Mossotti relation) is presented elsewhere.⁸ It shall not be repeated here. We also assume simple cubic lattices in the treatment of solids. Moreover, we employ periodic boundary conditions to eliminate surface effects and to facilitate the analysis in Fourier space.

4.1 Response to an external field

Eq. (6) is readily transformed into reciprocal space, because it is a second-order polynomial in the (split) charges. Thus, one merely needs to replace sums over \mathbf{R} with sums over wave vectors \mathbf{k} and follow the known rules for Fourier transforms. This leads to:

$$V = N \sum_{\mathbf{k}} \left[\frac{\kappa + \tilde{J}(\mathbf{k})}{2} \tilde{Q}^2(\mathbf{k}) + \{\chi + \tilde{\Phi}^{\text{ext}}(\mathbf{k})\} \tilde{Q}(\mathbf{k}) + \frac{\tilde{\kappa}_{\Delta\mathbf{R}}(\mathbf{k})}{2} \tilde{q}_{\Delta\mathbf{R}}^2(\mathbf{k}) \right], \quad (13)$$

where

$$\tilde{J}(\mathbf{k}) = \frac{1}{4\pi\epsilon_0} \sum_{\Delta\mathbf{R} \neq 0} \frac{S(\Delta\mathbf{R})}{R} \exp(-i\mathbf{k} \cdot \Delta\mathbf{R}) \quad (14)$$

represents the potentially screened Coulomb coupling of atoms in Fourier space. For unscreened interactions, the following approximation can be made for simple cubic lattices:⁵

$$\tilde{J}(\mathbf{k}) \approx \frac{1}{\epsilon_0 a} \frac{1}{(ka)^2} [1 - \alpha(ka)^2 + \beta \{(ka)^4 + K_4\}], \quad (15)$$

where a is the lattice constant, K_4 a fourth-order cumulant

$$K_4 = -(ka)^4 + \frac{3}{2} \sum_{\alpha=1}^3 (k_{\alpha}a)^4, \quad (16)$$

and $\alpha = 0.22578(1)$ while $\beta = 0.0037(1)$. The numerical values for the last two constants differ for other lattices and are affected by short-range screening corrections. However, the leading-order term of the Coulomb interaction is universal.

Eq. (13) formally allows for the possibility to have split-charges live not only between nearest neighbors. In other words, split charges q_{ij} can exist in addition to those for which

$\Delta \mathbf{R}_{ij} = a \mathbf{n}_\alpha$, where \mathbf{n}_α is one of the three unit vectors of the simple cubic lattice. Having these additional split charges adds flexibility when using SQE to model dielectric media with non-monotonic dielectric permittivity $\epsilon_r(\mathbf{k})$. However, until further notice we assume that split-charges only live between nearest neighbors, which simplifies the analytical treatment.

Using the integer triple (l, m, n) to index the simple-cubic lattice site

$$\mathbf{R}_{lmn} = a(l\mathbf{n}_x + m\mathbf{n}_y + n\mathbf{n}_z), \quad (17)$$

one can write the charge on that lattice site (assuming oxidation numbers are zero in a mono-atomic lattice)

$$Q_{lmn} = q_{lmn}^{(1)} - q_{(l-1)mn}^{(1)} + q_{lmn}^{(2)} - q_{l(m-1)n}^{(2)} + q_{lmn}^{(3)} - q_{lm(n-1)}^{(3)}, \quad (18)$$

where the (split) charge donated from lattice site $(l+1)mn$ to lmn is denoted as $q_{lmn}^{(1)}$, etc. This notation allows us to find the following continuum approximation to Eq. (18)

$$Q(\mathbf{R}) \approx a \partial_\alpha q_\alpha(\mathbf{R}), \quad (19)$$

or

$$\tilde{Q}(\mathbf{k}) \approx ia \sum_\alpha k_\alpha \tilde{q}_\alpha(\mathbf{k}), \quad (20)$$

for which we assume that the split charges are smooth functions in \mathbf{R} . However, when \mathbf{k} is not close to the center of the Brillouin zone, one should use

$$\tilde{Q}(\mathbf{k}) = \sum_\alpha 2 \sin\left(\frac{k_\alpha a}{2}\right) \tilde{q}_\alpha(\mathbf{k}) \quad (21)$$

instead of Eq. (19).

We can now insert Eq. (21) into Eq. (13) and minimise V with respect to the $\tilde{q}_\alpha(\mathbf{k})$ to yield

$$\tilde{E}_\alpha^{\text{ext}}(\mathbf{k}) = \{J_{\alpha\beta}^{\text{eff}} + \kappa_s\} \tilde{q}_\beta(\mathbf{k}), \quad (22)$$

κ_s being the nearest-neighbor split charge stiffness, and

$$\tilde{J}_{\alpha\beta}^{\text{eff}}(\mathbf{k}) = 4 \left\{ \kappa + \tilde{J}(\mathbf{k}) \right\} \sin\left(\frac{k_\alpha a}{2}\right) \sin\left(\frac{k_\beta a}{2}\right), \quad (23)$$

where $\tilde{J}(\mathbf{k})$ can be taken from Eq. (15).

We are now in a position to solve for the split charges, and thus for the polarisation \mathbf{P} , which in turn allows us to deduce the dielectric permittivity via

$$\tilde{\mathbf{P}}(\mathbf{k}) = \{\tilde{\epsilon}_r(\mathbf{k}) - 1\} \tilde{\mathbf{E}}^{\text{tot}}(\mathbf{k}). \quad (24)$$

Since $\mathbf{E}^{\text{tot}} = \mathbf{E}^{\text{loc}} + \mathbf{E}^{\text{int}}$, where \mathbf{E}^{int} is the electrostatic field due to the split charges, and $\mathbf{P}(\mathbf{R}) = \sum_\alpha q_\alpha(\mathbf{R}) \mathbf{n}_\alpha / a^2$ (assuming the $q_\alpha(\mathbf{R})$ are smooth functions), we can write in leading order, i.e., up to $O[(ka)^2]$,

$$\tilde{\epsilon}_r(\mathbf{k}) - 1 = \frac{1}{\epsilon_0 a \{\kappa_s + \kappa(ka)^2\}}. \quad (25)$$

See also Refs. 5 and 8 for alternative derivations of this relation. This term for the dielectric permittivity should be interpreted as the high-frequency permittivity, or more precisely, is

appropriate for frequencies that are large compared to those of lattice vibrations but small compared to electronic excitation frequencies.

Eq. (25) immediately reveals two important implications for charge-equilibration models: First, any approach neglecting bond polarisabilities, such as the original QE, assigns a divergent dielectric constant to a system in the thermodynamic limit, i.e., any materials responds to external charges like an ideal metal. This explains, for example, the superlinear polarisability of linear molecules with the degree of polymerisation P .²⁵ Second, any approach neglecting atomic hardnesses will find that $\epsilon(\mathbf{k})$ has little dispersion and thus, small systems exhibit little size effects in that approximation. This again explains why short oligomers do not increase their polarisability with P when modeled in terms of pure bond polarisable models, while SQE shows the correct scaling.²⁵

An additional consequence of Eq. (25) exists for bond-polarisable models. Since there are no atomic hardnesses, the bond or split-charge hardness has to be sufficiently high to ensure that the Hessian of the potential energy is positive definite. This limits the range of applicability to small values of ϵ_r , in particular when no screening for Coulomb interactions is used at small distances, i.e., to $\epsilon_r - 1 \lesssim 1$. Thus, if one wants to model the proper high-frequency dielectric response of a material with $1 < \epsilon_r - 1 < \infty$, both atomic and bond hardnesses need to be considered.

4.1.1 Penetration depth

As an advanced application of the SQE model, let us estimate the length after which the electrostatic field levels off to its value in the bulk. Obviously, this is an important number to reproduce if one is interested in predicting correct forces on ions embedded in a medium described by a polarisable force field.

The split-charge response to an external field can be deduced from Eq. (22). Within $O\{(ka)^2\}$, one obtains:

$$\tilde{q}_\alpha(\mathbf{k}) \approx \frac{\tilde{E}_\alpha(\mathbf{k})}{\kappa_s + \frac{1-\alpha(ka)^2}{\epsilon_0 a} + \kappa(ka)^2}. \quad (26)$$

The denominator on the r.h.s. can be factorised into a constant multiplied with $(k+ik_1)(k-ik_1)$. The inverse of k_1 , i.e., $\delta = 1/k_1$

$$\delta = a \sqrt{\frac{\epsilon_0 a \kappa - \alpha}{1 + \epsilon_0 a \kappa_s}} \quad (27)$$

then corresponds to a characteristic length scale. More detailed calculations as well as simulations reveal that the charge density of a dielectric placed between two parallel, ideal capacitor plates indeed falls according to $\exp(-\Delta z/\delta)$, where Δz is the distance from the surface.⁵

When κ is small, or even zero, δ is small as well, which implies that the electrostatic field drops to its bulk value close to the surface. This is in line with our conclusion that omitting atomic hardnesses eliminates size-dependence of the polarisability of short chain molecules. When κ is large, modification of the short-range Coulomb interaction can affect the value of α and thereby δ . If such a dependence can be used in a beneficial way or instead leads to undesired effects might depend on the system of interest.

4.1.2 Numerical examples

It is instructive to consider a simple model for the dielectric response in different QE models. For this purpose, we study a one-dimensional system, in which the electrostatic potential is altered on a single (super) atom, while that of all others is kept constant. Moreover, all oxidation states are set to zero so that the system is neutral. Results are summarised in Figure 2

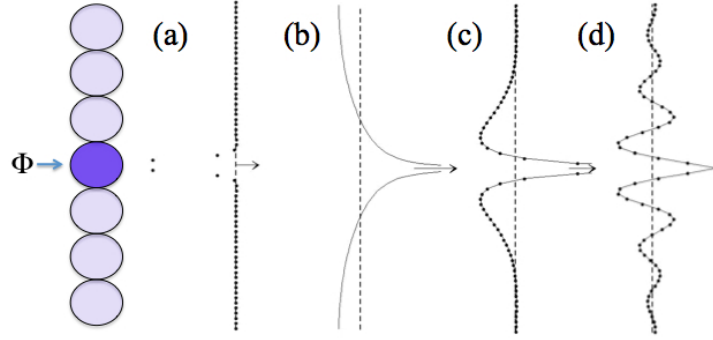


Figure 2. Induced atomic charges in a one-dimensional system sketched on the left-hand-side of the figure. The electrostatic potential is altered only on the central atom. The way in which charge is induced strongly depends on the details of the parameterisation. (a) $\kappa = 0$ corresponding to pure bond-polarisability, (b) $\kappa_s = 0$ reflecting the original QE model, (c) both κ and κ_s are finite as in SQE, and (d) similar to (c) but with next-nearest split charges stiffnesses having a relatively small associated stiffness.

As just discussed in Sect. 4.1.1, the polarisation response is localised in the immediate vicinity of the central atom, when atomic hardnesses are set to zero (as for example in the AACT model). Specifically, the central atom increases its charge dramatically mainly at the expense of the two nearest neighbors in the chain, as can be seen in panel (a) of Figure 2. Charges in the third-nearest neighbors are very small. When nearest-neighbor split-charge stiffness is set to zero while atomic hardnesses are finite, as in the original QE, not only the central atom is heavily charged but also its neighbors. As depicted in Figure 2(b), the positive charge near the central atom is compensated by negative charges near the (hyper-) surface. The SQE model in Figure 2(c) with finite atom and finite nearest-neighbor split-charge hardness produces responses in between the two previous limits. When we also allow second-nearest-neighbor split-charges, new features can arise. This is demonstrated in Figure 2(d), where oscillations arise due to small stiffnesses for split charges connecting next-nearest neighbors.

It is common to express susceptibilities of solids in Fourier space rather than in real space. This is done in Figure 3, where the same parameterisations are considered as in Figure 2. The $1/k^2$ behavior is clearly borne out for the regular QE model just like the weak dispersion for the AACT model, where deviations from the continuum limit require one to approach the border of the Brillouin zone. Interestingly, a non-monotonic dependence of $\tilde{\epsilon}(\mathbf{k})$ is predicted when split charges between next-nearest neighbors are introduced that have high polarisability or small hardnesses.

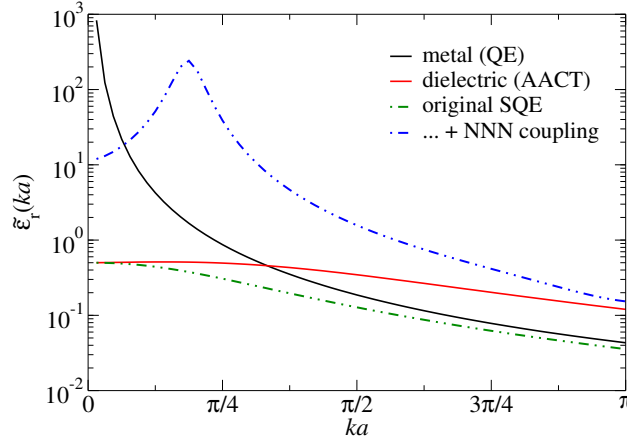


Figure 3. Dielectric constants of various QE models as a function of wave vector. The models are identical to those investigated in Figure 2.

If necessary, more complicated response functions can be modeled by adding split charges between even more distant atoms. To what degree this is helpful in the design of future force fields or the modeling of dielectric phenomena remains to be explored.

4.2 Response to an excess charge

Many processes in physics, chemistry, or biology involve the transfer of one or many electrons from one part of the system to another one. Examples are numerous but one that has had a particularly strong effect on the evolution of science is the charging of amber (Greek for “electron”) when rubbing it against lodestone (used by Thales of Miletus) or cat fur (used by your favorite physics high-school teacher). Thales’ experiment was a first hint that electricity and electric discharge, e.g., lightning, have their origin in nature. Today, electron transfer, particularly in the context of redox reactions play a crucial role, for example, in the development of batteries. Yet, conventional force fields cannot describe redox reactions, because they already fail to describe half reactions, i.e., the addition or subtraction of an electron to a (sub-) system. In order to advance the modeling of redox reactions, it is thus useful to know how the energy of a system changes when we add or subtract an extra integer charge. In this section, we briefly analyse the various QE models in this regard with an emphasis on diatomic molecules. Much of the insights obtained for molecules are also useful to rationalise the response of solids.

Following the presentation in Ref. 26, let us consider a heteronuclear, diatomic molecule in which an external charge $\Delta Q = n_1 e$ is placed onto atom number one, while the oxidation number of atom two is set to zero. The split charge q_{12} shall be denoted as q , the distance between the two atoms is given by a . Moreover, κ_i are atomic hardnesses, χ_i electronegativities, $\Delta\chi = \chi_1 - \chi_2$, and κ_s is the split-charge hardness. The split-charge

energy then becomes:

$$V = \frac{\kappa_1}{2}(\Delta Q + q)^2 + \frac{\kappa_2}{2}(-q)^2 + \frac{\kappa_s}{2}q^2 + \Delta\chi q + \chi_1 \cdot \Delta Q + \frac{J_C}{2}(\Delta Q + q)(-q), \quad (28)$$

where J_C reduces to

$$J_C = \frac{1}{2\pi\epsilon_0 a} \quad (29)$$

for unscreened Coulomb interactions. The split charge is chosen such that it minimises V , i.e., $\partial V/\partial q = 0$, which can be solved to yield:

$$q = -\frac{(\kappa_1 - J_C/2) \cdot \Delta Q + \Delta\chi}{\kappa_1 + \kappa_2 + \kappa_s - J_C}. \quad (30)$$

For the system to be positive definite, the denominator has to be greater than zero. At this point, one can reiterate an insight from the previous section: Omitting the atomic hardnesses means that large values for κ_s must be used, which implies small polarisabilities for molecules or small dielectric permittivities for solids. Eq. (30) can now be inserted into Eq. (28) and the result be sorted into powers of ΔQ . As a result, we obtain

$$V(\Delta Q) = \frac{\kappa_g}{2} \cdot \Delta Q^2 + \chi \cdot \Delta Q + V(0), \quad (31)$$

where

$$\kappa_g = \frac{\kappa_1 \cdot (\kappa_2 + \kappa_s) - (J_C/2)^2}{\kappa_1 + \kappa_2 + \kappa_s - J_C} \quad (32)$$

is the global hardness,

$$\chi = \frac{(\kappa_2 + \kappa_s - J_C/2) \cdot \chi_1 + (\kappa_1 - J_C/2) \cdot \chi_2}{\kappa_1 + \kappa_2 + \kappa_s - J_C} \quad (33)$$

is the global electronegativity, and

$$V(\Delta Q = 0) = -\frac{1}{2} \cdot \frac{\Delta\chi^2}{\kappa_1 + \kappa_2 + \kappa_s - J_C} \quad (34)$$

is the energy associated with the split charge for a neutral molecule in which both atoms have oxidation state zero. A variety of limits shall now be discussed.

In the limit of vanishing bond hardness, i.e., in the original QE model, it does not matter which atom receives the excess charge, because both χ and κ_g are unchanged if indices are inverted when $\kappa_s = 0$. This means that the ionisation energy of a molecule does not depend on which atom in a molecule is ionised. In other words, any molecule behaves like an ideal metal within the QE formalism.

In the limit of vanishing atomic hardnesses, the molecular hardness becomes negative. At the root of this unphysical behavior is that we can add charge of one sign to one atom without having to pay a (self-interaction) energy penalty. One can then add charge of opposite sign to the other atom, which results in Coulomb attraction between the atoms, thereby lowering the energy. This also means that the system becomes unstable to internal redox reaction, which obviously is undesired.

When properly parameterised, that is, for sufficiently large atomic hardnesses, SQE produces none of the just-mentioned artifacts. In this case, the global hardness is positive, and parameters (global hardnesses, ionisation energies, etc.) depend on which atom is ionised provided that the bond hardness is positive. We refer to Ref. 26 for a more detailed discussion, which also pertains to solids. Moreover, a semi-quantitative analysis of the NaCl molecule reveals that by choosing an appropriate value for the bond hardness, one obtains quite reasonable estimates (to within 20% accuracy) for the ionisation energy, $E[\text{Na}^+\text{Cl}^{(0)}]$, the dipole moment $\mu[\text{Na}^+\text{Cl}^-]$ and the first excitation energy $E[\text{Na}^{(0)}\text{Cl}^{(0)}] - E[\text{Na}^+\text{Cl}^-]$. (All other parameters entering the calculations are atom-based properties and Coulomb interactions.) Only the electron affinity $A_{\text{Na}^+\text{Cl}^-}$ does not turn out very accurate. This, however, can be rationalised quite easily: The calculation of that number necessitates the hardness of Cl^- ions, which cannot be readily determined and so using the atomic hardness explains the underestimation of $A_{\text{Na}^+\text{Cl}^-}$.

5 Applications

Most of the studies applying charge equilibration methods are concerned with finding parameters allowing one to reproduce results obtained with DFT or ab-initio methods as closely as possible, e.g., to reproduce electrostatic potential surfaces, partial charges, or interatomic forces. So far, each work found substantial improvements when using SQE, some of which will be discussed further below. However, to get acquainted with the intrinsic properties of QE models, we chose to emphasise the analysis of generic properties and the behavior of toy models, although comparison to DFT-based calculations are also explored.

5.1 Contact electrification

When two initially neutral but otherwise distinct solids touch they tend to exchange charge, which is also known as tribocharging. The tribocharging of dielectrics appears to be very complicated. At least, no consensus has been reached as to the detailed mechanism for the charge transfer.^{27,28} In contrast, tribocharging between metals is reasonably well understood. When two metals touch, electrons go from the metal with the higher work function to that having the smaller work function. Once the two metals are separated, they each carry a charge of same magnitude but opposite sign. As a consequence, they experience a long-range attraction that was not present before contact. Conventional force fields cannot account for the such history dependence. As in conventional DFT, forces on atoms are unique functions of their coordinates and the total charge. Even time-dependent DFT approaches cannot yet tackle problems like the one just mentioned, in which Landau-Zener level or Tully surface hopping is important.

Although descriptions are still at a rather generic level, redox-SQE allows one to mimic processes occurring during contact electrification. Figure 4 depicts the contact dynamics of two solid clusters whose partial atomic charges were calculated within the SQE framework. The figure shows that two initially neutral clusters transfer charge between them upon close approach. After the contact is broken again, the charge remains localised near the front atoms in the dielectrics, for which a finite value of κ_s is assumed inside the clusters. In

contrast to that, charge density delocalises in the metal and predominantly lives on the surfaces.

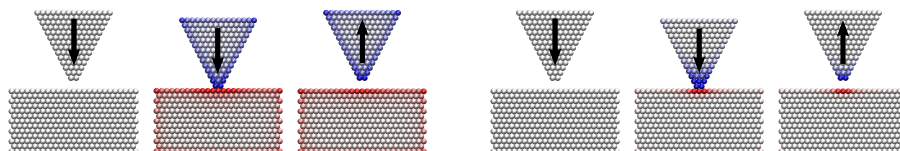


Figure 4. Generic features occurring during contact formation and break up of two metals (**left**) and two dielectric (**right**). Blue and red label positive and negative charge, respectively. In both systems, “dielectric” bonds are assumed to form between two front pairs of atoms. The associated bond stiffnesses are modeled to be very large at the point of bond formation and to decrease quite quickly as the distance between the atoms decreases. Integer charges can be donated through them when the atoms are sufficiently close, but no longer once the bond is considered to be broken. Finite bond hardnesses are used between atoms in the dielectric, preventing the charge to spread out over the surface of the cluster. The latter happens within the metal, where nearest-neighbor split charges are assigned a zero bond stiffness. From Ref. 7.

5.2 Battery discharge

Most atom-based simulations of processes relevant to Galvanic cells are reduced to half cells, because – according to the authors reporting such simulations – it suffices to consider half cells. However, upon close inspection, it turns out the used methods are intrinsically unable to simulate a full cell because they do not incorporate history dependence. A thought experiment shall support this claim:²⁹ Consider a fresh battery with a given voltage between the anode and cathode. We now bring a first demon into play, who keeps all atoms in the battery in place but not the electrons. A second demon connects the anode and the electrode through an external electric wire with a given Ohmic resistance. Charge will flow between the two electrodes and the voltage be reduced. After the second demon removes the wire, all atoms are still in their original place, but the voltage has changed. Methods assigning charges or charge densities as unambiguous functions of nuclear positions cannot predict by how much the voltage has changed. RedoxSQE, however, can overcome this limitation. In addition to their coordinates, atoms are assigned an oxidation state, and so the state of the battery has changed when one or more integer charges have passed through the external wire.

A first redox-SQE based simulation²⁹ of the discharge is sketched in Figure 5. It represents a classical wet cell roughly similar to the original cell designed by Volta. It contains a metallic anode and cathode, which have both been given identical properties, except that the electronegativity of anode atoms is the negative of that of cathode atoms. The cell also contains a molten salt making it energetically favorable for cations to go into solution. A salt bridge allows the salt ions to pass from one half cell to the other but not the metal ions – unless they pay a high price for the energy to pass the salt bridge. Interactions between metal atoms are modeled such that their bonds are considered metallic when their distance is below a threshold value, while a finite split-charge bond hardness is implemented for larger distances. The split-charge hardness is designed such that it increases quickly with

increasing bond length and ultimately diverges at a second cut-off. Redox moves are attempted only for “metal atoms” that are connected through a “dielectric bond”, i.e., when they share a split charge having finite bond hardness. Lastly, anode and cathode can exchange charge through an external wire having resistance R . This latter resistance is the only truly dissipative element in the setup. More details on the setup and the algorithm are given in the original article.²⁹

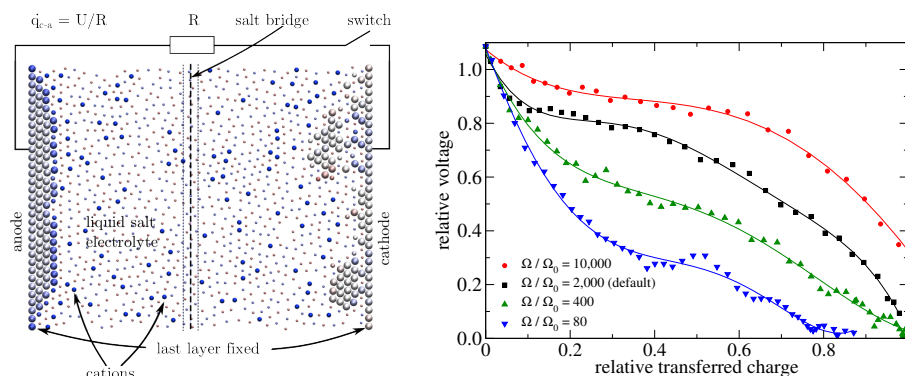


Figure 5. **Left:** Sketch of a redoxSQR battery. **Right:** Discharge characteristics for different external resistances as a function of the transferred charge. A relative voltage of one reflects the work function difference between cathode and anode assuming that both are neutral and placed in a non-polar medium, such as air. From Ref. 29.

Despite the simplicity of the model, the discharge characteristics reproduce many properties of real batteries. First, the initial voltage is slightly above the work function difference ΔV_0 of cathode and anode evaluated for neutral electrodes in a chemically inert, non-polarisable medium. Upon discharge, the voltage quickly drops below ΔV_0 . For large external resistors, the voltage remains slightly below ΔV_0 for longer times, i.e., it decreases relatively slowly until the active material is used up, in which case the voltage quickly goes down. Lastly, the larger the external resistance, the more efficiently the battery is used (unless the discharge is so slow that there is auto-discharge, i.e., migration of metal ions through the salt bridge). We refer again to the original literature²⁹ for more details on these simulations. At this point, it suffices to summarise that the redoxSQR model allows one to mimic the generic properties of electrochemical cells even outside of equilibrium situations, and as such, bears great promise to contribute to the modeling of energy materials.

5.3 Chemistry-specific parameterisations

When reproducing qualitative properties correctly, force fields should also be in a position to model interatomic interactions quantitatively, i.e., in a system-specific fashion. However, such descriptions can only be accurate if the used models are intrinsically able to represent the underlying physics correctly. A model that – such as standard QE – is automatically metallic, cannot be parameterised to reflect the dielectric response of non-metallic molecules, clusters, or solids. This implies that the original QE is intrinsically unable to

correctly account for changes of interatomic forces that are due to dielectric polarisation. In this chapter, we do not review works claiming the opposite. Since the SQE model is not even one decade old, element-specific parameterisations are still scarce, and are mostly concerned with the fitting of partial charges and dielectric response functions rather than with the design of complete force fields. This is why we will first focus on applications of charge equilibration approaches to electrostatic properties and partial charges.

In the original SQE paper,⁶ Nistor *et al.* considered a set of molecules containing sp^3 hybridised carbon and silicon, two-coordinated oxygen, and monovalent hydrogen. They found that the original QE could be parameterised to reproduce partial charges on atoms to within 34% error. A pure bond-based model was slightly better yielding a 28% error, while the combination of the two methods reduced it to 13%. This number could be further reduced to 8% by accounting for leading-order chemical induction, however, at the expense of one additional fit parameter for each bond pair. Mathieu found that the SQE method describes the partial charges during homolysis of a variety of molecules from equilibrium to the final separated fragments quite accurately.³⁰ The results were astonishingly good given the simplicity of the laws describing the divergence of (split-charge) stiffness with increasing bond length. From Warren *et al.*'s work,²⁵ it became clear that SQE can be parameterised to yield the correct polarisabilities of short and long alkanes, at least when using the well-known short-range Coulomb screening ensuring that the SQE Hessian remains positive definite.

The most systematic and exhaustive studies on QE models were conducted by Verstraelen and coworkers.^{14,31,32} A benchmark test on 500 organic molecules selected by an ingenious and autonomous protocol from an initial set of almost 500 000 small organic molecules found that SQE clearly outperformed QE in all 23 benchmark assessments.¹⁴ Moreover, Verstraelen *et al.*³¹ found transferable parameters for the simulation of silicates from isolated structures to periodic systems (both dense crystals and zeolites), although in some cases atomic hardness parameters necessitated environment-dependent corrections. An interesting result of that work is that polarisabilities are reproduced better when parameters are calibrated such that they reproduce the electrostatic potential surface as well as possible rather than the partial charges. Lastly, Verstraelen *et al.*³¹ found that the original formulation of SQE does not describe zwitter-ionic system. However, by introducing constraints, similar in philosophy to the addition of oxidation numbers, SQE also pertains to systems in which the sum of the formal oxidation numbers in a ligand differs from zero.

In the context of full force field development, two works deserve particular mention: Streitz and Mintmire added standard QE to embedded-atom method (EAM) potentials for metals and metal oxides, including their interfaces.³³ This extension lead to an accurate description of elastic properties, surface energies, and surface relaxation. The Streitz-Mintmire potential might have been even more successful if defects in addition to surfaces had been considered. Moreover, there is room for improvement by placing the idea of embedded-atom potentials with charge equilibration on a common footing rather than simply adding it on. For example, the derivative of the embedding functional could be related to the electronegativity of a given atom. Mikulski *et al.* pursued a similar approach as Streitz and Mintmire, this time by adding SQE to bond-order potentials (BOP).^{34,35} The resulting BOP/SQE potentials dramatically improved the transferability of the potentials so that accurate numbers for the heats of formation for isolated molecules, radial distribution functions of liquids, and energies of oxygenated diamond surfaces could be achieved.

6 Conclusions

In this chapter, we assessed different schemes allowing one to model charge distribution at small scales, in particular at atomic scales. Particular emphasis was placed on approaches in which the charge distribution is calculated self-consistently, i.e., through the minimisation of a model for the energy with respect to atomic charges for a given set of atomic coordinates. It turns out that difficulties in such schemes arise because the true energy functional (for the quantum-mechanical ground state) is non-local in space. As a consequence, a coarse-grained formulation, i.e., one in which electron density is approximated by atomic charges and electrostatic multipoles must be non-local as well – in addition to the “trivial” long-range Coulomb interaction. The non-locality can be incorporated through split charges, which describe the polarisation from one atom to another one. While split charges are usually local themselves, their presence – given proper parameterisation – ensures that long-range, fractional charge transfer is suppressed. Currently used functional for DFT calculations do not achieve this even if the kinetic energy is evaluated at the Kohn-Sham level.

An interesting aspect of SQE models is that they allow one to relate their adjustable parameters to collective response functions. For example, the dielectric permittivity of a solid is inversely proportional to the split-charge hardness. The wave-number dependence and the penetration depth – the relaxation length over which an external electrostatic field approaches its bulk value inside a solid – are mostly controlled by the (effective) atomic hardness. When including split charges beyond nearest neighbors, response functions can even be tuned to be non-monotonic in wave number. This makes SQE a promising candidate for the modeling of dielectrics not only at the atomic scale but also at the mesoscale. In contrast, the original QE and simple refinements thereof treat any system as an ideal metal. The computational overhead of SQE with respect to QE is minimal, if present at all. Due to the presence of a band gap for finite split-charge hardnesses, collective stiffnesses do not become small in the thermodynamic limit and so extended Lagrangians are effective for the same reasons why the Car-Parrinello method works well for systems with finite band gap. Thus, the SQE approach bears great potential for use in simulations in which polarisation is important. This includes, in particular, its use in multi-scale and multi-physics descriptions of systems ranging from molecules to condensed matter.

Last but not least, the SQE model allows one to introduce the concept of oxidation number in a way that has successfully guided the intuition of generations of chemists. This, in turn, opens the possibility to mimic Landau-Zener level dynamics or Tully surface hopping, and thus puts one into the position to address non-equilibrium situations as they occur in many systems where charge transfer is important. Examples presented in this chapter are the charge transfer during contact between two solids and the discharge of a battery. It remains a challenge for the future to parameterise the models such that they do not only show generic features but are chemistry or material specific.

Acknowledgments

The author thanks W. B. Dapp and J. Jalkanen for helpful comments on the manuscript.

References

1. R. G. Parr and W. Yang. *Density-Functional Theory of Atoms and Molecules*. Oxford University Press, New York, 1989.
2. A. K. Rappe and W. A. Goddard. *J. Phys. Chem.*, 95(8):3358–3363, 1991.
3. D. M. York and W. Yang. *J. Chem. Phys.*, 104:159–172, 1996.
4. J. Morales and T. J. Martinez. *J. Phys. Chem. A*, 105:2842, 2001.
5. R. A. Nistor and M. H. Müser. *Phys. Rev. B*, 79:104303, 2009.
6. R. A. Nistor, J. G. Polihronov, M. H. Müser, and N. J. Mosey. *J. Chem. Phys.*, 125:094108, 2006.
7. W. B. Dapp and M. H. Müser. *Eur. Phys. J. B* (in press).
8. M. H. Müser. In M. H. Müser, G. Sutmann, and R. G. Winkler, editors, *Hybrid particle-continuum methods in computational materials physics*, pages 171–186. John von Neumann Institute for Computing (NIC), Jülich, 2013.
9. J. Applequist, J. R. Carl, and K.-K. Fung. *J. Am. Chem. Soc.*, 94:2952, 1972.
10. B. Thole. *Chem. Phys.*, 59:341, 1981.
11. A. Papazyan and A. Warshel. *J. Phys. Chem. B*, 101:11254–11264, 1997.
12. K. G. Denbikh. *Trans. Faraday Soc.*, 36:936–948, 1940.
13. J. Meister and W. H. E. Schwarz. *J. Phys. Chem.*, 98:8245–8252, 1994.
14. T. Verstraelen, V. Van Speybroeck, and M. Waroquier. *J. Chem. Phys.*, 131:044127, 2009.
15. P. W. Fowler and P. A. Madden. *J. Phys. Rev. B*, 29:1035–1042, 1984.
16. J. Robles and L.J. Bartolotti. *J. Am. Chem. Soc.*, 106:3723–3727, 1984.
17. M. H. Müser. *Eur. Phys. J. B*, 85:135, 2012.
18. R. Chelli, P. Procacci, R. Righini, and S. Califano. *J. Chem. Phys.*, 111:8569–8575, 1999.
19. W. J. Mortier, K. van Genechten, and J. Gasteiger. *J. Am. Chem. Soc.*, 107:829, 1985.
20. R. G. Parr and R. G. Pearson. *J. Am. Chem. Soc.*, 105:7512, 1983.
21. P. Itskowitz and M. L. Berkowitz. *J. Phys. Chem. A*, 101:5687, 1997.
22. T. Verstraelen, P. W. Ayers, V. Van Speybroeck, and M. Waroquier. *J. Chem. Phys.*, 138:074108, 2013.
23. E. H. Lieb. *J. Chem. Phys.*, 104:159–172, 1996.
24. J. Cioslowski and M. Martinov. *J. Chem. Phys.*, 101:366–370, 1994.
25. G. L. Warren, J. E. Davis, and S. Patel. *J. Chem. Phys.*, 128:144110, 2008.
26. M. H. Müser. *Eur. Phys. J. B*, 85:135, 2012.
27. A. F. Diaz and R. M. Felix-Navarro. *Journal of Electrostatics*, 62:277–290, 2004.
28. L. S. McCarthy and G. M. Whiteside. *Angew. Chem. Int. Ed.*, 47:2188–2207, 2008.
29. W. B. Dapp and M. H. Müser. *J. Chem. Phys.* (in press).
30. D. Mathieu. *J. Chem. Phys.*, 127:224103, 2007.
31. T. Verstraelen *et al.* *J. Phys. Chem. C*, 116:490–504, 2012.
32. T. Verstraelen *et al.* *J. Chem. Theo. Comp.*, 8:661–676, 2012.
33. F. H. Streitz and J. W. Mintmire. *Phys. Rev. B*, 50:11996–12003, 1994.
34. P. T. Mikulski, M. T. Knippenberg, and J. A. Harrison. *J. Chem. Phys.*, 131:241105, 2009.
35. M. T. Knippenberg *et al.* *J. Chem. Phys.*, 136:164701, 2012.

Systematic Coarse Graining of Polymers and Biomolecules

Roland Faller

Department of Chemical Engineering & Materials Science
University of California at Davis, Davis, CA 95616, USA
E-mail: rfaller@ucdavis.edu

I will in this tutorial focus on systematic multiscale modeling of soft materials with applications to polymers and biomolecules. It will start out with an introduction into the fundamental concepts of modeling on multiple connected scales. These include the concepts of mapping models onto each other, fast and slow degrees of freedom etc. The concepts of systematic versus generic mapping will be discussed. Then I will introduce a variety of different techniques. These include both systematic coarse-graining techniques from a structural as well as a thermodynamic standpoint. Specifically techniques to be discussed include the Iterative Boltzmann Inversion, which will be the focus; Force Matching, as well as obtaining Lennard Jones parameters from thermodynamic considerations will be discussed as well. The advantages and limitations of all these techniques will be discussed in order to empower the students to make well informed choices in their own work. After the foundation has been laid I will be discussing several example applications; focusing on both heterogeneous polymer systems and biomembranes.

1 Introduction

Polymers both of the synthetic and the natural biopolymer variety are fascinating materials which are omnipresent in many modern materials applications, be it generating sustainable energy for a growing demand, developing safer solutions for health care, developing DNA and protein based materials as next generation drugs and many more. At the same time computational studies for soft materials are increasingly needed in order to design rather than find by chance new materials. But even with the most powerful computers it is unreasonable to hope that molecular modeling on the atomic scale will be able to predict large scale polymer properties like morphology from first principles.

So we have to devise techniques to develop intermediate and large scale models to predict large scale properties. These models, however, need to be rooted in the local chemistry in order as the large scale behavior is strongly tied to molecular arrangements on local scales. The relevant scales in polymers reach from the distance between bonded atoms on the order of Angstroms to the scales of supermolecular assemblies on the order of micrometers such that no single model can span this range.

For some applications it is sufficient to treat the different scales with fully independent models but in many cases that is not enough such that a true connection – a systematic mapping – between different scales is required. The idea is that we want to speed up the simulation while at the same time reproducing the polymer behavior except the atomistic detail.

One particular class of systems which will be discussed in detail below are organic photovoltaic systems (OPV) which are a cheaper alternative to silicon solar cells.¹ But these systems are still not efficient enough for commercial applications. In polymer solar cells polymers (typically of the thiophene family) are normally mixed with fullerene derivatives.

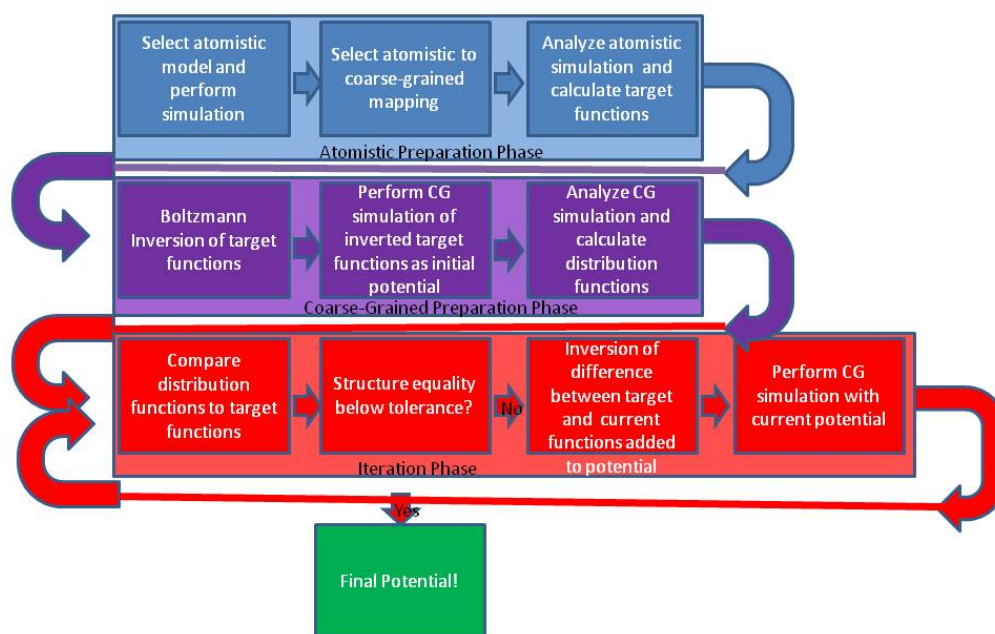


Figure 1. Flowchart of the Iterative Boltzmann Inversion

For best efficiency the local morphology has to consist of interconnected domains where ideally no point is more than 10–20 nm away from an interface in order to optimize charge separation as light first generates an exciton – a charge bound electron–hole pair – and only at an interface this exciton separates into an electron and an hole. A bi–continuous network to fulfill this condition without long–range order is called a bulk heterojunction (BHJ).² Molecular modeling has the potential to elucidate and eventually guide BHJ morphology development if accurate multiscale models exist. Coarse–graining is crucial for such an endeavor.

For the rest of this chapter we first will explain the fundamental theory several coarse–graining approaches. Then we show how they are used for a number of applications. The first of which is a in depth tutorial using the iterative Boltzmann inversion to develop a butanol model. Finally some general conclusions are drawn.

2 Fundamentals and Theoretical Basis of Different Coarse-Graining Techniques

2.1 Iterative Boltzmann Inversion

In order to develop a systematically multi-resolved model we need to use a technique to map models on different length (or time) scales onto another. We focus here on structure, i.e. length scales. An established procedure to coarsen an atomistic model to a mesoscale one is the iterative Boltzmann inversion (IBI).

The iterative Boltzmann inversion starts with an atomistic simulation of the system under study, i.e. one needs an atomistic model. We discuss the IBI here in terms of polymers but any other system can be described as well. One typically has to limit oneself to a small system of short chains. The corresponding distribution functions based on super-atoms are recorded. Super-atoms are linear combinations of atomistic positions.

$$\vec{R}_j = \sum_i w_{ij} \vec{r}_i \quad (1)$$

The capital letters mark the super-atom positions, the lower case the atomistic. Typical choices are centers of mass of a monomer or one central atom to represent a monomer but any other choice is possible. The weights should of course sum up to one, also normally one atomistic atom should not contribute to more than one super-atom. Lastly it is a good idea to choose super-atoms in a way that as many distribution functions (below) are single peaked and sharp.³

The recorded distribution functions include bond lengths, bond angles, torsions and radial distribution functions between super-atoms, i.e. super-atoms are treated as if they were atoms. Before we can generate a potential out of these distributions they may need be weighted by the corresponding Jacobians between internal and Cartesian coordinates. Then they are Boltzmann-inverted to obtain first generation interaction potentials between super-atoms. As the Boltzmann inversion leads to a free energy difference (and not a potential energy) this will need to be iterated to obtain a useful set of potentials. It is often a good idea if you have polymers to ignore the end monomers in this optimization.

$$V(\eta) = -k_B T \ln p(\eta) \quad (2)$$

Here η can stand for bond lengths, bond angles, torsions or non-bonded distances after Jacobian correction alike. This potential is completely numerical but in order to get useful derivatives it can be smoothed by splining. In concentrated solutions or melts the structure of the system is defined by an interplay of the interaction potentials and packing. Thus, a direct calculation of the potential of mean force is not enough and we need an iterative approach to correct the potential which gives the name to the iterative Boltzmann method. Figure 1 illustrates the different stages of a iterative Boltzmann procedure. We iterate this procedure until the desired distributions of the coarse-grained model and the atomistic model coincide within a described tolerance. The final potentials have no physical meaning except that they reproduce the same structure as the atomistic ones. Henderson's theorem⁴ guarantees the uniqueness (but not actually the existence) of an optimized two-body potential.

The optimization process should initially focus on the short distance region. Only after the meso-scale RDF of this region resembles the atomistic RDF well, the tuning process of the larger distances should start effectively. Also we may want to apply different weighting functions w_i for the correction terms depending on the difference between the resulting RDF from the atomistic RDF. E.g. the weighting function is set to 1 when the deviation is about 30 – 40% from the atomistic value. When we are getting closer a series of parallel runs with values of weighting function of 1/8, 1/4 and 1/2 are used to find an optimum starting point for the next step. Again, this procedure is technically only a mathematical optimization to reproduce the structure, we are just using a physically inspired method to do it. So every optimization step (except the first) follows the following equation.

$$V_{i+1}(\eta) = V_i(\eta) - w_i k_B T \ln p_i(\eta) \quad (3)$$

The optimizations for the inter- and the intra-chain interactions can be either performed at the same time or they can be done separately as the mutual effects between the two are negligible. Typically, the intra-chain optimization is much faster than the inter-chain optimization.

Up to now this description focussed on single component systems where each interaction site is equal. If we want to model multi-component systems (including co-polymers), we have to sort the interaction into self-interactions (A–A, B–B etc.) and A–B interactions. Since the self-interaction in the mixture is not the same as in the pure polymer (especially at larger distances homo-interactions are mediated by hetero-atom pairs), there are for binary systems three target RDFs to be optimized. We will have correspondingly more target functions if we are looking at ternary or more complex systems. It may be tempting to use analytical mixing rules for hetero-interactions but it has been shown that in that case even the simplest phase behavior is not correctly reproduced.⁵ Similarly to intra- versus inter-chain interactions we can optimize the hetero and homo interactions at the same time or after each other. There is today no generic scheme but it is preferred to start with the homo interactions as they are regularly similar to pure polymers. Also if a homo-model exists it is preferably used as a starting case.

The IBI only aims at the structure of the polymeric system and it is therefore not guaranteed that the thermodynamic state is in fact correctly described. This has been pointed out by Reith et al.⁶ In order to avoid such problems the co-adjustment of thermodynamic properties can be performed or more correctly the post-adjustment. The most abundant of such cases is pressure optimization. In order to do this a additional pressure correction ΔV_{pc} to the potential of the form

$$\Delta V_{pc}(r) = A_{pc} \left(1 - \frac{r}{r_{cut}} \right) \quad (4)$$

is added, where A is negative if the pressure is too high (which is the typical case) and would be positive if it was too low. As a linear potential corresponds to a constant force the structure does normally not deteriorate significantly.

2.1.1 Treatment of Surfaces

An additional degree of complexity is encountered if we want to treat a system under confinement, e.g. by a hard wall but even a free standing surface changes the interactions close to it. We clearly need an atomistic system under confinement to start. Now the problem is that the system behaves differently as a function of distance from the wall as the local density, packing etc. becomes a function of that distance. One also needs to pay attention to the effective volume while building a mesoscale model of a confined system according to a structure based coarse-graining technique due to the state–point dependence.^{7,8} Therefore, one has to preserve the effective (true) concentration of the molecules from the reference atomistic simulations in the mesoscale system. As the atomistic particles are smaller we might need to increase the system size in the CG case. Determining the effective concentration in a confined system is relatively straightforward: one can inspect the density profiles to deduce the exact vertical distance that the molecules occupy between the surfaces both in the atomistic and mesoscale systems.

It is often a good idea to first perform an unconfined optimization and then re-optimize starting from that already existing model as this speeds up convergence. The direct use of the Boltzmann inverted target distributions of a confined reference system as the initial trial potentials may cause a very slow, glass–like, behavior of the polymers.⁷ Normally one can find a potential which reasonably reproduces the behavior for a range of confinements.

Obviously, in addition to slightly changing the interactions within the system an interaction between wall beads and non-wall beads needs to be set up as well. One may use a set of Lennard–Jones parameters such that the density profile is reproduced. As target observables both the rdfs between super–atoms and the density profiles as a function of distance from the surface should be used. In a recent study of polystyrene–toluene solutions under confinement the final interaction between the wall and toluene beads was in the CG case almost three times the interaction between the PS monomers and wall beads because of the competition between the toluene and PS super–atoms for adsorption.⁸ As a monomer approaches the wall chain connectivity brings other monomers to the surfaces as well increasing adsorption in comparison to solvent which is one single interaction center.

2.2 A Brief Introduction to Force Matching

A widely used alternative to the IBI technique is force matching (FM). The basic idea is to not develop a potential which reproduces structure but to directly reproduce the forces from the atomistic simulation.^{9,10} FM does not use an iterative approach and its use is straightforward in mixed systems. The starting point is again an atomistic simulation and a linear combination of atomistic interaction sites into super–atoms. For FM we need to record and store forces from the atomistic simulation. This is non–standard in most atomistic simulation packages but can be done in some or the code can be adapted. One may also reconstruct the forces from the trajectory a posteriori.

These atomistic forces now have to be mapped onto the CG sites, i.e. we have to calculate the forces between super–atoms. For this every force between two atoms belonging to different super–atoms has to be projected onto the super–atom connection vector (cf. Figure 2) and all these pairs for the same distance have to be averaged.

We again obtain a numerical table, now for forces as a function of distance. This is typically splined or fit to some analytical function in order to avoid too much noise in the

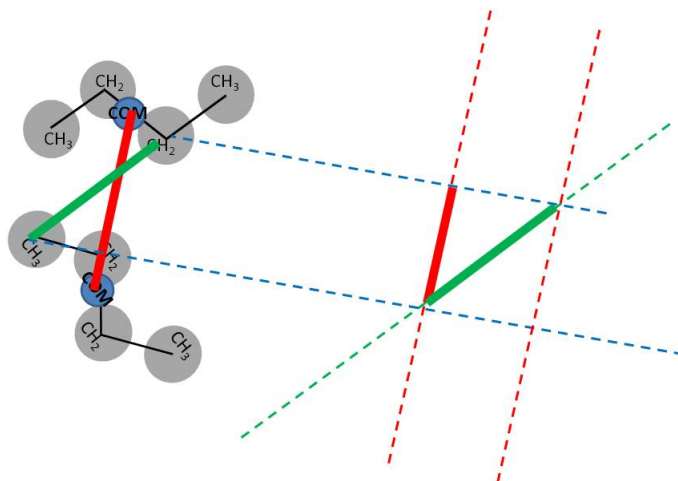


Figure 2. Mapping interactions of atoms of different super-atoms (sketched in green) onto the intermolecular vector (red). The right-hand side shows the projection process. United atom *n*-butane is used as an example.

data. Now these force are directly used in the CG simulations. Note, that no potential is ever constructed or even its existence assumed. Clearly the final forces will be different as the complete philosophy of the matching is different. In general IBI and FM work both very well.

2.3 Developing Parameters for Analytic Interaction Potentials

Another way to determine interaction parameters for a CG simulation is to try to fit the interactions on the large scale to a selected analytical form and reproduce a certain set of parameters; this is essentially very similar how to develop atomistic models. The most abundant form is the Lennard–Jones form but another one is a simple linear potential which is often used in DPD simulations.⁴⁵

The idea is to develop a set of parameters which can describe the behavior of super-atoms in any environment, i.e. one forgoes accuracy for generality. Often atomistic-like mixing rules are being used for simplicity.

In the biophysical Martini model e.g. Lennard–Jones parameters are fitted to reproduce thermodynamic properties of the system.^{11,12} In that case the relative solubility in water and alkanes is chosen. Here no mixing rules are being used but similar to IBI and FM above the hetero-interactions are being developed independently from the homo-interactions.

For DPD one normally selects the size of the bead based on its geometric size. We can measure the size in an atomistic simulation but most of the time it is just selected ad hoc. The interaction strength (potential height at $r = 0$) then can be selected to get the right pressure/density, i.e. essentially the equation of state.

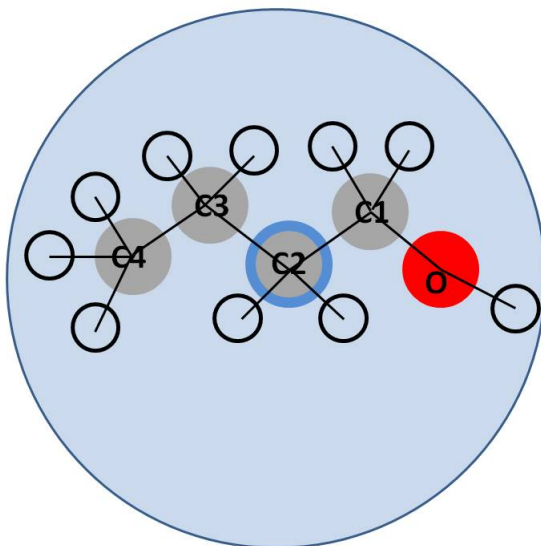


Figure 3. 1:1 Mapping of butanol. The marked carbon (C2) is the super-atom center. The hydrogens, except for the one in the OH group are not actually modeled in the provided atomistic trajectory.

3 Examples

3.1 Butanol Tutorial

A very simple example which we will discuss and also offer as downloadable tutorial is the Iterative Boltzmann coarse-graining of butanol where we only determine the non-bonded interactions. Butanol is a small chain alcohol with the chemical formula C_4H_9OH . We map the whole molecule onto one site. The C2 (see Figure 3) will be used as mapping site. We are using the gromacs¹³ simulation software here and assume that you are on a linux system and have VMD¹⁴ installed as well. Here you find the detailed instructions of how to develop a coarse-grained butanol model from an existing atomistic simulation.

1. Download atomistic data from <http://bit.ly/TNJVfe> or <http://www.chms.ucdavis.edu/research/web/faller/downloads> (IBI.tar or IBI.tar.gz)
2. Visualize trajectory data (BTL-atomist-traj.xtc) and/or final configuration (BTL-atomist-confout.gro) data with VMD
3. Look at the atomistic grompp file (BTL-atomist-grompp.mdp)
4. Make index file
- 4a. `make_ndx -f BTL-atomist-conf.gro -o BTL-atomist-C2.ndx`

4b. Enter "a C2" (selects the C2 atoms which we use as super-atom locations)

4c. Enter "q" (quits program and saves)

5. Make and visualize RDF of the super-atom positions calculated from the atomistic data

5a. `g_rdf f BTL-atomist-traj.xtc -n BTL-atomist-C2.ndx -o BTL-rdf-C2.xvg`

5b. Select "3" twice (make rdf of group 3 (the C2) with itself)

5c. Visualize with `xmgrace`, `gnuplot`, or `excel` (The rdf is just a text file)

6. Make input table for `gromacs`

6a. Compile `rdf2pot.c` (`gcc rdf2pot.c -lm`)

6b. Run and make the interaction table (`./a.out > table.xvg`). This smooths the rdf, calculates the potential and corresponding force and fills the zeros in the rdf with a linear potential. It assumes that `BTL-rdf-C2.xvg` is the atomistic rdf, and that there is no CG rdf, file `rdf.xvg` must not exist.

6c. Look at the tables with `xmgrace`, `gnuplot`, or `excel`

6d. We need the table twice (`ln -fs table.xvg table_BTL_BTL.xvg`)

7. Make a CG configuration of the atomistic configuration

7a. `editconf -f BTL-atomist-conf.gro -n BTL-atomist-C2.ndx -o BTL-CG-conf.0.gro`

7b. select group "3" (we only want C2 atoms)

8. Prepare the CG simulation (`grompp -f BTL-CG-grompp.mdp -c BTL-CG-conf.0.gro -p topol_BTL.top -maxwarn 1`) (You can ignore the warning as it

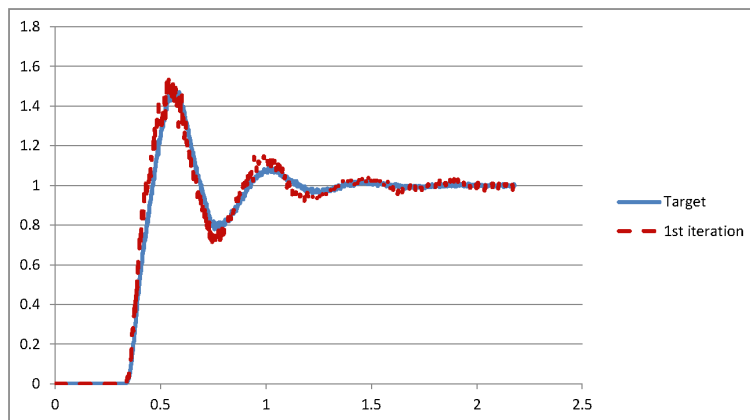


Figure 4. Butanol RDFs: Target from atomistic simulation (blue solid line) and First iteration (red dashed line)

is just a difference in names)

9. Run the CG system (mdrun)

10. Make CG index file

```
10a. make_ndx -f BTL-CG-conf.0.gro
-o BTL-CG-index.ndx
```

10b. Quit (q). This adds the standard groups.

11. Calculate and visualize coarse-grained RDF

```
11a. g_rdf -f traj.trr -n BTL-CG-index.ndx
```

11b. Enter "2" twice (This means calculate the rdf between all butanols)

11c. Visualize it (It got the standard name rdf.xvg)

12. Now you can play around with your CG and atomistic data and explore the communalities and differences.

Figure 4 shows the RDFs you calculated in steps 5 and 11. We see that for such a simple model even without iteration we obtain a good representation. This was now the initial direct Boltzmann inversion of the atomistic run. As we see a difference an iteration is now in order. We first have to determine the difference in rdfs and invert that and add to the potential. Then we rerun the CG and reanalyze it. We can skip a few steps as e.g. all the index files exist already. So in detail the steps 13-19 below are now one complete iteration.

13. Save the old interaction table for reference
(mv table.xvg table-iteration0.xvg)
- 14a. Run rdf2pot again and make the updated interaction table
(./a.out > table.xvg) This again smooths the
rdf, calculates the potential and corresponding
force and fills the zeros in the rdf with a linear potential.
It again assumes that BTL-rdf-C2.xvg is
the atomistic rdf, and now assumes that
there is a CG rdf which is called rdf.xvg.
You can use other filenames. Just run
a.out -h for its usage.
- 14b. We again need it twice but the
symbolic link should still be good.
15. Compare the table with the older one.
16. You can this time use the output configuration
of the first iteration as the next input.
cp confout.gro BTL-CG-conf.1.gro
17. Prepare the CG simulation
(grompp -f BTL-CG-grompp.mdp -c BTL-CG-conf.1.gro
-p topol_BTL.top)
(There should be no warning this time)
18. Run the CG system (mdrun). Gromacs uses for output
again its standard names but back up and
numbers all earlier files existing under these names.
19. Calculate and visualize coarse-grained RDF
- 19a. g_rdf -f traj.trr -n BTL-CG-index.ndx
- 19b. Enter "2" twice (This means calculate the
rdf between all butanols)
- 19c. Visualize it.

After each iteration we have to decide if the current state of the system as represented by its rdf is good enough or if we have to add another step. Clearly this can be automated e.g. by shell scripting. The trickiest part is to decide when the model is good enough. Essentially we have to define a tolerance. One integrates the difference between the target and the current rdf (potentially multiplied by a weighting function) and compares to the tolerance value.

After we have now actually done a simple IBI application let us now discuss a few recent applications where the IBI and other coarse-graining techniques were used for actual research.

3.2 Iterative Boltzmann Inversion of an Organic Photovoltaic System

An example where IBI was successfully applied is an organic photovoltaic system (see Figure 5 for snapshots).¹⁵ Poly(3-hexylthiophene) – P3HT – is the probably best studied OPV polymer although it is not the most efficient. As IBI has to start from an atomistic simulation we briefly describe the atomistic parent model from which the CG model was derived. It bases on the tetrathiophene model of Marcon and Raos¹⁶ and was used as 100% regioregular P3HT (rr-P3HT), in which all monomers are joined head-to-tail. The model originally was optimized using density functional theory calculations.^{16,17} Most Lennard-Jones potentials stem from the OPLS-AA model.^{18,19} The simulation parameters for the alkyl side-chain were taken directly from the OPLS-AA model,^{18,19} except for some charges.

It is always a good idea to validate the atomistic simulations against experiments before starting the CG procedure. Here, e.g. the density of a monomer liquid (0.931 ± 0.003 g/mL) from a short NPT simulation of 256 3HT monomers at 298 K and 1 atm agrees well with experiments (0.936 g/mL²⁰) at the same thermodynamic conditions. The simulated density (1.05 g/cm³) from a constant NPT simulation of a crystal of 3HT 12-mers also agrees with the measured density (1.10 ± 0.05 g/cm³²¹) of P3HT thin films.

After the atomistic simulations have been performed we have to choose the mapping. The mapping which was used to design a coarse-grained model of P3HT in a mixture with the simplest fullerene C60 used three sites for the P3HT monomer: A the center-of-mass of the thiophene ring B the COM of the carbon atoms of the first three and C last three side-chain methyl groups. A single site, the molecule's COM, was used for the CG model of C60. Figure 6 illustrates the coarse-graining scheme used.

The CG simulations in which the interactions were optimized were carried out at constant temperature and volume. After optimization a pressure correction was applied. These

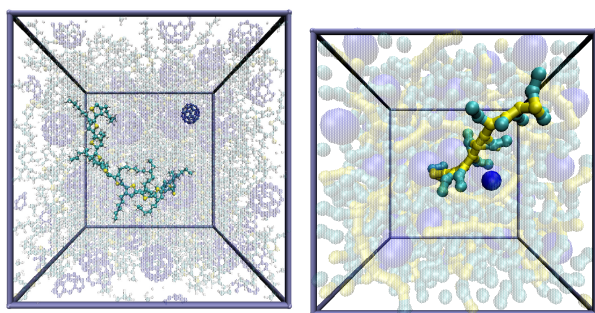


Figure 5. Snapshots of atomistic and coarse-grained representations of P3HT, relevant for organic photovoltaics. Reprinted with permission from David M. Huang, Roland Faller, Khanh Do, and Adam J. Moulé: “Coarse-Grained Computer Simulations of Polymer/Fullerene Bulk Heterojunctions for Organic Photovoltaic Applications” *J Chem Theor Comp* 2010, 6 (2), pp 526–537 Copyright (2009) American Chemical Society.

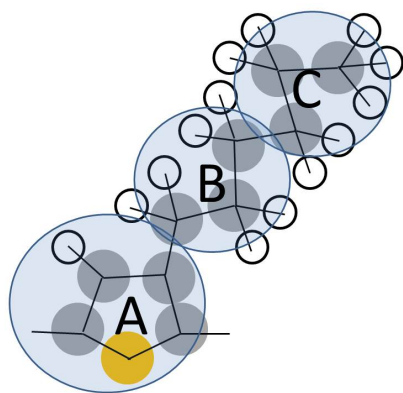


Figure 6. Chemical structure of a P3HT monomer and with coarse-grained mapping sites marked

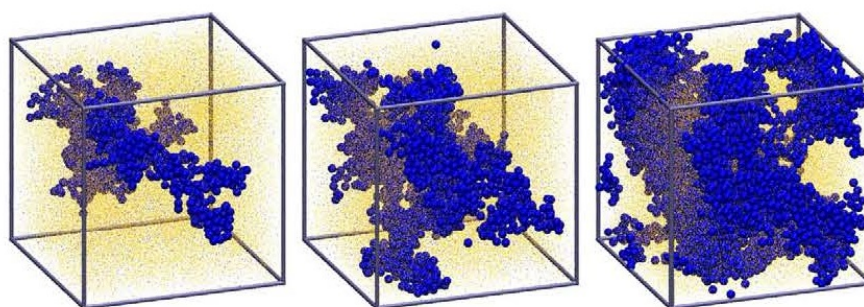


Figure 7. Clusters of fullerenes are growing in a bulk-heterojunction simulation. Reprinted with permission from David M. Huang, Adam Moule and Roland Faller: "Characterization of polymer-fullerene mixtures for organic photovoltaics by systematically coarse-grained molecular simulations" *Fluid Phase Equilibria* 2011, 301 (1-2), pp 21-25 Copyright (2011) Elsevier.

pressure corrections to the non-bonded potentials resulted in almost no change to the calculated RDFs.

The P3HT-P3HT interactions were optimized against pure P3HT at 550 K. Then, the P3HT-C60 and C60-C60 CG interactions were optimized in simulations of 1.85:1 w/w P3HT:C60 with the P3HT-P3HT CG interactions fixed. As these are quite complex polymers in addition to the non-bonded interaction potentials bonded potentials including bonds, angles, torsions and improper dihedrals were used. All bonded potentials were fit to analytical functions based on polynomials. End monomers were excluded from all distribution functions to minimize end effects. It took about 10 iterations to develop the models.

After the model has been developed we can perform simulations on much larger systems than possible atomistically. An example of that is shown in Figure 7 where the formation of fullerene clusters in a mixture of P3HT and C60 is calculated.²²

3.3 Tethered Lipid Bilayers using the Martini Model

This final example shows the use of a variant of the Martini model for a biophysical application. Biomembranes which consist of lipid molecules arranged in bilayers are crucial for compartmentalization of cellular systems. Simplified biomimetic systems containing only a few lipid types are used to understand fundamental membrane properties and at the same time can be used in bionanotechnology and drug delivery. Tethered lipid bilayer membranes are a useful membrane mimetic system where a lipid bilayer is chemically grafted to a solid substrate.

The GROMACS simulation suite version 4.5.2 was used with the MARTINI force field 1.4¹¹ where each interaction site has an 0.47 nm effective size and weighs 72 amu. The non-bonded interactions are described by Lennard-Jones potentials and a screened Coulomb potential with forces shifted smoothly from 0.9 nm (LJ) and 0 nm (Coulombic) to a cutoff of 1.2 nm, respectively. Harmonic potentials are used for bond and angle interactions.

The system contains three types of molecules DOPC lipids, tethered DOPC lipids, and water. Additionally there are immobile surface particles. We had to develop a new particle type for the surface particles in order to represent a hydrophilic surface. We had to be careful to avoid too strong an interaction as otherwise the surface acts as a nucleation site and leads to freezing of the complete water. The interaction between the surface type – P1 – and the water type – P – is 1/3 the value between P and P; the interactions between P1 and the other particle types are 12% of the standard value in the MARTINI model between P and the respective particle types.

In order not to change the interaction density of the surface we restrict ourselves to simulations under constant area but allow the box to fluctuate in z -direction. We have of course to ensure that the surface is tight and no particles leak through it. Another thing one has to be aware of in this case that periodic boundary conditions are normally still applied but the interaction between different replicas has to be completely negligible. As some of the molecules are now chemically attached to the surface (like in a polymer brush) we have to specify how they are geometrically ordered. We use a square lattice; in reality one does not have control over this but can only control the average grafting density. The actual grafting is done by fixing one interaction site in space close to the surface. Simulations were performed for 0.15, 0.31, 0.44 and 0.60 tether/nm² and different lengths of tethers. Dynamics is not the topic of this chapter but here we have to mention that for this model one normally has a speedup of 4, which means that the 20 fs time step used in the simulations is assumed to equal 80 fs in real time. This mapping comes from the diffusion coefficient of water. For details on the exact simulation conditions the reader is referred to articles using this model^{23–26}

Figure 8 shows snapshots of such a system with a tether density of 11% of all lipids. The bilayers remain planar for short tether length but we see instabilities for longer chains (15–20 PEG beads) where tethers start to aggregate. We clearly see a mechanism of instability. Atomistic simulations of such a system would be impossible, on the other hand for this particular case the exact interaction parameters are not so crucial as here largely the mechanisms as a function of more general properties like grafting density and chain length are investigated. In general the Martini model offers excellent insight into mechanisms with semi-quantitative agreement of numbers against experiments.

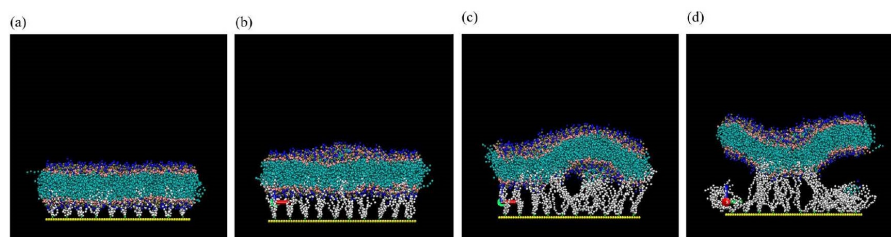


Figure 8. Visualizations of tethered bilayers from 5 – 20 monomer tether length (increasing from left). Reprinted with permission from Chueh Liu and Roland Faller: “Conformational, Dynamical, and Tensional Study of Tethered Bilayer Lipid Membranes in Coarse-Grained Molecular Simulations” *Langmuir* 2012, 28 (45), pp 15907–15915 Copyright (2012) American Chemical Society.

4 Conclusions

Coarse-graining today is not any more a matter of “if” but rather a matter of “how”. It is so clear that for many problems models on different scales have to be developed and adapted. In this chapter the focus was on the Iterative Boltzmann Inversion with an outlook to a few other techniques like force-matching and using semi generic models like the Martini model. Different techniques lead to different models. It is obvious that these different models will behave differently and therefore describe different aspects of the system. No technique will work for everything.

We explain how to do use the IBI in a simple system of butanol, discuss the general theory behind it and show a larger modern example of organic photovoltaics.

All the techniques discussed here and many more need to be in the portfolio of a molecular simulation team in order to be able to address modern questions.

5 Acknowledgements

I need to thank several students and collaborators over the years who were involved in working with me in this area: Dirk Reith, David Huang, Beste Bayramoglu, Chueh Liu, Qi Sun, Adam Moulé, and my former advisor Florian Müller-Plathe.

I also want to thank Sergio Pantano, Jim Pfaendtner and Cameron Abrams who organized a NSF PASI workshop in Uruguay in September 2012 where the butanol tutorial was thoroughly tested and were Cameron co-developed the rdf2pot program used here. I also thank Nithin Dhananjayan who developed and tested a very early version of the butanol mapping and model which eventually evolved into the tutorial.

References

1. Christoph J. Brabec, Jens A. Hauch, Pavel Schilinsky, and Christoph Waldauf, *Production Aspects of Organic Photovoltaics and Their Impact on the Commercialization of Devices*, *MRS Bulletin*, **30**, 50–52, 2005.
2. A.C. Mayer, S.R. Scully, B.E. Hardin, M.W. Rowell, and M.D. McGehee, *Polymer-Based Solar Cells*, *Materials Today*, **10**, no. 11, 28–33, 2007.

3. Roland Faller, *Reviews in Computational Chemistry*, vol. 23, chapter 4: Coarse-Grain Modeling of Polymers, pp. 233–262, Wilvy-VCH, 2007.
4. R. L. Henderson, *Uniqueness Theorem for Fluid Pair Correlation-Functions*, Phys Lett A, **49**, no. 3, 197–198, 1974.
5. Qi Sun, Florence R. Pon, and Roland Faller, *Multiscale modeling of Polystyrene in various environments*, Fluid Ph Equil, **261**, no. 1-2, 35–40, 2007.
6. Dirk Reith, Mathias Pütz, and Florian Müller-Plathe, *Deriving Effective Meso-Scale Coarse Graining Potentials from Atomistic Simulations*, J Comput Chem, **24**, no. 13, 1624–1636, 2003.
7. Beste Bayramoglu and Roland Faller, *Coarse-Grained Modeling of Polystyrene in Various Environments by Iterative Boltzmann Inversion*, Macromolecules, **45**, no. 22, 9205–9219, 2012.
8. Beste Bayramoglu and Roland Faller, *Modeling of Polystyrene under Confinement: Exploring the Limits of Iterative Boltzmann Inversion*, submitted to Macromolecules, 2013.
9. Sergei Izvekov and Gregory A. Voth, *A Multiscale Coarse-Graining Method for Biomolecular Systems*, J Phys Chem B, **109**, no. 7, 2469–2473, 2005.
10. Sergei Izvekov and Gregory A. Voth, *Multiscale coarse graining of liquid-state systems*, The Journal of Chemical Physics, **123**, no. 13, 134105, 2005.
11. Siewert J. Marrink, Alex H. de Vries, and Alan E. Mark, *Coarse Grained Model for Semiquantitative Lipid Simulations*, The Journal of Physical Chemistry B, **108**, no. 2, 750–760, 2004.
12. 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.
13. B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, *GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation*, J. Chem. Theory Comput., **4**, 435, 2008.
14. W. Humphrey, A. Dalke, and K. Schulten, *VMD - Visual Molecular Dynamics*, J. Molec. Graphics, **14**, no. 1, 33–38, 1996.
15. David M. Huang, Roland Faller, Khanh Do, and Adam J. Moulé, *Coarse-Grained Computer Simulations of Polymer/Fullerene Bulk Heterojunctions for Organic Photovoltaic Applications*, Journal of Chemical Theory and Computation, **6**, no. 2, 526–537, 2010.
16. Valentina Marcon and Guido Raos, *Free Energies of Molecular Crystal Surfaces by Computer Simulation: Application to Tetrathiophene*, Journal of the American Chemical Society, **128**, no. 5, 1408–1409, 2006.
17. Valentina Marcon and Guido Raos, *Molecular Modeling of Crystalline Oligothiophenes: Testing and Development of Improved Force Fields*, The Journal of Physical Chemistry B, **108**, no. 46, 18053–18064, 2004.
18. W. L. Jorgensen and J. Tirado-Rives, *The OPLS Force Field for Proteins. Energy Minimizations for Crystals of Cyclic Peptides and Crambin*, J. Am. Chem. Soc., **110**, 1657, 1988.
19. W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives, *Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids*, J. Am. Chem. Soc., **118**, 11225, 1996.

20. Sigma Aldrich, (Ed.), *Aldrich Handbook - a Catalog of Fine Chemicals and Laboratory Equipment 2012–2014*, Sigma Aldrich, 2012.
21. J. Mardalen, E.J. Samuelsen, O.R. Gautun, and P.H. Carlsen, *Chain configuration of poly(3-hexylthiophene) as revealed by detailed X-ray diffraction studies.*, *Solid State Communications*, **77**, no. 5, 337–339, 1991.
22. David M. Huang, Adam J. Moule, and Roland Faller, *Characterization of polymer-fullerene mixtures for organic photovoltaics by systematically coarse-grained molecular simulations*, *Fluid Phase Equilibria*, **302**, no. 12, 21 – 25, 2011.
23. Masaomi Hatakeyama and Roland Faller, *Coarse Grained Simulation of ABA Triblock Copolymers in Thin Films*, *Phys Chem Chem Phys*, **9**, no. 33, 4662–4672, 2007.
24. Ian G. Elliott, Tonya L. Kuhl, and Roland Faller, *Molecular Simulation Study of the Structure of High Density Polymer Brushes in Good Solvent*, *Macromolecules*, **43**, no. 21, 9131–9138, 2010.
25. Shou-Chuang Yang and Roland Faller, *Pressure and Surface Tension Control Self-Assembled Structures in Mixtures of Pegylated and Non-Pegylated Lipids*, *Langmuir*, **28**, no. 4, 2275–2280, 2012.
26. Chueh Liu and Roland Faller, *Conformational, Dynamical. and Tensional Study of Tethered Bilayer Lipid Membranes in Coarse-Grained Molecular Simulations*, *Langmuir*, **28**, no. 45, 15907–15915, 2012.

Theory and Simulation of Charge Transport in Disordered Organic Semiconductors

Peter A. Bobbert

Theory of Polymers and Soft Matter, Technische Universiteit Eindhoven, P.O. Box 513
NL-5600 MB Eindhoven, The Netherlands
E-mail: P.A.Bobbert@tue.nl

The understanding of charge transport in disordered organic semiconductors is crucial for the development of devices based on these semiconductors, like organic light-emitting diodes (OLEDs). The disorder leads to localization of the quantum-mechanical wave functions of charges at specific sites. Charge transport takes place by an incoherent hopping process involving phonon-assisted tunneling between those sites. Three approaches to calculate charge-transport properties in disordered organic semiconductors are described and used here. The first is a coarse-grained drift-diffusion (DD) approach, making use of a mobility function. This is the state of the art in the field. The second approach is a master-equation (ME) approach to calculate the *average* (and possibly time-dependent) occupational probabilities of sites. This turns out to be a very powerful approach, and can be applied to situations in which carriers of only one sign are present, such as in single-carrier devices. On-site Coulomb interactions are taken into account by demanding that not more than one carrier can occupy a site. Long-range Coulomb interactions between charges can be taken into account in an average way, which is important for describing space-charge effects in organic devices. The approach does not allow for an explicit treatment of Coulomb interactions, but this turns out not to be important in describing charge transport in single-carrier devices. The third approach is kinetic Monte Carlo (MC). This approach provides the most realistic description of charge transport, because it simulates the *actual* occupation of sites. Coulomb interactions can be taken into account explicitly in this approach. In principle, the implementation of the approach is straightforward, but special techniques are required to keep the CPU time under control, such as optimized look-up and update schemes, and the use of a cutoff on the Coulomb interaction. The MC approach is the appropriate one for describing charge transport in double-carrier devices, such as OLEDs, because Coulomb interactions play a crucial role in exciton formation in these devices. The application of the approaches to answer various theoretical questions around charge transport in disordered organic semiconductors and the simulation of various organic devices is demonstrated.

1 Hopping Transport

The organic semiconductors used in organic devices such as organic light-emitting diodes (OLEDs), which are now coming to the market, consist of π -conjugated semiconducting polymers or small semiconducting molecules. In either case, these semiconductors are almost always *disordered*. This may seem problematic, because it limits the mobility of charges. However, in OLEDs this is not a problem, since these devices are large-area light sources, so that the current density and therefore the mobility does not need to be large. Also, such devices do not need to switch very quickly: a display does not need to switch more quickly than the eye can follow. The relatively easy synthesis and deposition of organic semiconductors, their relatively low price, and their almost endless chemical variability make them competitive to crystalline inorganic semiconductors in several applications, among which most notably LEDs.

The consequence of the disorder is that the quantum-mechanical wave functions of charges in these semiconductors are localized, for example on a segment of a polymer

in a polymeric semiconductor or on a molecule in a small-molecule semiconductor. It is convenient to approximate these localization regions as point sites i located at positions \mathbf{R}_i . Charge transport then takes place by sudden events, called *hops*, where a charge jumps from site i to another site j . This is an incoherent process caused by phonon-assisted tunneling. The transition rate for this hopping is ω_{ij} . The hopping back from j to i is also possible and the rates ω_{ij} and ω_{ji} should follow the principle of detailed balance:

$$\frac{\omega_{ij}}{\omega_{ji}} = \exp(\Delta E_{ij}/k_B T), \quad (1)$$

where T is temperature and k_B is Boltzmann's constant. $\Delta E_{ij} = E_i - E_j$ is the difference in energy between the situations with the charge located at i or at j . This energy contains all electrostatic energies (due to Coulomb interactions with other charges and possibly an electric field), but also a random contribution because of the energetic disorder that will inevitably be present. In order to avoid confusion with signs our default carriers will be holes. Also, most comparisons with experiment will be for hole transport, because transport of holes is better documented than that for electrons. For electrons, appropriate signs should be introduced.

We will consider here two types of hopping rates that are often used in literature. Both of these rates of course satisfy the condition Eq. (1). The first one is the Miller-Abrahams (MA) hopping rate:¹

$$\omega_{ij} = \nu_0 \exp[-2\alpha R_{ij}] \exp[(\Delta E_{ij} - |\Delta E_{ij}|)/2k_B T], \quad (2)$$

where R_{ij} is the distance between sites i and j , ν_0 is an intrinsic hopping rate, and α is an inverse decay length of the localized wave functions. This hopping rate was derived for the case of coupling to a bath of acoustic phonons. For simplicity, we assume that the inverse decay length α and the intrinsic hopping rate ν_0 are the same for all (pairs of) sites. In principle, further than nearest-neighbor hopping can be considered with MA rates, but in practice (and at not too low temperatures) the nearest-neighbor hops are by far the most important, because the relevant values of α are large (several times the inverse nearest-neighbor distance). If all the nearest-neighbor distances are the same and equal to a , we can absorb the factor $\exp[-2\alpha a]$ into the prefactor and write for ME hopping

$$\omega_{ij} = \omega_0 \exp[(\Delta E_{ij} - |\Delta E_{ij}|)/2k_B T], \quad (3)$$

with

$$\omega_0 \equiv \exp[-2\alpha a] \nu_0. \quad (4)$$

The second hopping rate we will consider is the Marcus one:²

$$\omega_{ij} = \omega_0 \exp(-\Delta E_{ij}^2/4E_r k_B T) \exp(\Delta E_{ij}/2k_B T), \quad (5)$$

with

$$\omega_0 \equiv \frac{J_0^2}{\hbar} \sqrt{\frac{\pi}{E_r k_B T}} \exp(-E_r/4k_B T), \quad (6)$$

where J_0 is a transfer integral. E_r is the energy due to the deformation of the nuclear lattice upon charging (or decharging, the lattice deformation energies for charging and decharging are assumed to be equal) of a site. The Marcus hopping rate therefore considers coupling to local vibrations, or optical phonons. For simplicity, we only consider nearest-neighbor

hopping and equal values for J_0 for all nearest-neighbor pairs, and we assume E_r to be the same for all sites.

Studies of charge transport often focus on the *charge-carrier mobility* μ , which is the average speed of the charge carriers divided by the electric field (dimension: m^2/Vs). The electric field F appears in the hopping rate via a contribution $eFR_{ij,x}$ to ΔE_{ij} for a field in the x -direction, where e is the unit charge and $R_{ij,x}$ is the x -component of $\mathbf{R}_i - \mathbf{R}_j$. An important question is what the dependence is of the mobility on temperature and electric field. Many experimental and theoretical studies of charge transport in organic semiconductors have focused and are still focusing on this question.

2 The Disorder Energy Landscape

As mentioned in the previous section the site energies E_i contain a random contribution $E_{i,\text{rand}}$. If this random contribution can be considered as the summed result of many uncorrelated random effects, it is natural to assume that, because of the Central Limit theorem, the site energies have a normal distribution. Accordingly, a Gaussian density of states (DOS) is assumed for the random contribution:

$$g(E) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{E^2}{2\sigma^2}\right). \quad (7)$$

The standard deviation σ in this Gaussian is henceforth called the *disorder strength*. Consideration of this DOS has led to the *Gaussian disorder model* (GDM) for charge transport in disordered organic semiconductors, which was pioneered by Bässler.³ It was found in Ref. 3 that the energetic disorder is more important than the positional disorder, which has led to the consideration of hopping models on regular lattices with only energetic disorder.

In the original GDM no correlation was assumed between the random contribution to the site energies. In the comparison of the predicted and measured field dependence of the mobility in some organic semiconductors it was concluded, however, that the field dependence predicted by the GDM is not strong enough.⁴ It was suggested that the reason for this is that there is actually a correlation in the random contribution to the site energies. It was proposed that this correlation should be attributed to the presence of randomly oriented dipoles in the semiconductor.^{4,5} Placing a dipole \mathbf{d}_i with fixed magnitude d and random orientation on every site i , the random contribution to the site energies is given by:

$$E_{i,\text{rand}} = - \sum_{j \neq i} \frac{e\mathbf{d}_j \cdot (\mathbf{R}_j - \mathbf{R}_i)}{\epsilon_0\epsilon_r |\mathbf{R}_j - \mathbf{R}_i|^3}, \quad (8)$$

where ϵ_0 is the vacuum permittivity, and ϵ_r the relative dielectric constant of the semiconductor, which is usually around 3. Eq. (8) leads to spatial correlation in the random contribution to the site energies, which decays asymptotically as $1/R$. The resulting model for charge transport is called the *correlated disorder model* (CDM). It should be noted that the DOS that follows from Eq. (8) is not precisely a Gaussian, because the Central Limit theorem does not strictly apply.⁶ The relation between d and the disorder strength σ of the approximate Gaussian DOS is $d = \left(\sqrt{3/A}\right) \sigma \epsilon_0 \epsilon_r / e N_t^{2/3}$, where N_t is the density of sites. For a simple cubic (SC) lattice of sites the numerical factor $A \approx 16.532$, whereas for the face-centered cubic (FCC) lattice $A \approx 15.962$.⁷

3 The Master Equation

Using various techniques it is now possible to determine the charge-carrier mobility μ within the GDM or CDM and to determine its dependence on T and F . The most straightforward way to do this is to put a carrier in the lattice of sites and to simulate its motion by Monte Carlo (MC), where hops are chosen with a probability proportional to their rate.³ However, it was found a decade ago from studying the current density-voltage, J - V , characteristics of hole-only devices of π -conjugated polymers that, apart from the dependence of μ on T and F , also the dependence on the charge-carrier concentration, c , should be taken into account.⁸ It was suggested that in describing these J - V characteristics the dependence of μ on c is even more important than the dependence on F . In MC simulations this would mean that the lattice should be filled with many carriers that interact with each other, making these simulations complicated and CPU- time hungry (see Section 8). Instead, we can consider the *average occupation* p_i of a site i and its change with time. The most important effect of the interactions between carriers is that, because of strong on-site Coulomb repulsion, only one carrier can be present at the same site. This leads to the following equation of motion for p_i , the Pauli *master equation* (ME):

$$\frac{dp_i}{dt} = \sum_{j \neq i} [\omega_{ji} p_j (1 - p_i) - \omega_{ij} p_i (1 - p_j)]. \quad (9)$$

The first term in this equation corresponds to a *gain* in occupation due to carriers that hop from sites j surrounding i to i and the second term to a *loss* in occupation due to carriers that hop from i to surrounding sites. The factors $(1 - p_i)$ and $(1 - p_j)$ account for the maximum occupation of 1.

In a situation of stationary transport the left-hand side of Eq. (9) vanishes and the ME becomes

$$\sum_{j \neq i} [\omega_{ji} p_j (1 - p_i) - \omega_{ij} p_i (1 - p_j)] = 0. \quad (10)$$

Once the p_i are solved from the coupled equations (10) for all i , the current density J follows straightforwardly from the bond currents:

$$J = \frac{e}{L_x L_y L_z} \sum_{i,j} \omega_{ij} p_i (1 - p_j) R_{ij,x}, \quad (11)$$

where L_x , L_y and L_z are the dimensions of the lattice.

Instead of solving for p_i we can also solve for the electrochemical potential energy $\bar{\mu}_i$ (not to be confused with the charge-carrier mobility μ), which is defined at every site in terms of p_i :^a

$$p_i = \frac{1}{1 + \exp([E_i - \bar{\mu}_i]/k_B T)}. \quad (12)$$

Solving for $\bar{\mu}_i$ is mathematically equivalent to solving for p_i , but is in some cases more convenient.

^aWe use here the solid-state-physics definition of electrochemical potential energy. This means that it is not necessarily constant across a device in equilibrium. In electrochemistry, $\bar{\mu}$ would be referred to as the chemical potential energy.

The main advantage of the ME over the MC approach is that no time averaging is needed. As a result, it is often faster than the MC approach. The ME approach is also especially useful for time-dependent modeling, e.g., in describing transients and alternating currents. Another advantage is that, since p_i is known for every site, it is much easier to analyze the behavior of the system at the scale of single sites. The most important disadvantage of the ME approach is that it is not possible to take the Coulomb interactions between individual carriers into account, or to model real OLEDs with both holes and electrons. Another issue is that solving Eq. (10) reliably and in a stable way is often quite difficult, while the MC approach is guaranteed to converge eventually.

One could think that, accepting the assumption of only on-site Coulomb repulsion, the ME approach is exact. However, this is not the case. Technically speaking, a mean-field approximation has been made in the ME Eqs. (9) and (10), which neglects correlations between occupations of different sites. Even with only on-site Coulomb repulsion, such correlations are present. It turns out, however, that corrections due to non-zero correlations are very small⁹ and therefore of no concern to us here.

4 Master-Equation Calculations for a Complete Device

The ME approach also allows calculation of the current in complete organic sandwich devices, including charge-injecting and -collecting electrodes. In such calculations, a distinction must be made between the organic sites and sites describing the electrodes. Electrode sites, representing a metal or a metal-like layer, all have the same energy (the work function of the electrode material) and are neither occupied nor unoccupied: a carrier can always hop to or from one. This is implemented in the calculations by placing *two* layers of sites at each electrode, one unoccupied ($p_i = 0$) and one occupied ($p_i = 1$). Each of these layers is directly accessible from the adjacent layer of organic sites.

Coulomb interactions are taken into account through the long-range space-charge effect only, which lead to an electric field in the x -direction perpendicular to the electrodes. (As noted above, it is not possible to take into account explicit Coulomb interactions between individual carriers in the ME approach.) This leads to a self-consistency problem, since the hopping rates ω_{ij} in Eq. (10) depend on the occupations p_i . To compute this dependence explicitly, we determine the electric potential V_i at every site i , which is iteratively defined by

$$V_i = V_{i-1} - F_{i-1}(x_i - x_{i-1}), \quad (13)$$

where F_i is the electric field between sites i and $i + 1$, and x_i is the x -coordinate of site i , with the indices chosen so that the sites are ordered in the x -direction. We set $V_1 = 0$ by convention. The field F_i is determined by the space-charge approximation, i.e., we spread out the charge on site i over the full lateral layer:

$$F_i = F_{i-1} + \frac{ep_i}{\epsilon_0 \epsilon_r N_t L_y L_z}, \quad (14)$$

where F_0 must be chosen such that the total voltage over the device matches the desired voltage:

$$V_N = V - V_{bi}, \quad (15)$$

where N is the total amount of sites and V_{bi} is the built-in voltage, i.e., the difference in work function between the electrodes. We efficiently compute V_i by first setting $F_0 = 0$ and iteratively calculating the resulting electric potential, which we call $V_i^{(0)}$, from Eqs. (13) and (14). From there we can straightforwardly calculate the correct value of F_0 ,

$$F_0 = (V - V_{\text{bi}} - V_N^{(0)})/L, \quad (16)$$

where $L = L_x$ is the thickness of the organic layer. It is not necessary at this point to redo the calculation for the V_i 's; they are simply given by

$$V_i = V_i^{(0)} - F_0 x_i. \quad (17)$$

These resulting V_i 's obey Eqs. (13)-(15), and can be used to compute the energy difference used in the hopping rates. We note that this approach can also be applied when the sites are not ordered in layers.

The exact solution at zero voltage can also be determined using this approach. An extra wrinkle is now that p_i , required in Eq. (14), is initially unknown. However, we note that by the time we need it, V_i has already been computed. Since thermal equilibrium applies, we have $\bar{\mu}_i = eV_i + e\Phi_{\text{left}}$, with Φ_{left} the work function of the left electrode. We can then obtain p_i from Eq. (12). When we have done this for all sites, we check whether V_N satisfies Eq. (15); if not, we adjust our initial guess for F_0 and rerun the method until it does.

5 Master-Equation Calculations with Periodic Boundary Conditions

In periodic boundary conditions calculations, our goal is to determine the charge-carrier mobility μ . This requires uniform conditions, so instead of using electrodes as boundary conditions in the x -direction we use periodic boundary conditions, just like we do for all cases in the y - and z -directions. In addition, no space-charge effects are taken into account; the electric field F , which we take in the x -direction, is uniform throughout the lattice. An additional equation must be added to the system of equations given by Eq. (10) to fix the carrier concentration c :

$$\frac{1}{N} \sum_j p_j = c. \quad (18)$$

Without this additional equation the system is singular, i.e., it allows multiple solutions. This can be verified by summing Eq. (10) over all i , which yields $0 = 0$. After solving the set of equations (10) and (18) for the p_i 's, the current density and carrier mobility $\mu = J/ecN_t F$ follow from Eq. (11).

In equilibrium, i.e., $F = 0$, the solution is given by a constant electrochemical potential $\bar{\mu}_i = E_F$, with E_F the Fermi energy, which must be chosen such that Eq. (18) is obeyed. Written in terms of the p_i 's this solution is the Fermi-Dirac distribution:

$$p_i = \frac{1}{1 + \exp([E_i - E_F]/k_B T)}. \quad (19)$$

It can be verified straightforwardly that this solution indeed obeys Eq. (10) and leads to $J = 0$ in Eq. (11).

A significant problem in determining μ is that at low carrier concentration the charge transport is largely determined by the few sites with lowest energy. This is because these sites trap most of the charge carriers. For large disorder, the number of such sites can vary significantly between realizations of the disorder, even for very large lattice sizes up to $100 \times 100 \times 100$ sites in the case of, e.g., an SC lattice. We can, of course, average the mobility over multiple realizations, but this does not solve the issue because this average is not necessarily equal to the actual $L_x = L_y = L_z = L \rightarrow \infty$ value of the mobility in the thermodynamic limit. For low values of L the mobility is significantly higher than the actual value, even after averaging over multiple realizations of the disorder. Lukyanov and Andrienko proposed simulating small systems at high temperature and then extrapolating the low temperature behavior,¹⁰ but this requires *a priori* knowledge of the temperature dependence.

An approach to solve this problem is to fix the Fermi energy instead of the carrier concentration. We first determine the Fermi energy corresponding to the desired carrier concentration using the Gauss-Fermi integral:

$$c = \int_{-\infty}^{\infty} \frac{g(E)}{1 + \exp([E - E_F]/k_B T)} dE. \quad (20)$$

After solving this equation for E_F , we use as zero-current solution $\bar{\mu}_i = E_F$. This leads to a carrier concentration in the system that is not necessarily equal to the desired carrier concentration c . We simply accept this concentration as the one to use in Eq. (18), which then becomes:

$$\sum_j p_j = \sum_j \frac{1}{1 + \exp([E_j - E_F]/k_B T)}. \quad (21)$$

The method proceeds as usual from there, i.e., we solve the system given by Eqs. (10) and (21) for the desired field F . The advantage of this approach is that the effect of outlier sites trapping carriers is reduced. For example, suppose that a certain realization of the disorder has more outliers than usual. This simply leads to a higher right-hand side in Eq. (21). In other words, we are adding additional carriers to fill these trapping sites. With this approach, the dependence on lattice size is significantly reduced. This method is used for almost all results presented here. We do note that when considering large fields, the effectiveness of this method is reduced, because the Fermi-energy concept no longer applies. The dependence of the results on lattice size is then much stronger.

6 Solving the Master Equation Iteratively

Yu *et al.* introduced an explicit iterative method to solve the master equation Eq. (10),¹¹ which has been the dominant solution method for several years. In this method, we start with the equilibrium solution as described in Sections 4 and 5. The probabilities p_i are then updated one by one by solving Eq. (10) for p_i , yielding:

$$p_i = 1 / \left[1 + \frac{\sum_j \omega_{ij}(1 - p_j)}{\sum_j \omega_{ji}p_j} \right]. \quad (22)$$

Whenever a probability is updated according to this equation, that updated value is used for all further calculation within the same iteration. These iterations are repeated until satisfactory convergence is achieved.

Both calculations for complete devices and those with periodic boundary conditions require some specific modifications. For device calculations, updating p_i will also change all hopping rates ω_{ij} , because the space charge and the resulting electric field change. In practical calculations, we keep the rates fixed while applying the iterative method. Once this method has converged, we recompute the rates. These two steps are repeated until overall convergence is satisfactory.¹²

For calculations with periodic boundary conditions, we must make sure that the requirement of fixed carrier concentration Eq. (18) is satisfied. The initial equilibrium distribution satisfies this requirement, but the iterations defined by Eq. (22) do not conserve carrier concentration. This problem is solved by first allowing the iterative method to converge and then determining the electrochemical potential energy $\bar{\mu}_i$ from Eq. (12). We then shift $\bar{\mu}_i$ by a constant value for all sites, chosen such that Eq. (18) is satisfied. This running of the method of Yu *et al.* followed by rescaling the potential is repeated until both Eqs. (10) and (18) are satisfied to within specified tolerances.

Although the method of Yu *et al.* has been successfully applied in several cases,^{11–16} it does not reliably converge for large disorder ($\sigma/k_B T \gtrsim 6$). For those cases a combination of the method with Newton's method does lead to reliable results. For this combined method we refer to Ref. 17.

7 The Drift-Diffusion Equation

On a coarse-grained level the organic semiconductor can be viewed as a homogeneous material with a charge-carrier mobility μ of which in organic-device simulations at a fixed temperature only its dependence on the carrier concentration c and the electric field F is important. Instead of the carrier concentration c (dimensionless) it is often convenient to switch over to the carrier density $n = cN_t$ (dimension: m^{-3}). With a mobility function $\mu(n, F)$ the current density J in a single-carrier sandwich device consists on a coarse-grained level of the sum of a drift and a diffusion contribution:

$$J = e\mu(n, F)n(x)F(x) - eD(n, F)\frac{dn}{dx}, \quad (23)$$

where x is the distance from the anode (we consider the case of holes here, so that the anode is the injecting contact). The diffusion coefficient D is related to the mobility by the generalized Einstein expression:¹⁸

$$D(x) = \frac{\mu(n, F)n}{e} \frac{dE_F}{dn}, \quad (24)$$

where E_F is the Fermi energy. Insertion into Eq. (23) yields

$$J = \mu(n, F)n(x) \left[eF(x) - \frac{dE_F}{dn} \frac{dn}{dx} \right]. \quad (25)$$

The field and carrier density are also related by Gauss' law:

$$\frac{dF}{dx} = \frac{en(x)}{\epsilon_0\epsilon_r}. \quad (26)$$

Eqs. (25) and (26) together form a system of differential equations that can be solved for n and F . We will call this the *drift-diffusion* (DD) approach.

We also need to specify boundary conditions at the electrodes. The boundary conditions at the anode ($x = 0$) and cathode ($x = L$) are given by assuming thermal equilibrium with the electrodes. Omitting image-charge effects (a carrier close to an electrode generates an image charge in the electrode with which it interacts), we obtain

$$E_F(0) = \Phi_{\text{left}}, \quad E_F(L) = \Phi_{\text{right}}, \quad (27)$$

where Φ_{left} and Φ_{right} are the work functions of the electrodes. The densities $n(0)$ and $n(L)$ follow from these values of E_F via the Gauss-Fermi integral Eq. (20).

In an experiment one typically applies a chosen voltage and measures the current. When solving the above equations, it is easier to fix the current and determine the voltage, since otherwise J has to be treated as an unknown. At this point the problem is a well-posed boundary value problem. Typically one would now determine the solution on a grid in the x -direction using finite-element or finite-difference techniques. However, for our specific case of single-carrier devices one can convert the problem to an initial-value problem, which is easier to solve. To accomplish this, we guess $F(L)$ instead of specifying $n(0)$.^b Since we now know both $F(L)$ and $n(L)$, we can solve for $F(x)$ and $n(x)$ using a standard differential equation solver (for example in Mathematica). We then simply check if the value of $n(0)$ is consistent with Eq. (27). If not, we try a new guess of $F(L)$. Using this approach, it is possible to determine J - V characteristics and field/density profiles quickly, reliably, and accurately.

8 Monte Carlo

The kinetic *Monte Carlo* (MC) approach simulates the hopping model as described in Section 1 with no further approximations. We provide here a brief overview of the main features of an implementation of this approach. A complete description can be found in Ref. 19.

The method keeps track of the full state of the system, i.e., the locations of all charge carriers. A simulation step consists of choosing and carrying out one of the possible hops from an occupied site i to an empty one j (or to/from one of the electrodes), with the probability of each hop weighted by its rate ω_{ij} . This choice of hop and the necessary update of the hopping rates after the hop are made efficiently by keeping track of all hopping possibilities and their rates using a *binary search tree*, which reduces the look-up time for a hop to be executed as well as the update time of the hopping rates after the hop to a scaling as $M \ln M$, with M the number of carriers in the system. After each hop, time is advanced with a time step that is randomly drawn from an exponential distribution with a decay time equal to the inverse of the sum of the rates of all possible hops. One can mathematically prove that by following this procedure the evolution of the system is precisely simulated as it should be according to the hopping rates. Typically, we start with an empty system and run the simulation long enough to achieve a steady state. After that, we continue running the simulation for some time while measuring the desired quantities, such as the current density.

^bAlternatively, one could guess $F(0)$, but working from right to left turned out to be faster.

The method can take Coulomb interactions between individual charge carriers (and their image charges, if required) into account when computing the energy difference ΔE_{ij} associated with a hop. However, without a cutoff on the Coulomb interactions the hopping rates of all charges would have to be updated after a hop, which would lead to an undesirable scaling of the CPU time with M^2 . Instead, we split the Coulomb interaction into a short-range direct interaction and a long-range space-charge interaction, where the latter is only important in device simulations. Direct interactions are calculated explicitly only if two charges are within a spherical region around each other with a Coulomb cutoff radius R_C (a constant is subtracted from the Coulomb interaction within this sphere such that the interaction is zero at the sphere boundary). Outside this sphere, we consider a charge to contribute to a uniform sheet charge, which is considered as the space charge, giving rise to an internal x -dependent electric field that is treated in the usual way. To avoid double counting of interactions, disc-shaped regions forming the overlap between the spherical regions and the sheets have to be cut out from the sheet charge.¹⁹ The resulting scaling of the CPU time can be limited to $M \ln M$, which means that very large systems can be simulated. For $R_C \rightarrow \infty$, the method is exact (of course at the expense of CPU time) and this allows one to check what value of R_C provides sufficiently accurate results. For all results presented here for SC lattices with lattice constant $a = N_t^{-1/3}$, $R_C = 8a$ was used and found to be sufficient, i.e., increasing R_C further does not significantly affect the results.

The main advantage of the MC approach is that it can fully simulate the hopping model with no simplifications. Unlike the ME approach it can also handle actual OLEDs, where both holes and electrons hop through the device and generate excitons. The main disadvantage is that the method can be slow, since one needs to first allow the system to relax and then run long enough to collect sufficient statistics. This problem is especially severe when the current density is low. It also makes it more difficult to obtain detailed statistics at the site level, such as the occupation probabilities.

9 Example: a Hole-Only Device

All three approaches described above will now be applied to an example hole-only device with an SC lattice, MA nearest-neighbor hopping, disorder strength $\sigma = 0.122$ eV, site density $N_t = 4.28 \times 10^{26} \text{ m}^{-3}$ (corresponding to a lattice constant $a = N_t^{-1/3} = 1.33$ nm), hopping attempt rate $\omega_0 = 5.77 \times 10^9 \text{ s}^{-1}$, and relative dielectric constant $\epsilon_r = 3.2$. For the device length we take $L = 122$ nm, corresponding to 91 organic layers. At the anode we take no injection barrier, while at the cathode we take a work function 1.8 eV below the highest occupied molecular orbital (HOMO), i.e. $V_{bi} = 1.8$ V. These parameters correspond to a hole-transporting polyfluorene-triarylamine (PF-TAA) device studied by van Mensfoort *et al.*²⁰ The values used here are slightly different from those reported in that work; they were recently determined for the same device after three years of aging.

Room-temperature current density-voltage (J - V) characteristics for this device, obtained using the three discussed approaches, are shown in Figure 1. The mobility function $\mu(n, F)$ in the DD approach was obtained by including the carrier-concentration dependence in the GDM. We call the resulting model the *extended Gaussian disorder model* (EGDM). The dependences of μ on T , n , and F were obtained by performing ME calculations of the mobility on SC lattices with up to $100 \times 100 \times 100$ sites using periodic

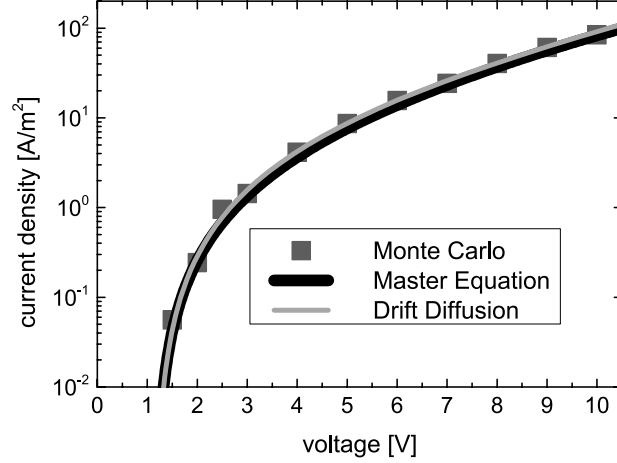


Figure 1. Current density as function of voltage for the example device at room temperature, computed using three different approaches.

boundary conditions. The resulting function $\mu(T, n, F)$ was parameterized in Ref. 14 and this parametrization was used in the DD approach.

We observe from Figure 1 that the three approaches essentially lead to the same result, which is very reassuring. The very small difference between the ME and MC results shows that the influence of short-range Coulomb interactions is insignificant. The very small difference with the DD results shows that the course-graining implied by the DD approach is in this case allowed, which means that it makes sense to speak about a local mobility in the device that varies with position. The experimental J - V curve is not included in Figure 1. It would essentially coincide with the theoretical data.²⁰

10 Transients

The above approaches also allow the calculation of time-dependent properties. Time-dependent experiments on organic devices provide information that cannot be obtained from stationary experiments, such as the measurement of a J - V curve. An easy time-dependent experiment is the measurement of the current transient after a voltage step. This is called a “dark injection” (DI) transient, because the voltage step injects extra carriers in an unilluminated device, which subsequently travel through the device and change the current.

For the example device the DI transient after a voltage step from 1.5 to 8 V is given by the thin black line in Figure 2. After the voltage step an initially large current flows because the current is not yet impeded by the space charge. Because of the build-up of space charge the current decreases. When the front of the space charge reaches the collecting electrode (the cathode), the current shows a maximum (indicated by an arrow) and then continues

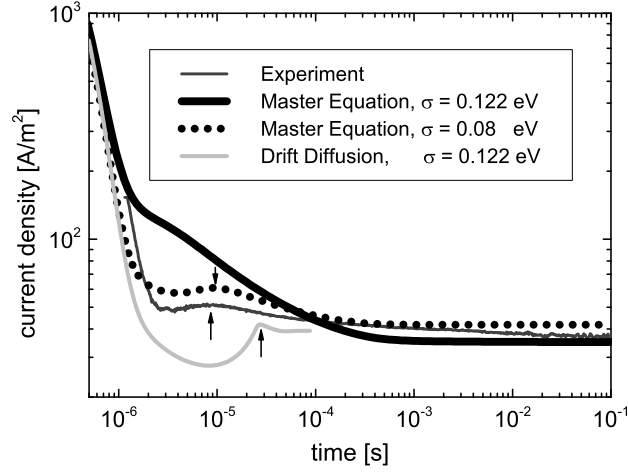


Figure 2. Dark-injection transient for the example device at room temperature. The voltage over the device is stepped from 1.5 to 8 V. Thin black curve: experiment. Thick black curve: master equation with disorder strength $\sigma = 0.122$ eV. Dotted curve: master equation with disorder strength $\sigma = 0.08$ eV. Gray curve: drift-diffusion with $\sigma = 0.122$ eV. The arrows indicate maxima in the current that signal the arrival of the front of the injected space charge at the collecting electrode.

to decrease to the steady-state value at 8 V. In a simple theory with a constant mobility μ_0 in which only drift is taken into account the maximum in the current appears at a time $0.786L^2/\mu_0V$,²¹ so that this time could be used to extract information about the mobility. The transient obtained from the DD approach is given by the gray line. In this approach the mobility at a position x is supposed to depend on the instantaneous charge density and electric field at that position. The maximum in the DD transient occurs about a factor three in time too late. The reason is that carrier-relaxation is not accounted for in this approach. When instead the ME approach is used (thick black lines) the maximum transforms into a shoulder that appears at about the right time. In the ME approach relaxation effects are properly accounted for, but too strong dispersive effects in the transport wash out the maximum. These dispersive effects can be reduced by reducing σ . In fact, when σ is reduced from the value $\sigma = 0.122$ eV obtained from the fit to the J - V curves to $\sigma = 0.08$ eV, the maximum appears and the transient agrees rather well with experiment. In this reduction of σ also the prefactor ω_0 in the MA hopping rates Eq. (3) should be reduced to obtain the right steady-state current density. This points at a possible overestimation of σ by the fit to the J - V curves. In fact, a reasonable fit to these curves can also be obtained with $\sigma = 0.08$ eV.²² This shows that transient currents can be fruitfully used in extracting information about charge transport in disordered organic semiconductors and provide additional information to that obtained from J - V characteristics.

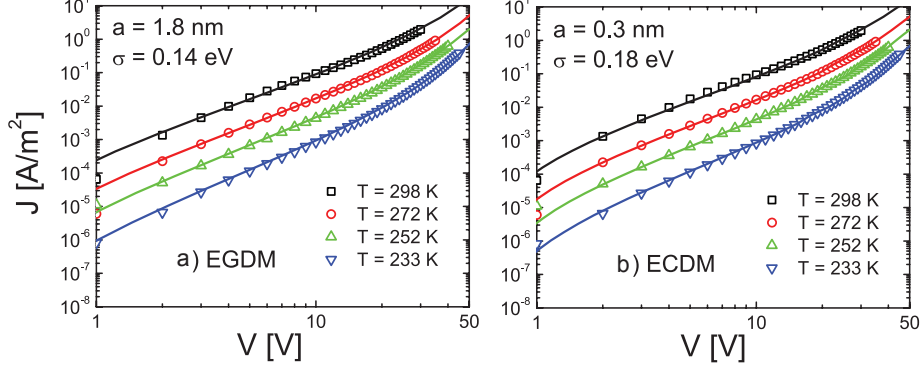


Figure 3. Experimental J - V characteristics (symbols) at various temperatures for an NRS-PPV hole-only device with a thickness $L = 560$ nm, and best fits (lines) with (a) the EGDM mobility and (b) the ECDM mobility model.

11 Uncorrelated or Correlated Disorder?

Just like the extension of the GDM to the EGDM, the CDM can be extended to the *extended correlated model* (ECDM) when the dependence on charge-carrier concentration is included.¹⁶ One can now ask the question with which model J - V curves can be fitted best: the EGDM or the EDCM? In Figure 3 fits to both models are shown of J - V characteristics of hole-only devices with $L = 560$ nm of poly[4-(3,7-dimethyloctyloxy)-1,1-biphenylene-2,5-vinylene] (NRS-PPV). The fit parameters were the intersite distance $a \equiv N_t^{-1/3}$ and the disorder strength σ . The theoretical curves were calculated with the DD approach using the EGDM parametrization from Ref. 14 and the ECDM parametrization of Ref. 16. It is clear that with both models excellent fits can be obtained. However, while the disorder strength $\sigma = 0.18$ eV obtained from the fit with the ECDM is rather large but still acceptable, the intersite distance $a = 0.3$ nm is much smaller than acceptable from the knowledge of the structure of the polymer. On this ground, the ECDM is rejected in this case and the EGDM is accepted as the most appropriate model. More studies of single-carrier (hole-only and electron-only) devices of polymeric^{20,23} and small-molecule^{24,25} semiconductors lead to the provisional conclusion that charge transport in polymeric semiconductors can be better described with the EGDM, but in small-molecule semiconductors with the ECDM.

12 Random-Resistor Network

The master equation, Eq. (10), can be written in terms of the electrochemical potential energy $\bar{\mu}_i$ defined by Eq. (12):

$$\sum_j \frac{e\omega_{ij,\text{symm}} \sinh \left[\frac{\bar{\mu}_i - \bar{\mu}_j + eFR_{ij,x}}{2k_B T} \right]}{2 \cosh \left[\frac{E_i - \bar{\mu}_i}{2k_B T} \right] \cosh \left[\frac{E_j - \bar{\mu}_j}{2k_B T} \right]} = 0, \quad (28)$$

where

$$\omega_{ij,\text{symm}} = \omega_{ji,\text{symm}} = \omega_{ij} \exp(-\Delta E_{ij}/2k_B T) \quad (29)$$

is the symmetrized rate, on the basis of Eq. (1). The electric field F is again applied in the x -direction. At low F , we can linearize Eq. (28) in F , $\bar{\mu}_i - E_F$, and $\bar{\mu}_j - E_F$, to obtain

$$\sum_j \frac{e\omega_{ij,\text{symm}}(\bar{\mu}_i - \bar{\mu}_j + eFR_{ij,x})}{4k_B T \cosh\left[\frac{E_i - E_F}{2k_B T}\right] \cosh\left[\frac{E_j - E_F}{2k_B T}\right]} = 0, \quad (30)$$

where the electric field is now no longer included in the definition of E_i and E_j , so these energies are now solely the random energies chosen from the DOS. Eq. (30) can also be read as Kirchhoff's law of current conservation, with $\bar{\mu}_i - \bar{\mu}_j + eFR_{ij,x}$ the voltage difference and the bond conductance G_{ij} given by

$$G_{ij} = \frac{e^2 \omega_{ij,\text{symm}}}{4k_B T \cosh\left[\frac{E_i - E_F}{2k_B T}\right] \cosh\left[\frac{E_j - E_F}{2k_B T}\right]}. \quad (31)$$

The problem of determining the charge-carrier mobility μ is now equivalent to determining the network conductance G_{network} of this *random-resistor* (RR) network. The relationship with μ is straightforward:

$$\mu = \frac{L_x}{L_y L_z e c N_t} G_{\text{network}}. \quad (32)$$

We note that the RR network approach to the ME is not an approximation. Up to this point, the RR and ME formulations of the hopping problem are mathematically identical in the limit of small F .

13 Percolation Theory and Scaling Ansatz

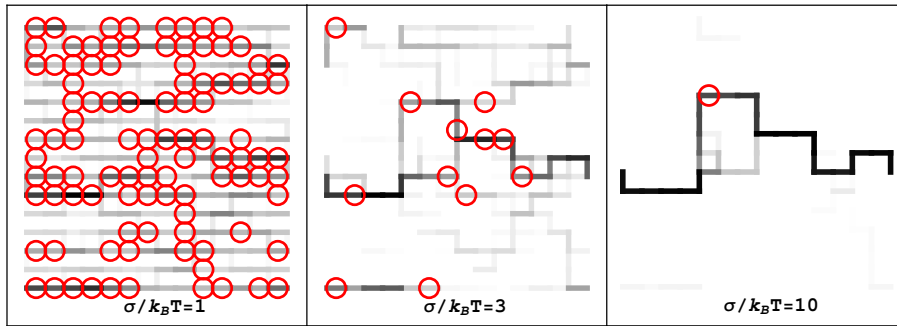


Figure 4. Normalized current (line opacity) in bonds of a 15×15 square lattice. The red circles indicate bonds with a power dissipation of at least 30% of the maximum power dissipation. The results shown are for uncorrelated Gaussian disorder, Marcus hopping with reorganization energy $E_r \rightarrow \infty$, and carrier concentration $c = 10^{-5}$. A small electric field has been applied from left to right.

To demonstrate the percolative nature of charge transport in disordered organic semiconductors, we consider the spatial distribution of current and power dissipation in the RR network, as shown for a 2D system in Figure 4. In the case of low disorder (left panel), the current and power distributions are very homogeneous. Although there are small local variations, these do not extend to a scale of more than a few bonds. This regime can be accurately described using effective-medium theory,²⁶ in which the average effects of the random resistors are described by an effective medium. This theory matches the simulation results for $\sigma/k_B T \lesssim 2$ in a 3D system (see the dashed curve in Figure 5). However, it is not accurate in the experimentally relevant regime $3 \lesssim \sigma/k_B T \lesssim 6$.

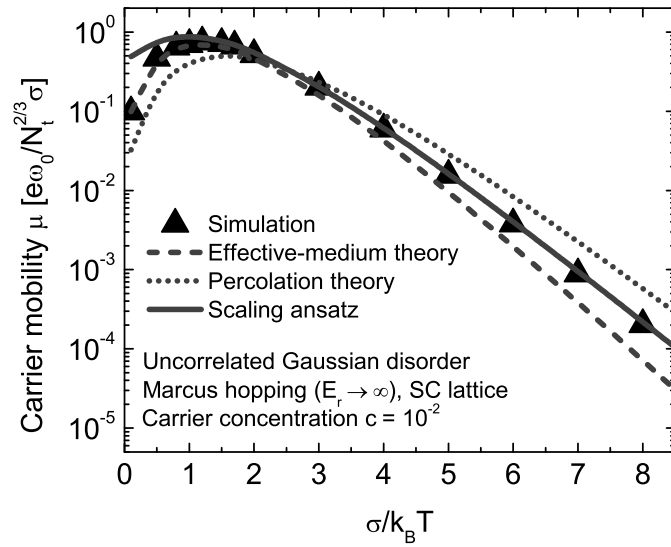


Figure 5. Dependence of the charge-carrier mobility μ on temperature T for Marcus hopping with reorganization energy $E_r \rightarrow \infty$, a simple-cubic (SC) lattice, and uncorrelated Gaussian disorder. Triangles: master equation (ME) simulation results. Solid curve: scaling Ansatz, Eq. (35), with $A = 1.8$ and $\lambda = 0.85$. Dotted curve: standard percolation theory, Eq. (33), with $H = 0.3$. Dashed curve: effective-medium theory (Eq. (5.4) in Ref. 26).

We now consider the opposite limit of high disorder (right panel in Figure 4). In this case, the current follows only the path of least resistance. Along this path, the bond with lowest conductance determines the overall conductance (this is the circled bond in the figure). We will call this bond the *critical bond*, and its conductance the *critical conductance*, G_{crit} . According to this reasoning, we should expect $G_{\text{crit}} = G_{\text{network}}$, but this would lead to a system-size dependence of the mobility; see Eq. (32). For this reason, percolation theories for the charge-carrier mobility generally take the following form:^{27–30}

$$\mu = \frac{H}{N_t^{2/3} e c} G_{\text{crit}}, \quad (33)$$

for some constant H that does not depend on T or c . This standard percolation approach, however, does not quantitatively match the simulation results (see the dotted curve in Fig-

ure 5).

We now focus on the case of intermediate disorder (middle panel in Figure 4). We clearly see the percolative nature of the transport here, with the current being funneled through high conduction pathways. However, there are now *multiple* bonds with high power dissipation. This indicates that the mobility is determined not only by the critical conductance, but also by the *amount* of bonds with such conductance. Dyre *et al.* introduced the term ‘fat percolation’ for this phenomenon.³¹ To quantify this ‘number of bonds’, we use the partial density function of the bond conductances f . We only use the value of this function at G_{crit} , $f(G_{\text{crit}})$. This is justified when the disorder is high enough; bonds with conductance well above G_{crit} can then be considered as perfectly conducting, and those with conductance well below G_{crit} as perfectly insulating. Concluding that the mobility only depends on G_{crit} and $f(G_{\text{crit}})$, and using the fact that it must scale linearly with G_{crit} , we find

$$\mu = \frac{1}{N_t^{2/3} e c} G_{\text{crit}} h[G_{\text{crit}} f(G_{\text{crit}})], \quad (34)$$

for some dimensionless function h . Since percolation can be viewed as a critical phenomenon, with a critical point at $f(G_{\text{crit}}) = 0$, it is logical to propose as a scaling Ansatz for this function h a power-law form:^{32,33}

$$\mu = \frac{A}{N_t^{2/3} e c} G_{\text{crit}} [G_{\text{crit}} f(G_{\text{crit}})]^\lambda, \quad (35)$$

where the constants A and λ do not depend on T or c .

This scaling Ansatz is tested by comparison to simulation results in Figure 5, with the values of A and λ fitted to the data. The values of G_{crit} and $f(G_{\text{crit}})$ were determined using the methods described below. We see that for $\sigma/k_B T \gtrsim 1$ the scaling Ansatz matches the ME simulation accurately. For $\sigma/k_B T \lesssim 1$, not only $f(G_{\text{crit}})$, but the whole distribution $f(G)$ becomes important and the approach fails.

14 Determining the Critical Conductance

To derive a simple expression for the charge-carrier mobility from Eq. (35) we need to compute G_{crit} and $f(G_{\text{crit}})$. To find G_{crit} , let us consider the percolation problem in detail. There is a percolation threshold p_{bond} , such that the portion p_{bond} of bonds with highest conductivity just forms an infinitely large connected network, the percolating network.²⁷ The critical conductance G_{crit} is the lowest conductance occurring in this network. G_{crit} and p_{bond} are related through

$$1 - \Phi(G_{\text{crit}}) = p_{\text{bond}}, \quad (36)$$

with $\Phi(G)$ the cumulative distribution function of the distribution of bond conductances, i.e., $\Phi(G)$ is the probability that a randomly chosen bond has a conductance lower than or equal to G . Since G_{ij} depends only on the energies of the bond sites E_i and E_j , we can work in the (E_i, E_j) -space to obtain

$$1 - \Phi(G_{\text{crit}}) = p_{\text{bond}} = \iint_{G(E_i, E_j) > G_{\text{crit}}} g(E_i) g(E_j) dE_i dE_j, \quad (37)$$

where $g(E)$ is the density of states.^c In words, G_{crit} is determined by the requirement that the contour defined by $G(E_i, E_j) = G_{\text{crit}}$ in the E_i - E_j plane encloses a portion p_{bond} of bonds. This concept is illustrated in Figure 6(a), where the black contour corresponds to $G = G_{\text{crit}}$.

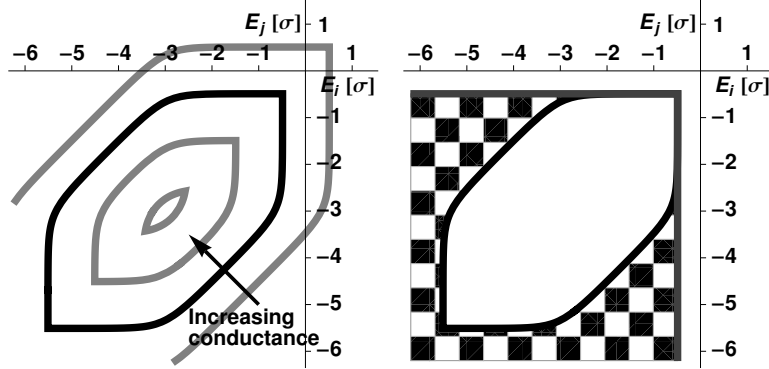


Figure 6. (a) Contours of constant bond conductance (using Eq. (31)) in the E_i - E_j plane for Miller-Abrahams (MA) hopping, Fermi energy $E_F = -3\sigma$, and $\sigma/k_B T = 4$ (corresponding to $c = 0.0033$). The black curve is the contour corresponding to the critical conductance G_{crit} for uncorrelated Gaussian disorder and an SC lattice. E_i and E_j are the energies of the sites linked by the bond. The arrow indicates the direction of increasing conductance. (b) Same G_{crit} contour, using the exact bond conductances as given by Eq. (31) (black) and using the approximation Eq. (38) (dark gray). The checkered area indicates the bonds erroneously considered to have conductance above G_{crit} by Eq. (38).

To proceed, we will approximate the exact bond conductances given by Eq. (31). If the Fermi energy E_F is well below the site energies E_i and E_j , the hyperbolic cosine terms become exponentials, leading to

$$G_{ij} = \frac{e^2 \omega_{ij, \text{symm}}}{k_B T} \exp \left(\frac{E_F}{k_B T} - \frac{E_i + E_j}{2k_B T} \right). \quad (38)$$

In general, E_F is not low enough for this approximation to be accurate for all bonds. However, this only matters in determining G_{crit} if bonds are incorrectly determined to be above or below G_{crit} . These bonds are indicated by the checkered area in Figure 6(b). This area is located at low energies and so contains few bonds (about 0.1% of all bonds in this example). This means that using Eq. (38) instead of Eq. (31) will not significantly affect the value of G_{crit} . We will therefore use Eq. (38) henceforth. We will see later that this is accurate for $c \lesssim 0.01$.

The final step in deriving G_{crit} is realizing that, for both MA and Marcus hopping, G_{ij} can now be written as

$$G_{ij} = \frac{e^2 \omega_0}{k_B T} \exp \left(\frac{E_F - E(E_i, E_j)}{k_B T} \right), \quad (39)$$

^cFor simplicity, we have assumed the energy disorder to be uncorrelated; in the case of correlated disorder, we would have to consider the joint density of states $g(E_i, E_j)$ in Eq. (37), but this does not affect our results.

where E is an energy function of E_i and E_j that does not depend on T or c . We note that the energy dependence of $\omega_{ij,\text{symm}}$ is also included in this function. This allows us to rewrite Eq. (37) as

$$1 - \Phi(G_{\text{crit}}) = p_{\text{bond}} = \iint_{E(E_i, E_j) < E_{\text{crit}}} g(E_i)g(E_j)dE_i dE_j, \quad (40)$$

with the critical energy E_{crit} related to the critical conductance G_{crit} by

$$G_{\text{crit}} = \frac{e^2 \omega_0}{k_B T} \exp\left(\frac{E_F - E_{\text{crit}}}{k_B T}\right). \quad (41)$$

Eq. (40) defines the percolation problem independently of T and c . Thus, E_{crit} is itself independent of T and c , and so Eq. (41) gives the dependence of G_{crit} on T and c . Eq. (40) also shows that the percolating network itself is independent of T and c , a fact that was until now assumed. We note that this does not imply that the structure in the current flow is independent of T and c ; indeed, it can be seen in Figure 4 that it does depend on temperature.

To complete our expression for the mobility we also need to find $f(G_{\text{crit}})$. By definition, the partial density function f is the derivative of the cumulative distribution function Φ , so to find $f(G_{\text{crit}})$ we can, slightly abusing the notation, take the derivative to G_{crit} of $\Phi(G_{\text{crit}})$ as found above:

$$\begin{aligned} f(G_{\text{crit}}) &= \frac{d\Phi}{dG_{\text{crit}}} = \frac{dE_{\text{crit}}}{dG_{\text{crit}}} \frac{d\Phi}{dE_{\text{crit}}} \\ &= \frac{k_B T}{G_{\text{crit}}} \frac{d}{dE_{\text{crit}}} \left[\iint_{E(E_i, E_j) < E_{\text{crit}}} g(E_i)g(E_j)dE_i dE_j \right]. \end{aligned} \quad (42)$$

Here, $dE_{\text{crit}}/dG_{\text{crit}}$ is found from Eq. (41) and $d\Phi/dE_{\text{crit}}$ from Eq. (40). The second factor can be computed numerically. Since it is independent of T and c anyway, we include it in a new constant B :

$$B \equiv AW^\lambda \left(\frac{d}{dE_{\text{crit}}} \left[\iint_{E(E_i, E_j) < E_{\text{crit}}} g(E_i)g(E_j)dE_i dE_j \right] \right)^\lambda, \quad (43)$$

with W the width of the DOS, which we introduce to make B dimensionless. For Gaussian disorder we use $W = \sigma$. The choice of W is somewhat arbitrary, but does not affect the final result.

Combining Eqs. (35), (41), (42) and (43) now yields a simple expression for the temperature and carrier concentration dependence of the zero-field charge-carrier mobility:

$$\mu(T, c) = B \frac{e\omega_0}{N_t^{2/3} W c} \left(\frac{W}{k_B T} \right)^{1-\lambda} \exp \left[\frac{E_F(T, c) - E_{\text{crit}}}{k_B T} \right]. \quad (44)$$

This is the central result of this section. The parameters B , λ and E_{crit} do not depend on T or c , although they typically do depend on the type of lattice, hopping, and energy disorder.

15 Application of the Scaling Expression to Different Hopping Models

In this section we will show how to apply the scaling expression derived above to different hopping models, i.e., different types of lattice, hopping rate and energy disorder. We first

show how to find the parameters in the scaling theory (A , B , λ , p_{bond} and E_{crit}), and list their values for several hopping models.

The simple scaling expression derived in the previous section, Eq. (44), applies to a wide range of hopping models, but we need to find the values of the parameters involved for each model. Specifically, we need to find the percolation threshold p_{bond} , the prefactor A and the scaling exponent λ (see Eq. (35)). From these we can also derive the critical energy E_{crit} (through Eq. (40)) and prefactor B (through Eq. (43)).

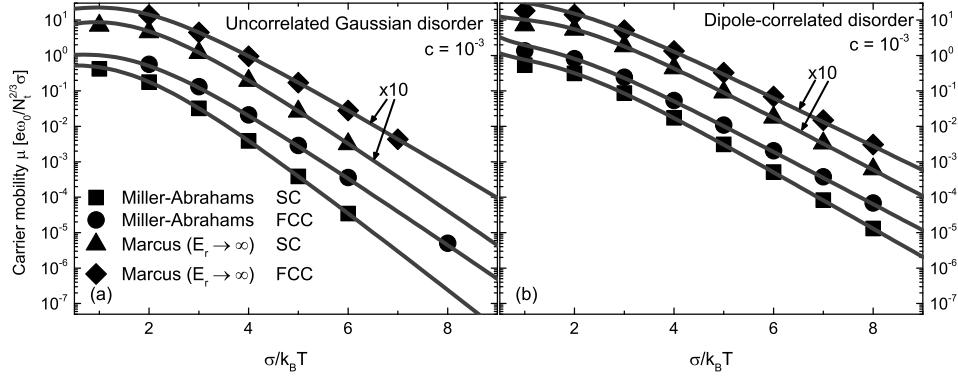


Figure 7. (a) Dependence of μ on T for different lattices and hopping rates, with uncorrelated Gaussian disorder. Symbols: ME. Curves: scaling expression, Eq. (44), with values of B , λ , and E_{crit} as given in Table 1. For clarity the mobilities for Marcus hopping have been multiplied by 10. (b) Same for dipole-correlated disorder, with parameter values as given in Table 2.

We first consider the case of uncorrelated disorder. We start by determining p_{bond} from a percolation analysis. Specifically, we generate the 3D lattice and energy disorder, and calculate the bond conductances using Eq. (38). We then find the critical path from left to right, defined as the path for which the minimum bond conductance along this path is the highest among all paths. To find the critical path we use a modified version of Dijkstra's shortest-path algorithm³⁴ with binary heap sorting.³⁵ The critical bond is the bond with minimum conductance along the critical path. p_{bond} is then simply the portion of bonds with conductance at or above the conductance of the critical bond G_{crit} . Note that this approach also directly gives the values of E_{crit} and $f(G_{\text{crit}})$. Next, we use the ME method to numerically determine the temperature dependence of the charge-carrier mobility. We then fit Eq. (35) to these values, with A and λ as fitting parameters. The value of B finally is calculated from Eq. (43). The parameter values thus obtained for uncorrelated Gaussian disorder are listed for different types of hopping and lattice in Table 1, and the accuracy of the resulting mobility is shown in Figure 7(a).

For Marcus hopping, the dependence of the parameter values on the reorganization energy E_r requires some extra attention. In principle we should consider each value of E_r as a separate hopping model, with its own values of the scaling parameters. However, we found that A and λ depend only weakly on E_r ; the values of A and λ for $E_r \rightarrow \infty$ given in Table 1 can also safely be used at finite E_r . The dependence of the percolation threshold p_{bond} on E_r cannot be neglected, but p_{bond} can be found from the percolation analysis

Uncorrelated Gaussian disorder

Lattice	Hopping	$E_r[\sigma]$	p_{bond}	A	λ	B	$E_{\text{crit}}[\sigma]$	C
SC	MA	N/A	0.097	2.0	0.97	0.47	-0.491	0.44
SC	Marcus	∞	0.139	1.8	0.85	0.66	-0.766	
SC	Marcus	10	0.131	1.8	0.85	0.63	-0.748	0.69
SC	Marcus	3	0.118	1.8	0.85	0.59	-0.709	0.49
SC	Marcus	1	0.104	1.8	0.85	0.51	-0.620	0.44
FCC	MA	N/A	0.040	8.0	1.09	0.7	-0.84	0.40
FCC	Marcus	∞	0.058	8.0	1.10	1.2	-1.11	
FCC	Marcus	10	0.054	8.0	1.10	1.1	-1.09	0.66
FCC	Marcus	3	0.048	8.0	1.10	1.0	-1.06	0.45
FCC	Marcus	1	0.042	8.0	1.10	0.8	-0.98	0.40

Table 1. Bond percolation threshold p_{bond} , prefactor A , critical exponent λ in Eq. (35), prefactor B , and critical energy E_{crit} in Eq. (44), for uncorrelated Gaussian disorder. The last column gives the value C in an optimal fit of the low carrier-concentration mobility $\mu_0(T)$, as given by Eq. (47), to $\exp(-C\hat{\sigma}^2)$ in the range $2 \leq \hat{\sigma} \leq 6$, with $\hat{\sigma} = \sigma/k_B T$. The number of digits given in each entry is compatible with the accuracy with which the parameters could be obtained.

described above, not requiring ME calculations. This also leads to different values of B and E_{crit} , as listed in Table 1. For typical values of T and c , Figure 8(a) shows that the dependence of μ on E_r is well described by this approach. We note that the dependence of ω_0 on E_r , not included in the figure, leads to a net decrease of μ with E_r .

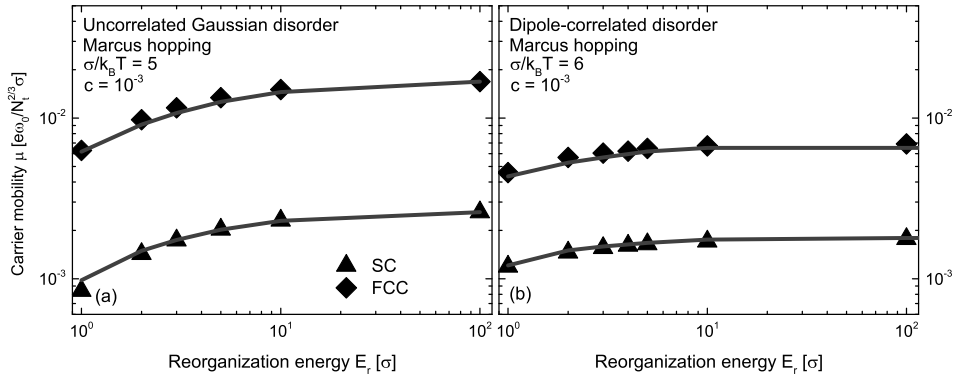


Figure 8. (a) Dependence of μ on E_r for different lattices, with uncorrelated Gaussian disorder. Curves: scaling theory, Eq. (44), with values of B , λ , and E_{crit} as given in Table 1. Interpolation was used for values of E_r not listed in this table. Note that the prefactor ω_0 depends on E_r , which leads to a net decrease of μ with E_r . (b) Same for dipole-correlated disorder, with parameter values as given in Table 2.

We now consider dipole-correlated energy disorder, with a disorder energy landscape obtained from Eq. (8). Figure 9 shows that the topologies of the percolating networks for uncorrelated and correlated disorder are very different; both the high-current bonds and

Dipole-correlated disorder

Lattice	Hopping	$E_r[\sigma]$	λ	B	$E_{\text{crit}}[\sigma]$	C
SC	MA	N/A	2.0	0.36	-1.26	0.33
SC	Marcus	∞	1.7	0.43	-1.37	
SC	Marcus	10	1.7	0.42	-1.37	0.61
SC	Marcus	3	1.7	0.38	-1.37	0.38
SC	Marcus	1	1.7	0.29	-1.37	0.32
FCC	MA	N/A	2.2	0.78	-1.43	0.31
FCC	Marcus	∞	2.2	1.1	-1.56	
FCC	Marcus	10	2.2	1.1	-1.56	0.60
FCC	Marcus	3	2.2	1.0	-1.56	0.38
FCC	Marcus	1	2.2	0.7	-1.56	0.31

Table 2. λ , B , E_{crit} for dipole-correlated disorder. The last column gives the value C in an optimal fit of $\mu_0(T)$, as given by Eq. (48), to $\exp(-C\hat{\sigma}^2)$ in the range $2 \leq \hat{\sigma} \leq 6$.

the critical bonds are much more clustered for correlated disorder. This clustering makes it very difficult to use the percolation analysis described above for correlated disorder; even lattices with $100 \times 100 \times 100$ sites are not big enough. In order to circumvent this problem, we fitted the parameters B , λ , and E_{crit} directly to ME mobility results, using Eq. (44). The results for the two different lattices and hopping types are listed in Table 2. The values of p_{bond} and A are not included in the table, since they are not used in this approach. We note that the different topology of the percolating network for correlated and uncorrelated disorder is reflected in the value of the critical exponent λ , which is around two for correlated disorder and around unity for uncorrelated disorder. The accuracy of the resulting mobility is shown in Figure 7(b).

The reorganization energy needs to be handled slightly differently for correlated disorder. For uncorrelated disorder, we assumed that A and λ are independent of E_r . This approach cannot be used for correlated disorder because we do not know the value of A . Instead, we keep λ constant and fit B and E_{crit} to ME calculations, using Eq. (44). The results are listed in Table 2. Interestingly, no dependence of E_{crit} on E_r is found, contrary to the case of uncorrelated disorder (compare with Table 1). In other words, the dependence of μ on E_r occurs only via the prefactor B . This can be understood by considering the effect of E_r on the hopping rates, as given by Eq. (5): a large value reduces the hopping rate when the energy difference between the sites involved is large. This energy difference is diminished by the correlation of the energy levels, thus reducing the effect of the reorganization energy. The validity of these results is demonstrated in Figure 8(b). Again, we must keep in mind that there is an additional dependence on E_r through the prefactor ω_0 .

16 Effect of Lattice Disorder

We consider here the effect of lattice disorder, which leads to varying distances between sites. Because of the exponential wave-function decay it is natural to replace the hopping

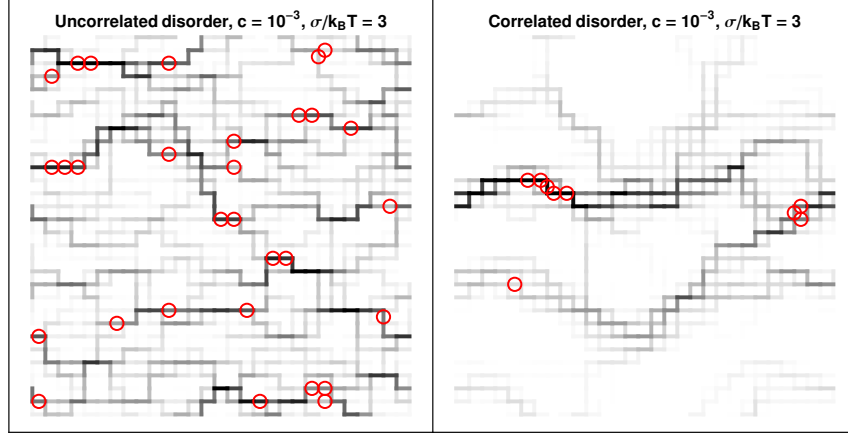


Figure 9. Normalized current (line opacity) in bonds of a 30×30 square lattice with uncorrelated Gaussian energetic disorder (left) and dipole-correlated energetic disorder (right). The red circles indicate bonds with a power dissipation of at least 30% of the maximum power dissipation. The results shown are for Marcus hopping with reorganization energy $E_r \rightarrow \infty$, $c = 10^{-3}$, and $\sigma/k_B T = 3$. A small electric field has been applied from left to right.

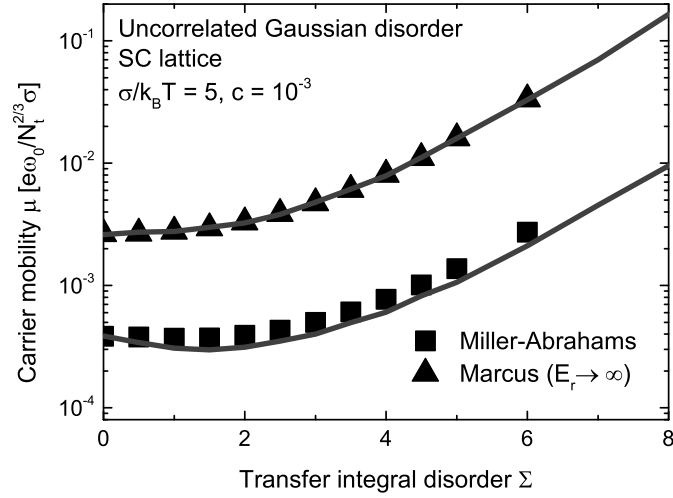


Figure 10. Dependence of μ on transfer-integral-disorder strength Σ . Symbols: ME. Curves: scaling Ansatz, Eq. (35), with values of A and λ as given in Table 1.

prefactor ω_0 in Eq. (3) and (5) by

$$\omega_{0,ij} = \omega_0 \exp(2u_{ij}), \quad (45)$$

where u_{ij} is a random number. We will choose $u_{ij} = u_{ji}$ from a uniform distribution between $-\Sigma$ and Σ for each bond i - j .

It is not a priori clear that Eq. (44) can now be applied, but we can still determine G_{crit} and $f(G_{\text{crit}})$ from the percolation analysis described in Section 15 and apply the basic scaling Ansatz Eq. (35), assuming no dependence of A and λ on the lattice disorder strength Σ . The results of this approach are compared with ME results for typical values of T and c in Figure 10; we see that the scaling theory still provides an excellent description of the mobility, even for large disorder $\Sigma = 6$. We also note that for $\Sigma \lesssim 3$ the mobility is almost independent of Σ , so that Eq. (44), valid for $\Sigma = 0$, can still be applied in this case. We can conclude from this analysis that lattice disorder does not change our results significantly and that energetic disorder is dominant.

17 Carrier-Concentration Dependence of the Mobility

An important conclusion drawn from Eq. (44) is that the dependence of the charge-carrier mobility μ on the concentration c is in all cases given by

$$\mu \propto \exp(E_F(T, c)/k_B T) / c, \quad (46)$$

containing no parameters depending on the type of hopping or lattice. For MA hopping this dependence was already found in Ref. 30. We now conclude that it also holds for Marcus hopping, at variance with another claim.³⁶ We note that our conclusion agrees with the numerically exact mobilities, as shown in Figure 11. When the carrier concentration is too high, the assumption of low Fermi energy used in deriving Eq. (38) no longer holds, and so the above dependence also fails. The requirement for uncorrelated Gaussian disorder is $c \lesssim 0.03$, and for dipole-correlated disorder $c \lesssim 0.01$. The higher threshold for uncorrelated disorder is caused by the higher value of E_{crit} .

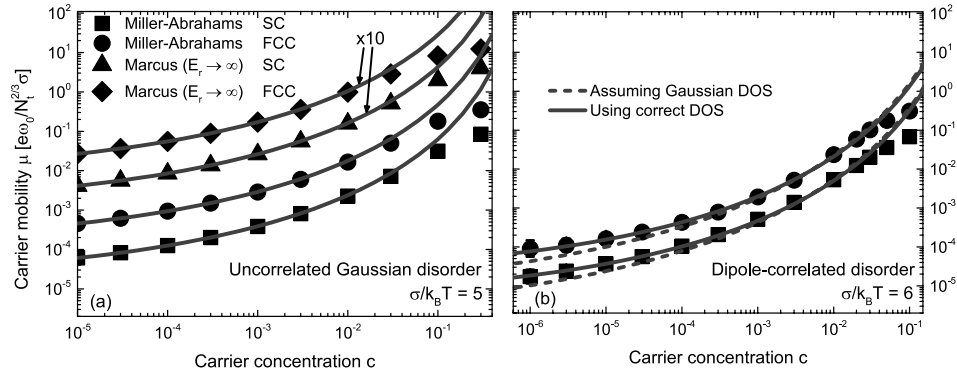


Figure 11. (a) Dependence of μ on carrier concentration c for different lattices and hopping rates, with uncorrelated Gaussian disorder. Symbols: ME. Curves: scaling expression, Eq. (44), with values of B , λ , and E_{crit} as given in Table 1. For clarity the mobilities for Marcus hopping have been multiplied by 10. (b) Same for dipole-correlated disorder, with parameter values as given in Table 2. The dashed curve indicates the result of Eq. (44) assuming a perfectly Gaussian DOS, while the solid curve uses the actual DOS, which is not precisely a Gaussian.

When applying Eq. (46) to the case of dipole-correlated disorder, we must keep in mind that the DOS is not perfectly Gaussian. With the correct DOS a slightly weaker con-

centration dependence is found than for uncorrelated Gaussian disorder, see Figure 11(b), consistent with the ECDM results found by Bouhassoune *et al.*¹⁶

18 Temperature Dependence of the Mobility

The temperature dependence of the charge-carrier mobility is typically analyzed in the limit of low carrier concentration $c \rightarrow 0$, i.e., for a single non-interacting carrier. For uncorrelated Gaussian disorder the mobility in this limit, $\mu_0(T)$, is given by (starting from Eq. (44)):

$$\begin{aligned}\mu_0(T) &= B \frac{e\omega_0}{N_t^{2/3}\sigma} \hat{\sigma}^{1-\lambda} \exp[-E_{\text{crit}}/k_B T] \lim_{c \rightarrow 0} \frac{\exp(E_F(T, c)/k_B T)}{c} \\ &= B \frac{e\omega_0}{N_t^{2/3}\sigma} \hat{\sigma}^{1-\lambda} \exp[-\hat{\sigma}^2/2 - E_{\text{crit}}/k_B T].\end{aligned}\quad (47)$$

This expression does not apply to the dipole-correlated case because the DOS is not exactly Gaussian for that case. In that case, the following approximation can be made:⁷

$$\mu_0(T) \approx B \frac{e\omega_0}{N_t^{2/3}\sigma} \hat{\sigma}^{1-\lambda} \exp[-0.56\hat{\sigma}^{1.9} - E_{\text{crit}}/k_B T].\quad (48)$$

We have to keep in mind that in the case of Marcus hopping ω_0 depends on T via Eq. (6), leading to an additional temperature dependence that is not explicitly shown in Eqs. (47) and (48).

In Ref. 30 the expression $\mu_0(T) \propto T^\gamma \exp(-b\hat{\sigma}^2 - a\hat{\sigma})$ ($\hat{\sigma} \equiv \sigma/k_B T$) was derived with $a = 0.566$, $\gamma = -1$ and $b = 1/2$ for nearest-neighbor MA hopping with an SC lattice and uncorrelated Gaussian disorder. Our expression for $\mu_0(T)$ is of the same form, also with $b = 1/2$. However, the values of a and γ differ: for MA hopping we have $a = E_{\text{crit}}/\sigma$ and $\gamma = \lambda - 1$, and for Marcus hopping, accounting for the T dependence of ω_0 , $a = (E_{\text{crit}} + E_r/4)/\sigma$ and $\gamma = \lambda - 3/2$. Note that the sign of a found by us for MA hopping (see E_{crit} in Table 1) is *opposite* to that in Ref. 30, leading to a significantly different T dependence.

The temperature dependence of the mobility is often expressed as $\mu_0(T) \propto \exp(-C\hat{\sigma}^2)$. We find that this provides a quite accurate description of Eqs. (47) and (48) when considering a limited temperature range $2 \leq \hat{\sigma} \leq 6$. To facilitate the comparison with earlier work, we have included the value of C in such a fit in Tables 1 and 2, taking into account the dependence of ω_0 on T for the case of Marcus hopping. For correlated disorder the much lower value of E_{crit} leads to a significantly weaker temperature dependence, i.e., a lower value of C . This is consistent with the ECDM results.¹⁶ For the case of an SC lattice with uncorrelated Gaussian disorder and MA hopping, the obtained value of C (0.44) is similar to the best-fit value $C = 4/9$ found from an MC simulation of this system by Bäessler.³ This result is often interpreted as if the temperature dependence of the mobility is determined by the rate of hops from the average carrier energy $-\sigma^2/k_B T$ to a ‘transport level’ with an energy around $-(5/9)\sigma^2/k_B T$. We note that the origin of the similar factor $\exp(-(1/2)\hat{\sigma}^2)$ in Eq. (47) is very different: it originates from the limit taken in deriving this equation and results purely from the physics of carriers obeying Boltzmann statistics in a Gaussian DOS and not from the transport properties.

19 Monte Carlo Modeling of Electronic Processes in a White Multilayer OLED

We now make a big jump and discuss the application of the MC approach in modeling of electronic processes of a white OLED with a design that is similar to present commercial white OLEDs. These OLEDs consist of a multilayer stacks of different organic small-molecule semiconductors, where each layer has a specific function. Light of different colors is emitted in different emissive layers, together composing white light. The electronic processes taking place in such a multilayer stack involve the injection of electrons and holes from suitable electrodes, the transport of electrons and holes to the inner layers in the stack, the formation of excitons by mutual capture of electrons and holes, the diffusion of excitons to the place where they should decay, and the final decay of the excitons under the emission of a photon. Monte Carlo (MC) is in principle ideal to model these processes, because

- All these processes involve incoherent sudden events, which can be ideally simulated with MC.
- It is in principle possible to model every molecule in the OLED by a site in the MC computer program.
- If proper rates of charge hopping and exciton hopping (and possibly other events, such as interaction events between excitons or between charges and excitons) are implemented one can just run the simulation and trust the outcome, because no approximations are made. This could save OLED manufacturers development time, because they do not need to worry about the effects of approximations that have to be made in other approaches, like the DD approach.

The big problem could of course be that it is simply not feasible to simulate realistic OLEDs with MC, because of excessive computational demands. We will see, however, that we can be optimistic at this point.

In Ref. 22 the multilayer OLED stack of Figure 12 was studied. The OLED has been fabricated by thermal evaporation in ultra-high vacuum of the organic materials displayed in this figure. The structure is ideal for a fundamental study, because most of the used materials have been well characterized in literature and all relevant processes can be addressed. The generation of the primary colors in this OLED is based on a *hybrid* principle, used extensively nowadays in commercially available white OLEDs. Green and red light are generated in layers of a host organic semiconductor doped by green and red phosphorescent dyes. A heavy metal atom in such dye molecules (in this case iridium) opens up, by its strong spin-orbit coupling, a radiative decay pathway for triplet excitons, next to singlet excitons. Hence, almost all excitons formed in such layers decay under the emission of a photon. Hybrid OLEDs avoid the use of blue phosphorescent dyes (of which the stability is still an issue) by using instead blue-emitting molecules without a heavy metal atom, at which only the singlet excitons decay radiatively by fluorescence. This compromises the internal quantum efficiency, because the triplet excitons (75% of all formed excitons) have no efficient radiative decay pathway and are thus wasted. Still, the power efficiency of today's commercial hybrid white OLEDs is already a factor of three to four higher than

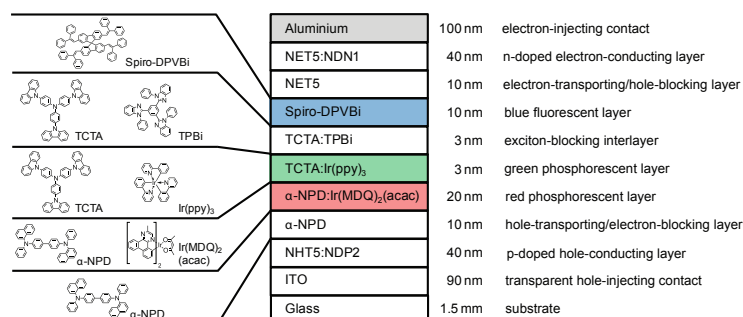


Figure 12. Schematic of the studied multilayer white OLED stack. Given are the chemical structures of the used organic molecular semiconductors and dyes (apart from proprietary materials from the company Novaled), and the thickness and function of the layers. See Ref. 22 for more details.

that of incandescent light bulbs ($\sim 40\text{--}60\text{ lm/W}$ vs. $\sim 15\text{ lm/W}$), combined with operational lifetimes exceeding 10,000 hours.

Of crucial importance to the functioning of multilayer OLEDs is that excitons are generated at the right place by encounter of electrons and holes, and that their subsequent motion until the moment of radiative decay is precisely controlled. In the OLED of Figure 12 an exciton-blocking interlayer has been inserted in between the blue and green layer. This interlayer prevents the motion of singlet excitons from the blue to the green layer as well as that of triplet excitons from the green to the blue layer. These are unwanted energetically downward processes that have to be blocked. On the other hand, motion of excitons from the green to the red layer can take place because of the direct contact between these layers. This is a desired process, because it leads to the right color balance in this OLED, as will become clear.

Next to the control of the exciton motion, the control of the motion of electrons and holes is crucial. This control is achieved by using organic semiconductors with appropriate energy levels of electrons and holes; see Figure 13. First, electrons and holes have to be injected from suitable electrodes (of which at least one has to be transparent, in this case indium-tin-oxide, ITO) into the organic layers. Highly n- and p-doped organic layers adjacent to the electron- and hole-injecting electrodes provide an almost barrier-free contact with these electrodes. From these doped layers the electrons and holes smoothly enter the electron- and hole-transporting layers, via which they move to the inner layers of the stack. The electron and hole energies of the organic semiconductors used in the transport layers are such that charge carriers of only one polarity can enter these layers. This guarantees that electrons and holes meet in the inner emissive layers of the stack and form excitons there. The OLED functions as an optical microcavity, in which exciton formation close to metallic electrodes must be avoided because this would lead to non-radiative decay.

A special role is again played by the interlayer between the green and blue layer. This very thin (3 nm) layer should block excitons, but allow passage of both electrons and holes, in order to guarantee exciton formation in both the green and blue layer. In order to achieve

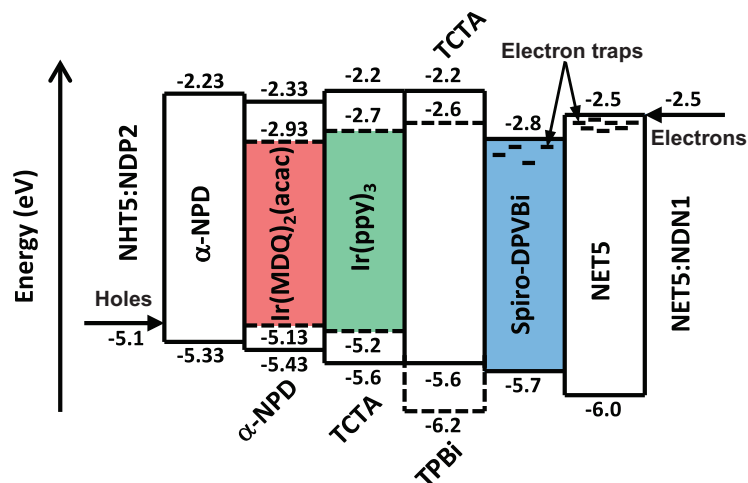


Figure 13. Energy-level scheme of the OLED at open circuit. Indicated are the hole and electron energies of the highest occupied and lowest unoccupied molecular orbitals of the corresponding molecules, in eV. Due to disorder, these energies are broadened by approximately 0.1 eV. Electron traps are indicated in the two layers where they matter.

this, the interlayer consists of a mixture of an electron transporter and a hole transporter, where the electron energy of the former matches well to that of the blue fluorescent material, while the latter material is the same as the hole-transporting host in the green layer.

It is possible to reconstruct the emission profile of the different colors *within the OLED* from the angle-, wavelength-, and polarization-dependent emitted light intensity, with a nanometer-scale accuracy.³⁷ The result is given in Figure 14(a). The balance between emission of the primary colors, with a strong red component, leads for this OLED to the emission of warm-white light. Resolution of the emission profile within the very thin (3 nm) green layer is just beyond the limits of the reconstruction approach. The profiles in the red and blue layer are on the scale of a few nanometres confined to the interfaces with the green layer and interlayer, respectively.

MC simulations of the charge and exciton motion have been performed by modeling the OLED stack as an array of hopping sites representing all the different molecules in the stack, including the dyes.²² Electron traps occur in many organic semiconductors. These are taken into account in the layers where they matter: the electron-transporting and blue fluorescent layers. We take an SC lattice with a lattice constant of $a = 1$ nm, which is the typical distance between the molecules. All molecules are given an electron and a hole energy according to the energy-level scheme of Figure 13. Random energies should be added because of the disorder. Because we are dealing with small-molecule semiconductors, we assume that the disorder is correlated. We therefore generate a disorder landscape from the electrostatic potential of randomly ordered dipoles placed at the lattice points (Eq. (8)). The size of the used SC lattice is $56 \times 50 \times 50$ sites. The disorder strength is taken to be $\sigma = 0.1$ eV, which is the value found for hole transport in α -NPD,²⁴ an

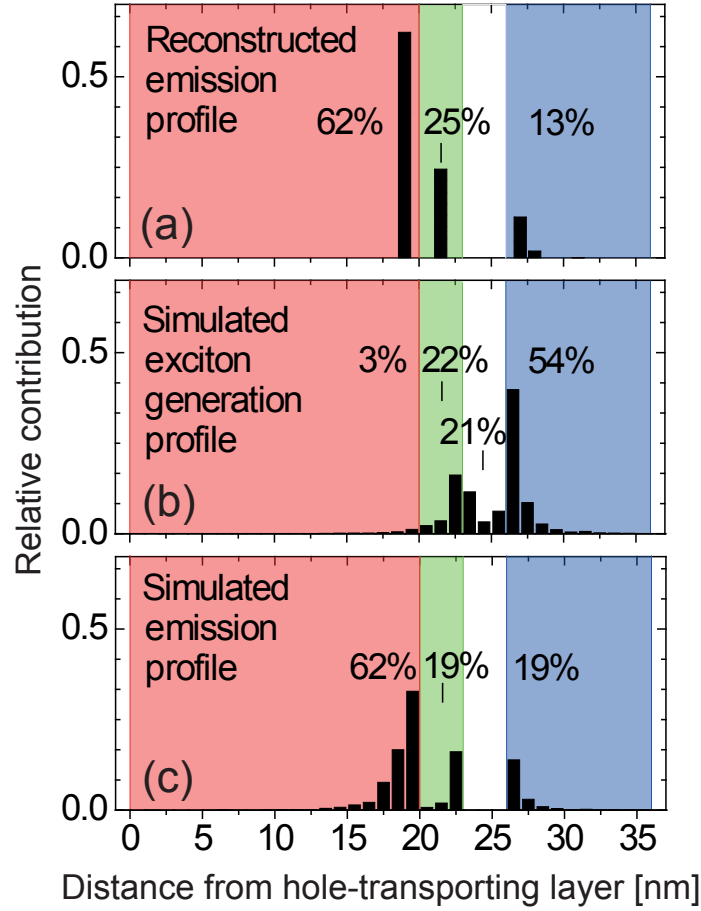


Figure 14. (a) Reconstructed light-emission profile of the OLED, at a bias voltage of 3.6 V. (b) Simulated exciton generation profile. (c) Simulated light-emission profile. The difference between (b) and (c) is caused by motion of excitons from the green to the red layer and by radiative emission probabilities smaller than unity. Excitons generated in the interlayer (white) are lost.

important hole-conductor used in the stack (see Figure 12). The doped electron- and hole-conducting layers are modeled as metallic-like contacts. The hopping rates in the various layers are chosen to reproduce available experimental information about the mobility of electrons, holes, and excitons in each material in the stack; see Ref. 22 for details. Coulomb interactions between all charges are taken into account. Electrons and holes attracting each other by the Coulomb force form excitons. Subsequent exciton motion is simulated within the green and red layer, and, importantly, from the green to the red layer. Excitons formed in the blue layer will stay there, because of the adjacent exciton-blocking interlayer. Information about the radiative decay efficiencies of the blue fluorescent and the green and

red phosphorescent emitters, determining the fraction of excitons that decay by emitting a photon, is taken from experiments. This information is needed to predict the light-emission profile from the simulations. Excitons formed in the interlayer are assumed to be lost.

A first check of the validity of the simulations is the comparison between the calculated and measured current density in the OLED. At the operating voltage of 3.6 V the current densities agree to within 25%,²² which is a gratifying result in view of the rather drastic approximations and assumptions made. In Figure 14(b) the simulated exciton generation profile is given. We find that indeed almost all injected electrons and holes form excitons. As desired, most excitons form in the emissive layers, with the majority of excitons (54%) formed in the blue layer. Also, a considerable fraction of excitons is formed in the green layer (22%), while almost no excitons are formed in the red layer (3%). Figure 14(b) reveals an important loss mechanism caused by excitons formed in the interlayer (21%). This leads to a suboptimal efficiency, as is indeed observed in a measurement of the external quantum efficiency (EQE) of the OLED.²²

After taking into account the excitonic motion and radiative decay efficiencies the simulated light-emission profile of Figure 14(c) is obtained. We observe the same large component of emitted red light as found in the reconstruction of the experimental light-emission profile of Figure 14(a), which is almost completely caused by transfer of excitons from the green to the red layer. Also the green and blue component of the simulated emission profile are in fair agreement with the reconstructed emission profile. Like the reconstructed profiles, the simulated profiles are confined to nanometer-scale regions close to the interfaces. The overall agreement between the reconstructed and simulated emission profiles is striking.

20 Concluding Remarks

In the present work it has been shown how computational approaches of various degree of sophistication can be used to model charge transport in disordered organic semiconductors and sandwich devices of these semiconductors. This modeling has also contributed to the theoretical understanding of this charge transport, in particular the percolative nature of this transport. The approaches discussed in this work, the drift-diffusion (DD), master-equation (ME), and Monte Carlo (MC) approach all have their advantages and disadvantages. Researchers in the field will keep using the fast DD approach for quick device calculations. We have seen that this approach is perfectly suitable to obtain current-voltage (J - V) characteristics of single-carrier devices. However, the approach is not suitable for describing transport in situations that are far out of equilibrium, such as in the case of dark-injection (DI) transients. The master-equation approach is more involved, but very powerful in the description of charge transport in single-carrier devices, also in situations that are far from equilibrium. Moreover, the approach provides powerful insight into the percolative nature of the transport and understanding of this transport in the context of scaling arguments. However, Coulomb interactions cannot be explicitly included in the approach. This does not appear to be a problem in the description of charge transport in single-carrier devices, but the approach cannot (or at least not straightforwardly) be applied to double-carrier devices, where Coulomb interactions between electrons and holes play a crucial role in the formation of excitons. With MC one can simulate in principle precisely what is happening in a real organic device and include all the effects of Coulomb interactions. Therefore,

MC simulations provide the most powerful approach to describe electronic processes in organic devices. Because of their CPU-time hungriness MC simulations have their limitations, but efficient algorithms and the ever increasing computing power allow simulations for impressively large systems.

The MC approach seems to have opened the road towards rational design of multilayer OLED stacks based on molecular-scale modeling of electronic processes. Extensions of the present approach in various directions are possible. Inclusion of exciton-exciton and exciton-charge quenching processes will be important to assess efficiency loss and material degradation by these processes. Another important extension is the incorporation of information about the microscopic morphology of the stack materials, obtained with molecular dynamics or Monte Carlo modeling, and about hopping rates obtained from quantum-chemical calculations.^{38,39} This will finally allow complete predictive multi-scale modeling of electronic processes in OLEDs.

Acknowledgments

I am indebted to many colleagues who have made this work possible. In particular, I would like to mention my colleague Reinder Coehoorn from Philips Research, who has come up with many research questions from his environment of commercial OLED research and fabrication. Also, he has contributed to almost all the work presented here. Material has been used from the work of the following PhD students and postdocs (alphabetically): Mohammed Bouhassoune, Marco Carvelli, Jeroen Cottaar, Harm van Eersel, Jeroen van der Holst, Siebe van Mensfoort, Murat Mesta, Frank van Oost, Frank Pasveer, and Rein de Vries. Large parts of the text come from the thesis of Jeroen Cottaar.¹⁷

References

1. A. Miller and E. Abrahams, *Impurity Conduction at Low Concentrations*, Phys. Rev., **120**, no. 3, 745, 1960.
2. R. A. Marcus, *Electron transfer reactions in chemistry. Theory and experiment*, Rev. Mod. Phys., **65**, no. 3, 599, 1993.
3. H. Bässler, *Charge Transport in Disordered Organic Photoconductors*, Phys. Status Solidi B, **175**, 15, 1993.
4. Y. N. Gartstein and E. M. Conwell, *High-field hopping mobility in molecular systems with spatially correlated energetic disorder*, Chem. Phys. Lett., **245**, no. 4–5, 351–358, 1995.
5. S. V. Novikov, D. H. Dunlap, V. M. Kenkre, P. E. Parris, and A. V. Vannikov, *Essential Role of Correlations in Governing Charge Transport in Disordered Organic Materials*, Phys. Rev. Lett., **81**, 4472–4475, 1998.
6. S. V. Novikov and A. V. Vannikov, *Distribution of the electrostatic potential in a lattice of randomly oriented dipoles*, Sov. Phys. JETP, **106**, 877, 1994.
7. J. Cottaar, R. Coehoorn, and P. A. Bobbert, *Scaling theory for percolative charge transport in molecular semiconductors: Correlated versus uncorrelated energetic disorder*, Phys. Rev. B, **85**, 245205, 2012.

8. C. Tanase, E. J. Meijer, P. W. M. Blom, and D. M. de Leeuw, *Unification of the Hole Transport in Polymeric Field-Effect Transistors and Light-Emitting Diodes*, Phys. Rev. Lett., **91**, no. 21, 216601, 2003.
9. J. Cottaar and P. A. Bobbert, *Calculating charge-carrier mobilities in disordered semiconducting polymers: Mean field and beyond*, Phys. Rev. B, **74**, no. 11, 115204, 2006.
10. A. Lukyanov and D. Andrienko, *Extracting nondispersive charge carrier mobilities of organic semiconductors from simulations of small systems*, Phys. Rev. B, **82**, 193202, 2010.
11. Z. G. Yu, D. L. Smith, A. Saxena, R. L. Martin, and A. R. Bishop, *Molecular geometry fluctuations and field-dependent mobility in conjugated polymers*, Phys. Rev. B, **63**, no. 8, 085202, 2001.
12. J. J. M. van der Holst, M. A. Uijtewaald, B. Ramachandhran, R. Coehoorn, P. A. Bobbert, G. A. de Wijs, and R. A. de Groot, *Modeling and analysis of the three-dimensional current density in sandwich-type single-carrier devices of disordered organic semiconductors*, Phys. Rev. B, **79**, no. 8, 085203, 2009.
13. Z. G. Yu, D. L. Smith, A. Saxena, R. L. Martin, and A. R. Bishop, *Molecular Geometry Fluctuation Model for the Mobility of Conjugated Polymers*, Phys. Rev. Lett., **84**, 721–724, 2000.
14. W. F. Pasveer, J. Cottaar, C. Tanase, R. Coehoorn, P. A. Bobbert, P. W. M. Blom, D. M. de Leeuw, and M. A. J. Michels, *Unified Description of Charge-Carrier Mobilities in Disordered Semiconducting Polymers*, Phys. Rev. Lett., **94**, no. 20, 206601, 2005.
15. F. Jansson, S. D. Baranovskii, F. Gebhard, and R. Österbacka, *Effective temperature for hopping transport in a Gaussian density of states*, Phys. Rev. B, **77**, no. 19, 195211, 2008.
16. M. Bouhassoune, S. L. M. van Mensfoort, P. A. Bobbert, and R. Coehoorn, *Carrier-density and field-dependent charge-carrier mobility in organic semiconductors with correlated Gaussian disorder*, Org. Elec., **10**, no. 3, 437, 2009.
17. J. Cottaar, *Modeling of charge-transport processes for predictive simulation of OLEDs (PhD thesis)*, Technische Universiteit Eindhoven, Eindhoven, 2012.
18. Y. Roichman and N. Tessler, *Generalized Einstein relation for disordered semiconductors - implications for device performance*, Appl. Phys. Lett., **80**, no. 11, 1948, 2002.
19. J. J. M. van der Holst, F. W. A. van Oost, R. Coehoorn, and P. A. Bobbert, *Monte Carlo study of charge transport in organic sandwich-type single-carrier devices: Effects of Coulomb interactions*, Phys. Rev. B, **83**, no. 8, 085206, 2011.
20. S. L. M. van Mensfoort, S. I. E. Vulto, R. A. J. Janssen, and R. Coehoorn, *Hole transport in polyfluorene-based sandwich-type devices: Quantitative analysis of the role of energetic disorder*, Phys. Rev. B, **78**, no. 8, 085208, 2008.
21. A. Many and G. Rakavy, *Theory of transient space-charge limited currents in solids in the presence of trapping*, Phys. Rev., **126**, 1980–1988, 1962.
22. Murat Mesta, Marco Carvelli, Rein J. de Vries, Harm van Eersel, Jeroen J. M. van der Holst, Matthias Schober, Mauro Furno, Björn Lüssem, Karl Leo, Peter Loeb, Reinder Coehoorn, and Peter A. Bobbert, *Molecular-scale simulation of electroluminescence in a multilayer white organic light-emitting diode*, Nature Mater., **12**, 653, 2013.
23. S. L. M. van Mensfoort, J. Billen, S. I. E. Vulto, R. A. J. Janssen, and R. Coehoorn,

Electron transport in polyfluorene-based sandwich-type devices: Quantitative analysis of the effects of disorder and electron traps, Phys. Rev. B, **80**, 033202, 2009.

24. S. L. M. van Mensfoort, V. Shabro, R. J. de Vries, R. A. J. Janssen, and R. Coehoorn, *Hole transport in the organic small molecule material α -NPD: evidence for the presence of correlated disorder*, J. Appl. Phys., **107**, no. 11, 113710, jun 2010.
25. S. L. M. van Mensfoort, R. J. de Vries, V. Shabro, H. P. Loeb, R. A. J. Janssen, and R. Coehoorn, *Electron transport in the organic small-molecule material BAQ — the role of correlated disorder and traps*, Org. Elec., **11**, no. 8, 1408–1413, 2010.
26. Scott Kirkpatrick, *Percolation and Conduction*, Rev. Mod. Phys., **45**, no. 4, 574, 1973.
27. Vinay Ambegaokar, B. I. Halperin, and J. S. Langer, *Hopping Conductivity in Disordered Systems*, Phys. Rev. B, **4**, no. 8, 2612, 1971.
28. B. I. Shklovskii and A. L. Efros, *Electronic properties of doped semiconductors*, Springer-Verlag, Berlin, 1984.
29. S. D. Baranovskii, O. Rubel, and P. Thomas, *Theoretical description of hopping transport in disordered materials*, Thin Solid Films, **487**, no. 1-2, 2, 2005.
30. R. Coehoorn, W. F. Pasveer, P. A. Bobbert, and M. A. J. Michels, *Charge-carrier concentration dependence of the hopping mobility in organic materials with Gaussian disorder*, Phys. Rev. B, **72**, no. 15, 155206, 2005.
31. Jeppe C. Dyre and Thomas B. Schröder, *Universality of ac conduction in disordered solids*, Rev. Mod. Phys., **72**, no. 3, 873, 2000.
32. S. Tyč and B. I. Halperin, Phys. Rev. B, **39**, 877, 1989.
33. J. Cottaar, L. J. A. Koster, R. Coehoorn, and P. A. Bobbert, *Scaling Theory for Percolative Charge Transport in Disordered Molecular Semiconductors*, Phys. Rev. Lett., **107**, 136601, 2011.
34. E. W. Dijkstra, *A note on two problems in connexion with graphs*, Numerische Mathematik, **1**, 269–271, 1959.
35. J. Williams, Comm. Assoc. Comput. Mach., **7**, 347, 1964.
36. I. I. Fishchuk, V. I. Arkhipov, A. Kadashchuk, P. Heremans, and H. Bässler, *Analytic model of hopping mobility at large charge carrier concentrations in disordered organic semiconductors: Polarons versus bare charge carriers*, Phys. Rev. B, **76**, no. 4, 045210, 2007.
37. S.L.M. van Mensfoort, M. Carvelli, M. Megens, D. Wehenkel, M. Bartyzel, H. Greiner, R. A. J. Janssen, and R. Coehoorn, *Measuring the light emission profile in organic light-emitting diodes with nanometre spatial resolution*, Nat. Phot., **4**, 329, 2010.
38. J. J. Kiatkowski, J. Nelson, H. Li, J. L. Bredas, W. Wenzel, and C. Lennartz, *Simulating charge transport in tris(8-hydroxyquinoline) aluminium (Alq3)*, Phys. Chem. Chem. Phys., **10**, 1852, 2008.
39. V. Rühle, A. Lukyanov, F. May, M. Schrader, T. Vehoff, J. Kirkpatrick, B. Baumeier, and D. Andrienko, *Microscopic Simulations of Charge Transport in Disordered Organic Semiconductors*, J. Chem. Theor. Comp., **7**, no. 10, 3335–3345, 2011.

Multiscale Modeling Methods for Electrochemical Energy Conversion and Storage

Alejandro A. Franco^{1,2}

¹ Laboratoire de Réactivité et de Chimie des Solides (LRCS)
Université de Picardie Jules Verne & CNRS, UMR 7314 - Amiens, France
E-mail: alejandro.franco@u-picardie.fr

² Réseau sur le Stockage Electrochimique de l'Energie (RS2E), FR CNRS 3459, France
E-mail: a.a.franco.electrochemistry@gmail.com

Energy conversion and storage through electrochemical devices, such as fuel cells and batteries, are called to play an important role for the development of future sustainable energy networks. With the impressive progress reached by the computational facilities in the recent past years, physical modeling and numerical simulation start nowadays to be recognized as crucial tools for the understanding-based, and thus controllable-based, development of efficient, stable and inexpensive energy conversion and storage materials and components, and the optimization of the operation conditions at the device level.

This tutorial comprehensively covers both theoretical and practical aspects of multiscale modeling of electrochemical power generators.

1 Introduction

1.1 Towards a sustainable energy conversion and storage: the promising high-tech aura of the electrochemical power generators

With the modern times, the humanity entered into an existential crisis arising from multiple factors, among them

- the heterogeneity of the distribution of natural resources between the countries;
- the increase of the global population with the consequent increasing demand for energy;
- the strong dependence of the countries on fossil fuels and the consequent international economic and political tensions;
- the global warming, the consequent climate and geographical changes and the population migration.

As the common factor besides these problems is the availability of energy, this encourages scientists and industrialists to invest in innovative technology for new energies production (or conversion from natural resources) and storage. Over the last decades, great efforts have been deployed on the common quest for an alternative to the depleting natural sources of fossil fuels. For the design of new technologies, the quantity of energy that can be produced, the cost, and the impact on the environment (especially the quantity of emission of CO₂) are of major concerns. One expects such a technology to be competitive over the

whole range of applications: transportation, stationary power generation (for residence, public buildings), and portable applications (like mobile phones, portable computers, auxiliary power unit in cars, etc.).

Within the spectrum of power generators suitable in a sustainable energetic network, electrochemical devices for energy conversion and storage are called to play a very important role in the future. These technologies present a great potential to become cost-competitive (because they can be applied to nomad systems), highly efficient (because energy can be produced and stored at room temperature), and environmentally benign (because of the zero-emissions and of the no noise).

An excellent example of this follows from modern electronic equipment and electric vehicle applications which have been rapidly developing, resulting in a growing demand for high energy density power sources such as rechargeable lithium ion batteries (LIBs).

Among the large diversity of electrochemical power generators (EPGs) under study within the scientific and industrial communities, we will focus here only on some of the most attractive ones because of their application potentialities and because of the remaining challenging but scientifically exciting technical issues to be overcome, for instance

- Hydrogen-feed Polymer Electrolyte Membrane Fuel Cells (PEMFCs) and Polymer Electrolyte Membrane Water Electrolysers (PEMWEs) for energy conversion;
- Electrochemical capacitors – known also as supercapacitors – (ECs), LIBs and Lithium Air Batteries (LABs) for energy storage.

The general operation principles of such devices are presented in Figure 1, except for PEMWEs which actually operate in the reverse way than PEMFCs, and the typical specific power and energy densities which can be delivered by some of these technologies is presented in Figure 2.

Porous electrodes are the pivotal components of modern PEMFCs, PEMWEs, ECs, LIBs and LABs. Porous electrodes are inherently multiscale systems as they are made of multiple coexisting materials, each of them ensuring a specific function in their operation. Such electrodes structural complexity has been historically driven by the needs of reducing the device cost and of enhancing its efficiency, stability and safety. The use of nano-engineered materials and chemical additives (in the case of batteries) allowed a significant progress toward these goals.

For instance, in the case of modern PEMFCs, the electrodes are constituted by metallic nanoparticles of few nanometers size having the role of electrocatalyst, and are supported on carbon particles of few microns size having the role of electronic conductor. The resulting complex structure, arising from more than 30 years of research efforts to enhance the efficiency and reducing the loading by precious metals in these devices², is in turn embedded within perfluorosulfonic acid (PFSA) proton conducting polymers, the Nafion[®] ionomer from Dupont being the most used one, arising into a composite electrode of few micrometers thick (Figure 3).

During the PEMFC operation, a strong non-linear multi-scale dynamical coupling between several physicochemical phenomena takes place within the Membrane-Electrodes Assembly -MEA-: reactant transfers (hydrogen and oxygen through the Gas Diffusion Layer -GDL- and Catalyst Layer -CL- pore phases), water transfers (bi-phasic water in CL and GDL meso/macro-pores, dissolved water in the Polymer Electrolyte Membrane -PEM- and CL ionomer), electrochemistry (hydrogen oxidation producing electrons and protons,

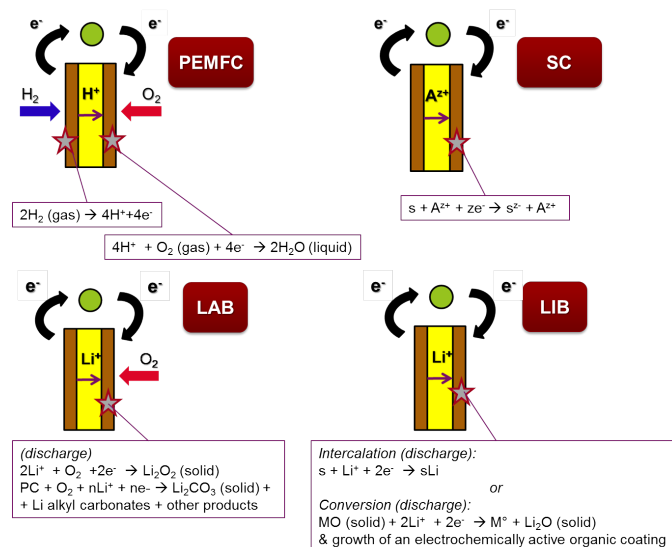


Figure 1. Operation principles of PEMFCs, SCs, LIBs and LABs.

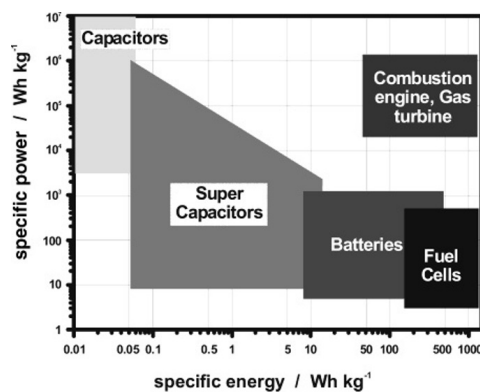


Figure 2. Typical specific power vs. specific energy for different electrochemical power generators and comparison with combustion engines. Source: Ref. 1.

and oxygen reduction producing water), and charge transfer (proton within the CL ionomer and PEM, electron within the CL and GDL).

In fact, processes at the smaller scales (e.g. Oxygen Reduction Reaction -ORR- on the cathode platinum nanoparticles) dominate the processes at the larger scales (e.g. liquid water transport through the cathode carbon support secondary pores) which in turn affect the processes at the smaller ones (e.g. through the water flooding limiting O_2 transport in the cathode). PEMFC technologies have not yet reached all the required characteristics to be competitive, in particular regarding their high cost and their low durability ^{4, 5}.

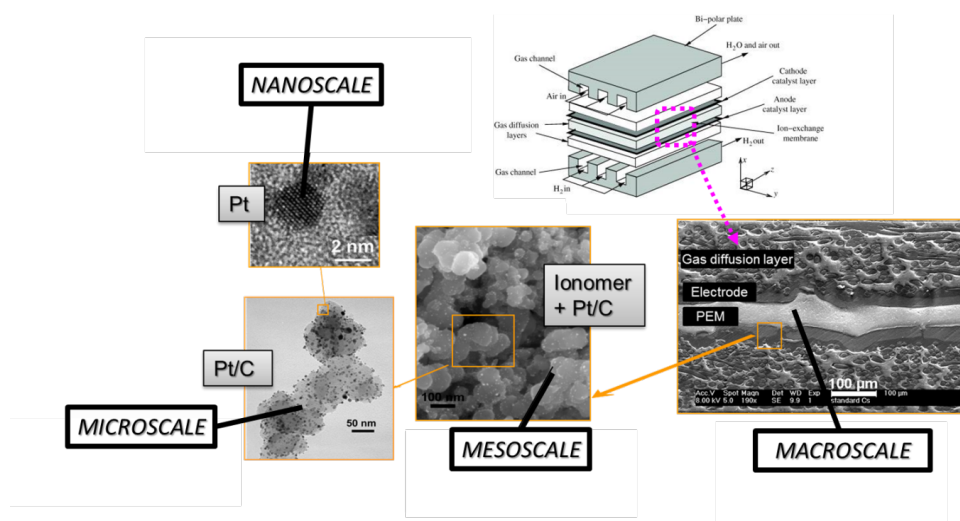


Figure 3. Multiscale structure of a PEMFC electrode. Source: ³.

In addition to the electrochemical reactions, reactants and biphasic water transport, other mechanisms limiting optimal platinum utilization are charge transfer, thermo-mechanical stresses and irreversible materials degradation. For instance, microstructural degradation leading to the PEMFC components aging is attributed to several complex physicochemical phenomena not yet completely understood:

- dissolution and redistribution of the catalyst: mainly due to the high potentials of the cathode electrode ⁶. This phenomenon reduces the specific catalyst surface area leading to the loss of the electrochemical activity ^{7,8,9,10};
- corrosion of the catalyst carbon-support: carbon is thermodynamically unstable at typical cathode operating conditions. Furthermore, carbon degrades more rapidly during transient startup and shut-down conditions and high humidification levels. Indeed, oxygen permeation locally increases the cathode potentials accelerating the cathode damage ¹¹;
- loss or decrease of the hydrophobicity: caused by an alteration of the PTFE ¹², which is used to give hydrophobic properties to the CLs as well as to the GDLs and the Micro Porous Layers (MPLs). This affects the water management in the cell and thus the electrochemical performance;
- apart from mechanical degradations such as thinning and pinhole formations, chemical and electrochemical degradations could also take place in PerFluoroSulfonated Acid (PFSA) PEM. Hydrogen peroxide (H_2O_2) and radical species can be formed in the CLs. H_2O_2 , a highly oxidative reagent, may deteriorate ionomer in the MEA, but the mechanism is not yet fully understood ¹³. Furthermore, PEM degradation facilitates reactants cross-over between the CLs, and hence the performance and durability decay.

These spatiotemporal nano/microstructural changes translate into irreversible long-term cell power degradation. Moreover, the ways of how aging mechanisms occur are expected to be strongly sensitive to the PEMFC operation mode. Understanding the relationship between operation mode and degradation mode remains a challenging task. The PEMFC response can be even more complex if the reactants are contaminated with external pollutants (e.g. in the anode: CO from hydrocarbons reforming fabricating H₂; cathode: NO₂ or SO₂ from air)^{14,15}. The competitions and synergies between all these “non-aging” and “aging” mechanisms determine the effective instantaneous electrochemical performance and durability of the cell. Present life time of PEMFC under automotive solicitations rarely exceeds 1000 hours. A maximum rate of potential degradation from 2 to 10 μ V/hour with less than 10% power global decay for 5000 operational hours are required for automotive applications¹⁶.

As PEMFCs work with hydrogen, hydrogen needs to be fabricated. A promising device for the production of pure hydrogen from renewable energy sources is the PEMWE¹⁷. Although PEM technology was introduced in the 1960s¹⁸, PEMWE has started to receive more attention from the scientific community, because of the problems cited above, only from the mid of the 1990s and the beginning of the 2000s¹⁹. Such a device, in comparison with alkaline and high temperature solid oxide electrolyzers, offers several advantages including ecological cleanness, higher efficiency from both current density and energy, compactness and low temperature operation^{20,21}.

The overall decomposition reaction of water into oxygen and hydrogen taking place in a PEMWE is $\text{H}_2\text{O} \rightarrow \frac{1}{2}\text{O}_2 + \text{H}_2$. The hydrogen evolution reaction (HER), $4\text{H}^+ + 4\text{e}^- \rightarrow 2\text{H}_2$, takes place at the cathode side and the oxygen evolution reaction (OER), $2\text{H}_2\text{O} \rightarrow \text{O}_2 + 4\text{H}^+ + 4\text{e}^-$, at the anode side. One of the main drawbacks of the PEMWE is that the electrodes are based on expensive precious-metal-based catalysts. In the cathode side, platinum nanoparticles supported on percolated carbon nanoparticles are used (as in PEMFC electrodes) while in the anode side, rutile oxides like IrO₂ and RuO₂ are currently used, having the role of both catalyst and electronic support. Proton conduction within and between the electrodes is ensured by Nafion®-like polymers. PEMWEs present also substantial technical challenges related to their efficiency and lifetime. For example the catalyst oxidation leads to an evolution of the catalyst layers microstructure properties²².

To enhance the performance and durability of the PEMWE, a deep understanding of the physicochemical processes related to the nano and microstructural properties of the electrodes is crucial. With this aim, the complex mechanisms occurring at multiple scales in PEMWE operation (Figure 4), have been extensively studied from different experimental approaches: (i) synthesis and characterization of the different materials used as electrocatalysts²³; (ii) investigation of OER and HER mechanisms^{24,25,26}; (iii) influence of MEA components on the global electrochemical measurements^{27,28}; (iv) the water flow effect on the performance²⁹; (v) stack development; (vi) Nafion® properties in the PEMWE environment³⁰.

LIBs are also multiscale and multiphysics systems, as illustrated in Figure 5 for a positive electrode made of LiFePO₄, one of the most popular electrode materials because of its high power density and safety among a large variety of compounds developed since the 1970s^{32,33}. In this case, non-porous LiFePO₄ crystals represent the smallest scale observed, whereas the porous agglomerates represent the second size scale. The third size scale is the positive electrode itself, which consists of carbon black, binder and the porous

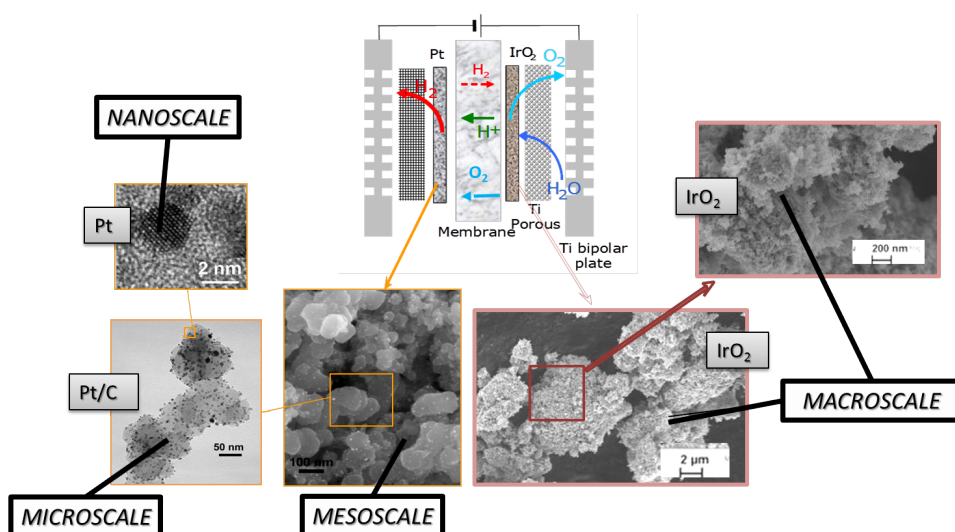


Figure 4. Multiscale structure of the PEMWE electrodes. Source: adapted from ³¹.

agglomerates of the LiFePO_4 particles ³⁴. Carbon black is usually added to reduce the formation of agglomerates during the synthesis process and to enhance the electronic conductivity properties ³⁵. The binder is usually made of Polyvinylidene difluoride (PVDF) polymer, which provides to the electrode an aspect of “polymer composite”. The binder spatial distribution depends on the electrode preparation method, coating and drying process, and surface properties of each compound, such as the active material. In early work without careful morphology control, considerable crystal agglomeration occurred ³⁶, and even in more recent materials, agglomerated particles are often still present ^{37,38,39}.

LIB negative electrodes are typically made of carbonaceous materials and also present a complex multiscale structure. The application of carbonaceous materials instead of Li-metal has several advantages such as better cycle-life and reliability preventing severe degradation problems such as Li dendrites formation during cycling. Carbon has in fact the ability to reversibly absorb and release large quantities of lithium ($\text{Li:C}=1:6$) without altering the mechanical and electrical properties of the material. On the first charge of the battery a polymeric layer, the so-called *solid electrolyte interphase* (SEI), forms from the electrolyte decomposition. This “passivation” or “protective” layer is of crucial importance for the battery operation in terms of safety as it prevents the carbon from reacting with the electrolyte and helps on avoiding graphite exfoliation ⁴⁰.

Moreover, several materials have also been proposed as alternatives to replace graphite in the negative electrode, also showing more or less a multiscale structure ⁴¹.

As the electrodes structure of rechargeable LIBs can be seen as a complex ensemble of lithium sources and sinks embedded in an electrolyte medium, the rate-determining processes during charge and discharge will depend on the Li^+ concentration on the (e.g. intercalation, conversion) electrode active material surface, Li^+ concentration in the electrolyte, the potential drop between the active material and the electrolyte and the lithium

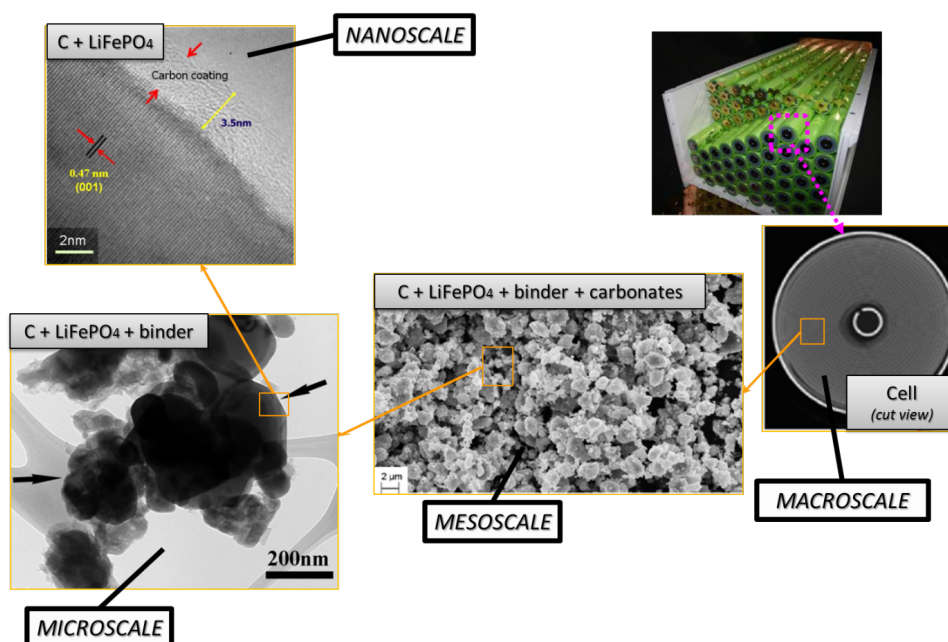


Figure 5. Multiscale structure of an intercalation LIB electrode. Source: ⁴².

concentration inside the active material ^{43,44,45}. The LIB operation may thus be limited by Li^+ transport in the electrolyte, lithium transport in the electrode material or by the ionic or electronic conductivity of the electrolyte or electrodes.

Electrochemical reaction of lithium intercalation and/or conversion takes place on a nanometer scale and strongly depends on the chemistry and on the nano- and microstructural properties of the intercalation/conversion material. Charge transport, heat transport and mechanical stresses take place from the material level up to the cell level and also depend on the materials and components structural properties. Time scales vary from sub-nanoseconds (electrochemical reactions) over seconds (transport) up to days or even months (structural and chemical degradation).

LABs are one type of metal air batteries (with metals such as Zn ⁴⁶, Na ⁴⁷, Mg ⁴⁸ and Al ⁴⁹) which are receiving a growing interest as they theoretically achieve a specific energy significantly higher than current lithium-ion batteries with two intercalation electrodes ⁵⁰.

LABs are conceptually a mix between PEMFCs and LIBs (Figure 6). Abraham and Jiang were the first on reporting a practical LAB with the use of a Li/C cell in which a gel polymer electrolyte membrane served as both the separator and the ion-transporting medium ^{51,52}. Their cell theoretical specific energy was of up to $\sim 3400 \text{ Wh}\cdot\text{kg}^{-1}$. The reason of such high specific energies is that the positive electrode active material, *i.e.* oxygen, is not stored internally in the battery. Oxygen actually enters a porous carbon electrode from air for the ORR, as a similar functional process to what one has in PEMFC cathodes. Lithium and oxygen then react to form metal oxides during the discharge process. During the charge process, the oxides decompose to release lithium ions and oxygen again.

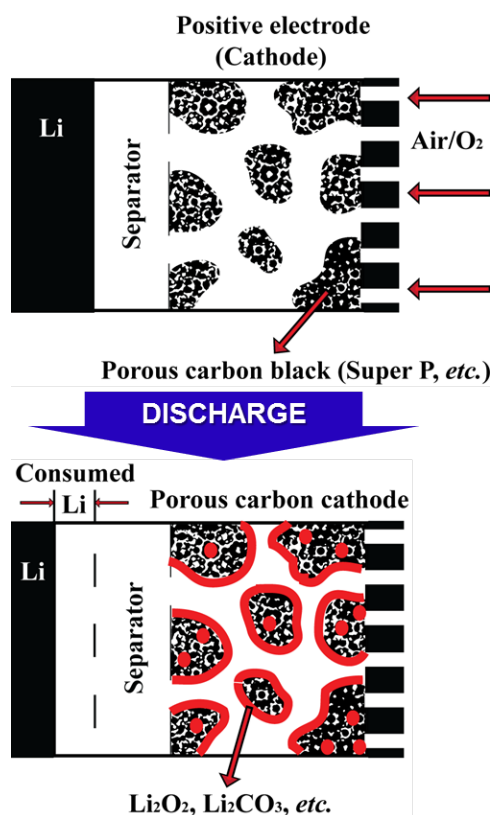


Figure 6. Schematics of a LAB discharge process.

Abraham and Jiang's LAB was actually the first non-aqueous LAB. In modern non-aqueous LABs the electrolyte is typically made of lithium salts (*e.g.* LiPF_6) mixed with carbonate-based solvents such as propylene carbonate (PC), ethylene carbonate (EC) and dimethyl carbonate (DMC), and the carbon electrode can support or not catalyst nanoparticles (*e.g.* RuO_2 , Pt, Au, or MnO_2)⁵³.

The performance of LABs has been reported to be affected by many factors such as the air relative humidity, the oxygen partial pressure⁵⁴, the choice of catalysts,⁵⁵ the electrolyte composition,⁵⁶ the micro- to nanostructure of carbonaceous materials, the macrostructure of the positive electrode,^{57,58} and the overall cell designs⁵⁹.

In practice, LABs suffer from poor cyclability (up to few cycles) and reversibility between the discharge and charge (with discharge voltages around 2.5-3.0 V and charging voltages around 4.0-4.5 V)^{60,61,62,63}. Typical LAB capacity fades twice as fast after 50 cycles (compared to 25% capacity fade after 300 cycles for ordinary LIBs). The high positive electrode polarization (sharp voltage drop-off with increasing current) is frequently believed to be due to the oxygen diffusion limitations. Recent studies have also identified that a possible cause of the high-voltage hysteresis is due to side reactions of the electrolyte with the discharge product of the ORR, Li_2O_2 , which can form lithium carbonate

and lithium alkyl carbonates with the carbonate species in the electrolyte^{64,65,66}. These side reactions are believed to deplete the electrolyte during cycling, limiting the reversibility of LABs.

Moreover, O₂ reduction products are mostly insoluble in non-aqueous electrolytes. They precipitate on the surface of the porous carbon electrode^{67,68}. This ultimately hinders the discharge reaction and also leads to a lower specific capacity than the theoretical value.⁶⁹

Analogies between discharge in LABs and water generation in PEMFC operation can be done on several aspects. The impact of pore clogging on O₂ transport in LAB positive electrodes, can be within some extent assimilated to the impact of liquid water on O₂ transport in PEMFC cathodes.⁷⁰

Furthermore, pore clogging by solid oxides in LABs is unfavorable to Li⁺ transport whereas pore clogging by liquid water is favorable to H⁺ transport in PEMFCs.

ECs have a greater power density and a longer cycle life than batteries do, and a higher energy density than that of conventional capacitors^{71,72}; therefore, they have attracted a lot of research attention in recent years^{73,74,75}.

The storage mechanism in ECs consists mainly of two types of processes, a purely capacitive and a pseudo-capacitive process. The former is based on the electric charge separation at the electrode/electrolyte interface (double layer), the latter on electrochemical reactions occurring on the electrodes (faradaic process). In the later, the electrode material is electrochemically active, e.g. metal oxides, which can directly store charges during the charging and discharging processes^{76,77}.

In ECs, the capacitance performance exhibited by the devices is strongly dependent on the nature of the electrode/electrolyte interface (Figure 7). Generally, the larger the specific surface area of carbon in the electrodes, the higher the capability of accumulation of electric charges at the interface, and thus the higher the capacitance. However, high surface area is not a sufficient condition to achieve high capacitance; the carbon must also contain a large fraction of mesopores. The charge (discharge) mechanism in an EC must involve an easy access of electrolyte into the carbon pores, possible only in the presence of macro- and mesopores, which allow a high rate of charge and discharge, to obtain a large amount of electric charge. A typical supercapacitor has two electrodes, made of high surface area carbon, and an aqueous or non-aqueous electrolyte with a porous separator between them. In most commercial supercapacitors, tetraethylammonium tetrafluoroborate in acetonitrile or propylene carbonate is used as the organic electrolyte^{78,79}, while in others sulfuric acid or potassium hydroxide is used as the aqueous electrolyte⁸⁰.

One of the key challenges for ECs is their limited energy density, which has hindered their wider application in the field of energy storage. To overcome this challenge, a major focus of ECs research and development should be to discover new electrode materials with high capacitance and a wide potential window. In the design of EC electrode materials, properties to be favored by the research efforts include:

- high specific surface area (i.e. large amount of active sites);
- suitable PSD, pore network, and pore length for facilitating the diffusion of ions at a high rate;
- low internal electrical resistance for efficient charge transport in the composite electrode;

- good electrochemical and mechanical stability for good cycling performance.

Nano-micropores are necessary to achieve higher specific surface area, and these micropores must be ensured to be electrochemically accessible for ions. Hence, pore network, the availability and wettability of pores, with dimensions matching the size of solvated anions and cations are crucial aspects to be considered in the design of EC electrode materials.

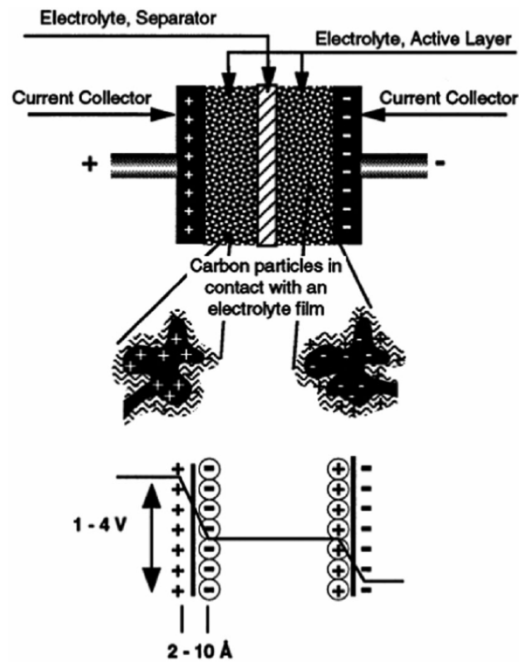


Figure 7. Schematics of a non-faradaic EC. Source: ⁸¹.

In spite of excellent technological prospects in all these electrochemical technologies, commercialization of advanced electrochemical devices for power generation in transportation, electronics, and stationary applications is far from being guaranteed. For the commercialization of such technologies, the concomitant reduction in cost of the materials, efficiency and their stability would be the decisive breakthrough. Critical progress hinges on new concepts in the design of advanced materials as well as fundamental understanding of basic electrochemical processes.

1.2 Lecture objectives

From the discussions above, it is obvious that EPGs are multiphase systems as they involve at least liquids (electrolytes) and solids (electrochemically active surfaces), and sometimes gas (case of PEMFCs). EPGs are also multiphysics systems as they involve multiple competing mechanisms behind their operation principles, such as electrochemistry, ionic and

liquid transport (e.g. water in the case of PEMFCs), mechanical stresses and heat management. All these mechanisms are strongly and nonlinearly coupled over the various scales, and thus processes at the nano- and microscale can therefore dominantly influence the macroscopic behavior. For example, the materials spatiotemporal microstructural changes leads into irreversible long-term cell power degradation, and the ways of how aging mechanisms occur are expected to be strongly sensitive to the cell operation mode. For instance, understanding the relationship between operation mode and degradation mode remains a challenging task.

In consequence, because of the structural complexity and multiphysics character of modern electrochemical devices for energy conversion and storage, interpretation of experimental observations and ultimate cell optimization remain a challenge. An analysis through a consistent multiscale physical modeling approach is required to elucidate the efficiency limitations and their location, the degradation and failure mechanisms.

From a practical point of view, it is crucial to accurately predict their performance, state-of-health and remaining lifetime. For that purpose, it is necessary to develop diagnostic schemes that can evaluate electrochemical cell performance and state-of-health adequately. In order to achieve this, several steps are required:

- to develop via physical modeling a better understanding of several individual processes in the cell components;
- to understand the interplay between individual scales over the spatiotemporal hierarchies with their possible competitive or synergetic behavior;
- to identify the contribution of each mechanism into the global cell response under dynamic conditions;
- to design separated controllers for an online control of the EPGs behavior to enhance its durability under specific operation conditions (e.g. by controlling the dynamics of the alcohol fuel, the temperature, etc.).

A detailed understanding of the relevant processes on all these materials and components scales is required for a physical-based optimization of the electrochemical cell design regarding its efficiency, durability and safety.

Based on this, this lecture has two major objectives:

- first, providing conceptual and epistemological tools aiming to help the EPG modelers in their approaches choice and in their communication with experimentalists for model validation;
- then, providing practical tools to develop multiscale models for electrochemical devices for energy conversion and storage. Only some basic features are provided here but the readers are strongly encouraged to explore other features with the help of the large database of bibliographic references provided by the author.

At the end of the lecture, remaining main challenges in the multiscale modeling of EPGs are also discussed

2 Modeling Experiments and Experimenting Models

The word “modeling” is inherently connected with the concept of “theory”. But...what is theory exactly? Theory can be defined as a “contemplative and rational type of abstract or generalizing thinking”, or “the results of such thinking”. One can develop theories for example within a large diversity of disciplines, such as philosophy or physics. Physical theories aim to correlate diverse experimental observations (Figure 8, top).

A mathematical model is a transcription of a physical theory describing a system into mathematical concepts and language (Figure 8, bottom). The process of developing a mathematical model is termed mathematical modeling. Thus, a theory does not necessarily translate as a mathematical model. As mathematics is the most logical and organized way of thinking, it becomes natural using it when one wants to rationalize and to predict the behavior of physical systems.

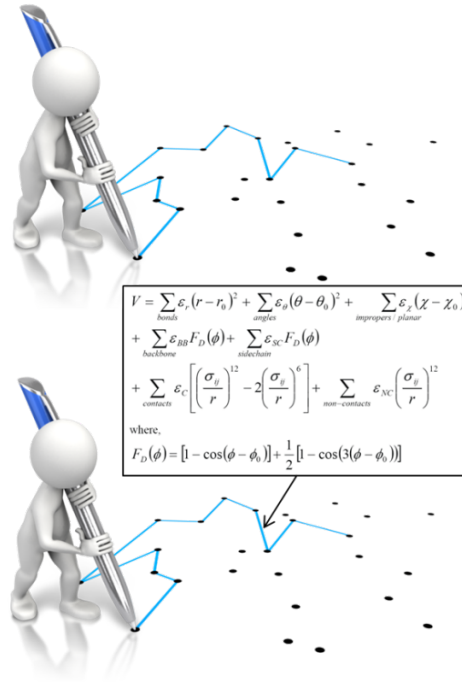


Figure 8. Theory vs. mathematical model.

However, mathematical models are always imperfect. According to the mathematician Kurth Gödel⁸² any formal system which contains arithmetics (i.e. the mathematics of whole numbers) is *incomplete*. By *incomplete* is meant, that the system contains *undecidable* statements, i.e. statements which are neither provable nor disprovable (by means of the system itself).

The *consistency* of such a system implies the existence of undecidable statements (first incompleteness theorem) and that the consistency of the system itself is undecidable (sec-

ond incompleteness theorem). By *consistency* is meant, that it is excluded to prove a statement together with its negation.

From this it arises that all mathematical models are essentially incomplete, and thus will never represent perfectly the physical system.

To enhance the representation by models of the physical system, comparison between the modeling outcomes and experimental data is crucial, but still by keeping in mind the following important premises:

- theory without experiments is just “speculation”: theory needs experiments for validation;
- experiments without theory becomes just a “trial/error” method: experiments need theory as guideline;
- both theorists and experimenters are “simultaneously theorists and experimenters”. For instance, an experimenter, as a theorist, search on isolating a part of the “real system” to study some specific mechanism or set of mechanisms at the lab scale. Furthermore, experimenters always use implicitly or explicitly theories or mathematical models for the interpretation and for the report of their experimental data;
- Most theorists and experimenters do not speak the same language (or do not use the same *theories*...): this is a barrier usually making difficult the research. However, a theorist with a good understanding of experiment, or an experimenter with a good understanding of theory, will progress more efficiently than others who do not have this double profile.

2.1 The modeling method

The current scientific method used in the modeling discipline can be schematized as the process in Figure 9. In this process, the first and second steps consist respectively of defining the physical problem (i.e. the system which will be modeled: the electrode alone? the complete cell? an active particle?...etc.) and on identifying the observables one would intend to simulate with the model (e.g. electrode potential? cell potential? active area evolution?...etc.). Then, the third and fourth steps consist respectively of defining the structural model which will be used (i.e. the geometrical assumptions: e.g. 1D, 2D or fully 3D representation of the electrode?) and the physics to be treated (e.g. detailed electrochemistry? ions transport? both coupled?). These two steps are crucial as they strongly determine the choice of the simulation approach in step five (e.g. Quantum Mechanics? Molecular Dynamics? Kinetic Monte Carlo? Coarse Grain Molecular Dynamics, Continuum Fluid Dynamics? or combination of several of such approaches?). Then this determines the choice of the mathematical formulations for performing calculations, of the appropriate numerical algorithms, software and hardware (e.g. parallel computing or not?) to proceed with the simulation of the observables. These observables are then compared with the available experimental data, preferentially obtained with model experiments, i.e. experiments designed to be representative of the model (e.g. geometrical) assumptions. This comparison will allow the model validation or its improvement in terms of its structural/geometrical assumptions or physics accounted for. After several iterations between theory and experiment one could expect achieving on “producing” a model with predictive capabilities of

the electrochemical device operation. The key step in this modeling process, is the choice of the modeling simulation approach among three categories: multiphysics, multiscale and multiparadigm modeling approaches, which are discussed below.

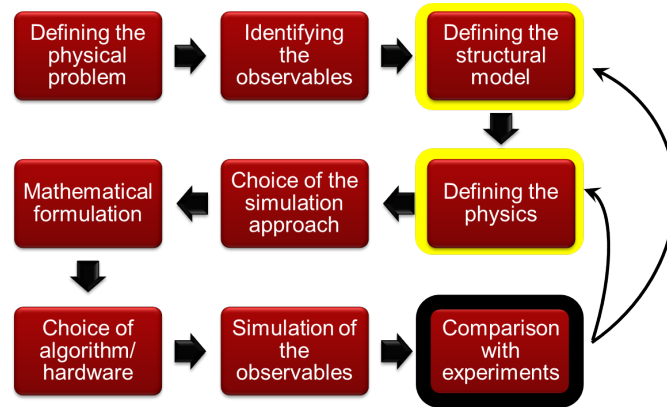


Figure 9. The modeling method.

2.2 Multiphysics, multiscale and multiparadigm models: definitions

The word “multiphysics” usually characterizes, in published literature, models that mathematically describe the interplaying of mechanisms belonging to different physical domains. Multiphysics models include models describing these multiple mechanisms within a single and unique spatial scale (e.g. a model describing the impact of heating on the mechanical stress of a material)⁸³. Multiphysics models can be by construction “multiscale” on time, as they can be built on the basis of mathematical descriptions of multiple mechanisms with different characteristic times (e.g. when the heat dissipation time constant is different to the material deformation time constant). The majority of models developed to describe the operation of EPGs fall within this category as they have to consider as least two different physical domains: the electrochemistry and the charge transport. For example, several groups have developed various rigorous LIB models based on the porous electrode theory, coupled with concentration solution theory and modified Ohm’s law, which allow treating thermal, mechanical and capacity fade mechanism aspects, as discussed later in this lecture.

“Multiscale models” typically refer to models accounting for mathematical descriptions of mechanisms taking place at different spatial scales⁸⁴. Multiscale models aim, by construction, to considerably reduce empirical assumptions than can be done in simple multiphysics models. This is because they explicitly describe mechanisms in scales neglected in the simple multiphysics model. Actually, multiscale models have a hierarchical structure: that means that solution variables defined in a lower hierarchy domain have finer spatial resolution than those solved in a higher hierarchy domain. Consequently, physical and chemical quantities of smaller length-scale physics are evaluated with a finer spatial

resolution to resolve the impact of the corresponding small-scale geometry. Larger-scale quantities are in turn calculated with coarser spatial resolution, homogenising the (possibly complex) smaller-scale geometric features.

A large diversity of multiscale models exists in numerous domains, such as climate science, geology, nuclear energy and physical chemistry^{85,86,87}. In the case of EPGs, model geometry decoupling and domain separation for the physicochemical process interplay are valid where the characteristic time or length scale is “segregated”. Assuming statistical homogeneity for repeated architectures typical of EPGs devices is often adequate and effective for modeling submodel geometries and physics in each domain. For example, the so called *multiple-scale technique*^{88,89,90} provides a systematic way for accounting for the EPGs mechanisms which occur within a microscopic quasi-periodic microstructure in terms of a macroscopic system of equations, as used in^{91,92,93,94,95}. Such method allows deriving the macroscopic equations and determining the corresponding parameters from a local problem for the microscopic behavior. Model coefficients are calibrated in terms of the microstructure, and thereby provide a tool for improving EPGs particles design. This approach can be contrasted with averaging methods as for example in⁹⁶ where spatial averages are taken of the microscopic equations resulting in equations on a macroscopic scale for the microscopically averaged variables: these macroscopic equations are closed by making *ad hoc* assumptions about the mathematical closure conditions and fitting these to empirical data.

Depending on the development context of these models (engineer or physicist based), they would be built following top-down or bottom-up viewpoints. Top-down models connect detailed macroscopic descriptions of mechanisms with global parameters representing microscopic mechanisms. On the other hand, bottom-up cell models scale up detailed descriptions of microscopic mechanisms onto global parameters to be used in macroscopic models. It is important to develop approaches which synergistically combine these two complementary views, as the former provides “a closer” comparison with macroscopic experiments, and the latter predictability towards the materials chemical and structural properties (Figure 10).

Finally, the mathematical descriptions in a multiscale model can be part of a single simulation paradigm (e.g. only continuum) or of a combination of different simulation paradigms (e.g. stochastic model describing a surface reaction coupled with a continuum description of reactants transport phenomena). In the latter, one speaks about “multiparadigm” models. Multiparadigm models can be classified in two classes: “direct” or “indirect”.

Direct multiparadigm models are multiparadigm models which include “on-the-fly” mathematical couplings between descriptions of mechanisms realized with different paradigms: for example, coupling continuum equations describing transport phenomena of multiple reactants in a porous electrode with Kinetic Monte Carlo (KMC) simulations describing electrochemical reactions among these reactants. Several numerical techniques are well established to develop such a type of models applied in the simulation of physicochemical processes e.g. catalytic and electro-deposition processes^{97,98}. In the field of catalysis, KMC simulations have been used to calculate instantaneous kinetic reaction rates on a catalyst calculated iteratively from concentrations which are in turn calculated from Computational Fluid Dynamics (CFD)-like continuum transport models⁹⁹: the calculated reaction rates are in fact sink/source terms for the transport models.

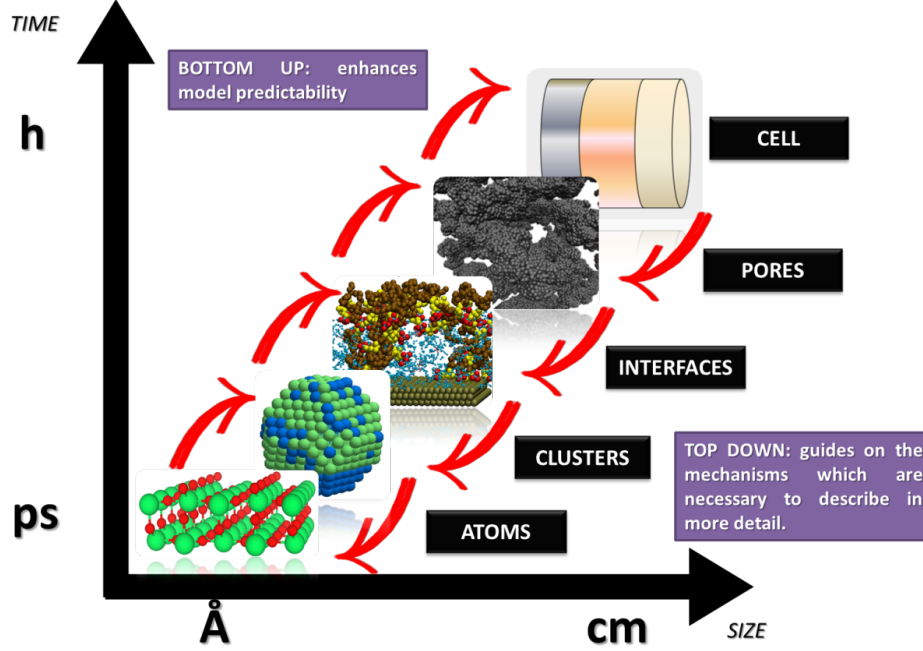


Figure 10. Bottom-up and top-down multiscale modeling.

Even if very precise, these methods can reveal themselves to be computationally expensive. For this reason, indirect multiparadigm models consisting in injecting data extracted from a single scale model into upper scale models via their parameters, can constitute an elegant alternative. For example, in the field of catalysis, one can use Nudged Elastic Band (NEB) calculations¹⁰⁰ to estimate the values of the activation energies E_{act} of single elementary reaction kinetic steps, and then inject them into Eyring's expressions to estimate the kinetic parameters k

$$k = \kappa \frac{k_B T}{h} \exp \left(-\frac{E_{act}}{RT} \right) \quad (1)$$

where κ refers to the frequency pre-factor, k_B and R the Boltzmann and ideal gas constants, T the absolute temperature and h the Planck constant. These expressions are used for the calculation of the individual reaction rates at the continuum level¹⁰¹,

$$v_i = k_i \prod_y a_y^v - k_{-i} \prod_{y'} a_{y'}^{v'} \quad (2)$$

where a refers to the activity of the reactants and products and v the stoichiometry coefficients. Equations (2) are in turn used for the calculation of the evolution of the surface or volume concentrations of the reaction intermediates, reactants and products, following

$$K_n \frac{da_y}{dt} = \sum_i v_i - \sum_j v_j \quad (3)$$

where K_n is the number of reaction sites per mol of reactants.

More precisely, NEB method aims in fact to find reaction pathways when both the initial and final states are known. The pathway corresponding to the minimal energy for any given chemical process may be calculated, but however, both the initial and final states must be known. NEB method consists in linearly interpolating a set of images between the known initial and final states, and then minimizing the energy of this string of images. Each “image” corresponds to a specific geometry of the atoms on their way from the initial to the final state, a snapshot along the reaction path. Thus, once the energy of this string of images has been minimized, the pathway corresponding to the minimal energy is found.

Another example of multiparadigm model results from the use of Coarse Grain Molecular Dynamics (CGMD) for the calculation of the materials structural properties (e.g. tortuosity and porosity) as function of the materials chemistry, which are used in turn for the estimation of the effective diffusion parameters used in continuum reactants transport models¹⁰²:

$$D_{eff} = \frac{\epsilon}{\tau} D_0 \quad (4)$$

where ϵ refers to the material porosity and τ to the material tortuosity.

One could then imagine building up in this way an EPGs model with macroscopic equations based on materials parameters extracted from atomistic and molecular level calculations (Figure 11). This approach gives a method for systematically investigating the effect of different materials designs on the LIB efficiency and durability.

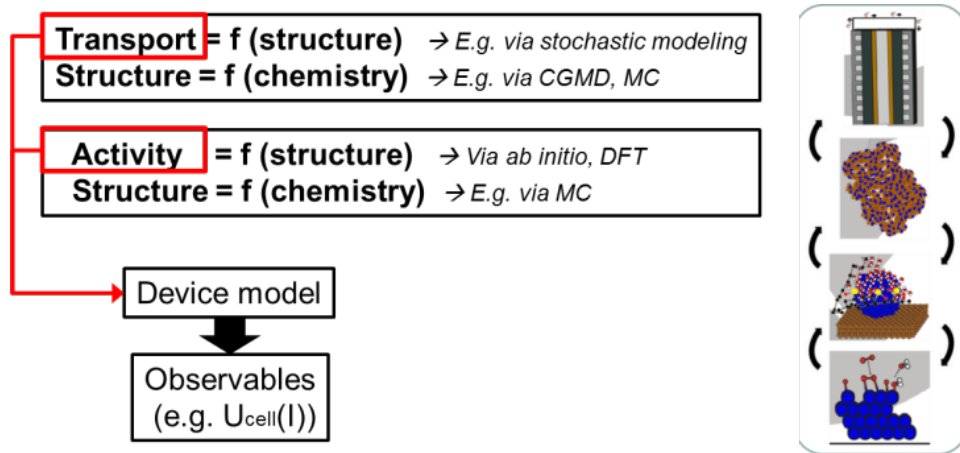


Figure 11. Schematics of an indirect multiparadigm approach for the simulation of EPGs.

Figure 12 summarizes the logical interdependencies of the three concepts revisited in this section.

2.3 Modular models: programming aspects and concepts

Application of multiscale modeling to LIBs is quite recent but this is also true for the modeling of other EPGs such as fuel cells^{103,104}. As it is computationally expensive to per-

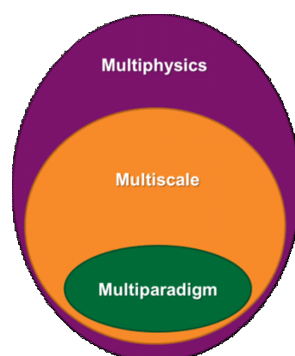


Figure 12. Schematics of the logical interdependencies between the multiphysics, multiscale and multiparadigm terminologies: a multiphysics model is not necessary “multiscale” either “multiparadigm”; a “multiscale” model is necessarily “multiphysics” but not necessarily “multiparadigm”.

form a predictive numerical simulation of a LIB operation response while capturing all the possible couplings among the different physicochemical processes in varied characteristic length and time scales in complex geometries using a single computational domain, the majority of the ongoing research efforts concentrate on developing indirect multiparadigm models.

Globally speaking, governing equations in indirect multiparadigm model include numerous nonlinear, coupled and multidimensional partial differential equations (PDEs) that are needed to be solved simultaneously in time along with some highly nonlinear algebraic expressions for transport and kinetic parameters. Rigorous EPGs models need from several seconds to a few minutes to simulate a discharge curve depending on the computer, solver, etc.

From a programming point of view, different languages, in-house or commercial software have been used to solve such model equations, e.g. Matlab, Simulink, C, Fluent, Comsol¹⁰⁵, or even combinations of those software and languages. Each class of software presents advantages and disadvantages depending on the desired application of the model developed. For example, Simulink is more adapted for system level simulation and Comsol is more dedicated to multiphysics models with detailed spatial resolution at the single cell level.

The numerical solver is also a critical aspect for robust simulations. As commercial software such as Simulink and Comsol propose a limited number of numerical solvers, or limited spatial meshing capabilities, numerous groups develop their own numerical solvers of Ordinary Differential Equations (ODEs) and PDEs, such as PETSc¹⁰⁶, LIMEX¹⁰⁷ or FiPy¹⁰⁸. In-house codes are usually more flexible and can be integrated within High Performance Computing (HPC) frameworks for sequential or parallel calculations¹⁰⁹.

The generation of parameters for such indirect multiparadigm models can be done with any kind of software available for *ab initio* calculations (e.g. VASP¹¹⁰, CRYSTAL¹¹¹, ADF¹¹², Gaussian¹¹³, BigDFT^{114,115,116}) or for molecular dynamics calculations (e.g. GROMACS¹¹⁷, LAMMPS¹¹⁸, AMBER¹¹⁹, CHARMM¹²⁰), the choice depending on the particularities of the material being studied and the kind of information one wants to extract with. A complete introduction to *ab initio* and MD methods including fundamental

concepts and detailed algorithms is beyond the goal of this paper and can be easily found in text books^{121,122,123,124}.

Multiscale simulation of device materials exhibits extreme complexity due to the variation of a huge number of possible compounds, device morphology and external parameters. Such simulation requires new hierarchical concepts to connect simulation protocols with the computing infrastructure, particularly with HPC architectures, on which the simulations can be executed (e.g. in the case of *ab initio* codes). Automatization of the generation of database libraries and their integration in indirect multiparadigm models is also an important aspect to be considered as highlighted by Bozic and Kondov¹²⁵. Some software platforms allowing to create data flows (or *pipelines*), to selectively execute some computational steps and to automatically inspect the results, are already available, such as KNIME and the platform UNICORE^{126,127}. Particularly for LIBs, an in-house flexible and scalable computational framework for integrated indirect multiparadigm modeling is reported by Elwasif *et al.*¹²⁸. The framework includes routines for the codes execution coordination, computational resources management, data management, and inter-codes communication. The framework is interfaced with sensitivity analysis and optimization software to enable automatic LIB design.

Further than the choice of the programming language and the software characteristics, there are also other key issues related to the mathematical formulation of the models, such as their modularity and their parameters identifiability.

2.3.1 Modularity

In building up multiscale models, a process engineering viewpoint is mandatory in order to provide models which are acausal (i.e. not requiring an action to be modeled on the basis of input data from a previous action), modular (i.e. consisting on an interconnected network of “modules”, each “module” describing an unique physicochemical mechanism) and reusable (i.e. where physics can be exchanged without changing the mathematical formulation and/or interconnection between the “modules”).

Numerous models in process engineering are based on a structured approach using sets of balance equations, constitutive equations and constraints¹²⁹. Mangold *et al.*¹³⁰ proposed a block-diagram approach which also applies to distributed parameter systems. This block-diagram is constituted of the components elements (representing the storage of conserved quantities) and the coupling elements (defining the fluxes between components) related by bidirectional signal flows (composed of potentials and fluxes). In this approach causality is thus assigned once for all which harms the reusability of the submodels, the submodels have to be re-defined for each new set of boundary conditions (i.e. for each new configuration of the interconnection with the environment). Maschke *et al.* proposed a port-based model using a novel extension of the bond graph language to multiscale and non-uniform models (also known as *infinite-dimensional bond graphs*)⁸⁴. Their work extends previous work on structured modeling for chemical engineering using bond graph for finite dimensional systems^{131,132,133,134,135,136,137}. This approach leads to a simple and easily re-usable graphical description of the system which is an interesting modeling alternative to sets of PDEs and boundary conditions. The infinite dimensional bond graph models represent the basic thermodynamic properties, conservation laws at each scale and the multiscale coupling in terms of a network of multiport elements acausally related by

edges (bonds) indicating the identity of pairs of power conjugated variables (intensive and variations of extensive variables). This network includes energy dissipative elements (“R” elements) and energy cumulative elements (“C” elements). Cumulative elements concern the balance equations such as

$$\frac{\partial C}{\partial t} = \nabla \cdot J + S \quad (5)$$

with a source/sink term interconnected with a smaller or higher scale through the boundary conditions, e.g. via the flux,

$$S = \gamma \times J_{\partial V} \quad (6)$$

where γ is the specific surface area between the scales (e.g. in $\text{m}^2 \cdot \text{m}^{-3}$).

Dissipative elements concern the constitutive equations such as,

$$J = -\Gamma(C) \times \nabla \tilde{\mu} \quad (7)$$

Because of the multidisciplinary property and the “universal” formulation inherent to the bond graph approach, it provides a framework which facilitates the collaboration between experts working on different physical domains.

The modularity of infinite dimensional bond graphs provides to the multiscale models a hierarchical, flexible and expandable mathematical architecture.

An example of infinite dimensional bond graph representation is provided in Figure 13 for the case of the modeling of a chemical reactor⁸⁴. Because of the reusability of the approach, the same representation will be valid, for example, for the modeling of ion transport across the EPG porous electrodes.

Specific software can be used to build up these types of models and to calculate the propagation of the causality, e.g. 20-Sim¹³⁸ or using similar concepts, AMESim¹³⁹, Dymola¹⁴⁰ and OpenModelica¹⁴¹.

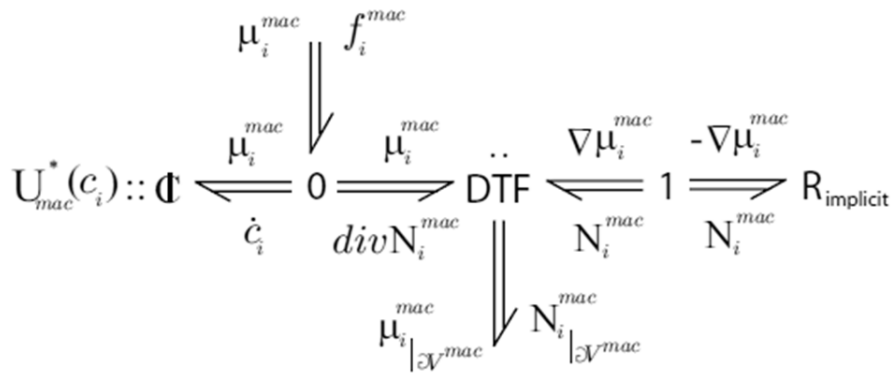


Figure 13. Example of an infinite dimensional bond graph model, representing the transport of a reactant across a porous media with a sink/source term connected to a microscale model (not shown here). Source: ¹³².

Only very few efforts have been reported on modeling batteries paying particular attention on their a-causality, modularity and reusability by using the bond graph approach^{142,143}.

For PEM Fuel Cells, significant progress within this sense has been achieved: the models developed by Franco are fully based on the use of infinite dimensional bond graphs^{144,145,146}. The models represent explicitly the different physical phenomena as nonlinear sub-models in interaction. Such developed models are multi-level ones in the sense that it is made of a set of interconnected sub-models describing the phenomena occurring at different levels in the PEMFC. However, this description remains macroscopic (suitable for engineering applications) in the sense that it is based on irreversible thermodynamic concepts as they are extensively used in chemical engineering: use of conservation laws coupled to closure equations. Such an approach allows to easily modify the sub-models and to test new assumptions keeping the mathematical structure of the model and the couplings.

The infinite dimensional bond graph structure of a generic multiscale model for the numerical simulation of electrochemical devices for energy conversion and storage is presented in¹⁴⁷ where several application examples are discussed, including fuel cells, electrolyzers and batteries.

In the oral presentation associated to this lecture, it will be shown how to build up bond graph-based models by using both 20-Sim and Simulink software. For instance, Simulink (from “Simulation” and “Link”) is a graphical extension of MATLAB by Mathworks for modeling and simulation of dynamical systems. The construction of a model under Simulink environment is done with click-and-drag mouse operations under a graphical user interface (GUI) environment. Systems are drawn on screen as block diagrams with inputs and outputs which can be 1 to 1 associated to the ports of a Bond Graph element, from a customizable set of block libraries (Figure 14). Many elements of block diagrams are available, such as transfer functions, summing junctions, etc., as well as virtual input and output devices such as function generators. Simulink provides an interactive graphical environment offering on one side a quick programming approach to develop models in contrast to text based-programming language such as e.g., C, and in the other side it has integrated fixed and variable time step solvers (in text based-programming language such as e.g., C one needs to code the solver).

However, it is in practice very helpful to combine Simulink capabilities, with the compactness of C and/or Python programming languages. For instance, some specific blocks can be programmed in C and/or Python and then embedded in Simulink.

Finally it should be noticed that Simulink is a software originally devoted to the development of system-level models, in particular for control-command purposes. Franco was pioneering its adaptation and its use for the numerical simulation of electrochemical processes¹⁴⁴. In the oral presentation associated to this lecture, several practical examples and exercises on the use of Simulink for the modeling of EPGs will be provided.

2.3.2 Identifiability and parameters estimation

While EPG models that are trained to experimental data provide great benefit in their ability to be integrated into vehicle models, because of their simple construction and fast computational speed, they have several shortcomings. In particular, these models are only as

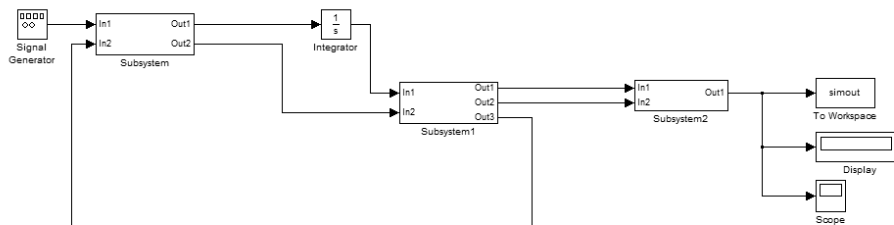


Figure 14. Example of model built up in Simulink.

good as the experimental data they are trained to, and thereby do not provide the ability to extrapolate beyond the range of this data. Moreover, changes in the cell design do not allow the use of the same models, and the task of building prototype cells, collecting data and training the model has to be repeated. Additionally, as these models are empirical in nature, they provide little, if any, insight into the operation principles of the cell.

By construction, bottom-up multiparadigm and multiscale models are not designed for fitting, especially if they contain parameter values estimated by atomistic or molecular level calculations. Predicting observable trends with these models with good order of magnitude in a large diversity of conditions and materials properties can be sufficient enough to consider them as “validated”. But estimation of all the parameters of such a type of models from atomistic or molecular calculations is impossible and experimental fitting of some empirical parameters, inherent to the non-ideality of the real system being simulated (cf. Section 1), is always necessary.

For this, identifiability is a crucial aspect to be treated when developing multiscale models. When these models are used for EPG optimization, the estimation of accurate physical parameters is important in particular when the underlying dynamical model is nonlinear. Identifiability concerns the question of whether the parameters in a model mathematical structure can be uniquely retrieved from input-output data. Literature on identifiability and techniques to check identifiability is extensive^{148,149,150,151}. Being able to first assess the identifiability of a model without going through the estimation work (e.g. by using iterative methods such as Gauss-Newton or Steepest-Descent, or the Bayesian method) allows gaining time in the model development. Identifiability analysis can result in structurally non-identifiable model parameters. Furthermore, practical non-identifiability can arise from limited amount and quality of experimental data. In the challenge of growing model complexity on one side, and experimental limitations on the other side, both types of non-identifiability arise frequently, often prohibiting reliable prediction of system dynamics. Once non-identifiability is detected, it can be resolved either by experimental design, measuring additional data under suitable conditions, or by model reduction, linearization¹⁵², tailoring the size of the model to the information content provided by the experimental data, or by more model refinement based on lower scale calculations.

EPG multiscale models usually present a large number of equations that result from finite difference reformulation of the mathematical expressions. Until recently^{153,154,155,156,157,158,159,160,161,162}, there were no significant efforts in developing ef-

ficient techniques for estimating parameters in multiscale EPG models because of computational constraints. In particular, Boovaragavan et al. reports a numerical approach for a real-time parameter estimation using a reformulated LIB model ¹⁶³. It is to be noted in this work that the estimation of parameters using LIB models are performed only for up to 2C rate of discharge. Further reformulation of the authors' multiscale LIB model is required to enable estimation of parameters at high rate of discharge.

3 Multiscale Models of EPGs: Examples and Practice

In the following the practical use of the different concepts introduced in the previous sections is illustrated through some examples of indirect multiscale models of EPGs. This section does not intend to be exhaustive and only exposes some few relevant application cases: the oral presentation associated to this lecture will introduce other examples. For further application examples, the reader is invited to visit the home page of Prof. Franco and to read the publications within: www.modeling-electrochemistry.com

3.1 Modeling PEMFC electrochemical reactions

Different simulation approaches of the PEMFC performance have been developed during the last 20 years from the pioneering papers of Springer et al. and Bernardi and Verbrugge ^{164,165,166,167,168}. These models quite well describe water management and thermal phenomena occurring in PEMFC for different operating conditions ^{169,170,171,172}. A number of CL models have been then developed, including the interface models ^{164,165, 166}, the thin film models ¹⁷³, the agglomerate models ^{174,175,176,177,178,179}, and the thin film agglomerate models ^{176,177}. Optimum performance of PEMFC for a number of parameters (type of agglomerate, CL thickness, CL porosity, distribution of Nafion[®] content, Pt loading, etc.) has been already impressively investigated ^{180,181,182,183,184,185}.

The mean feature of these models is that the kinetic rates associated to the electrochemical reactions are described via *Butler-Volmer equations* with empirical parameters, not connected with atomistic processes and thus describing reactions through effective global steps ¹⁸⁶. The numerical estimation of the values of parameters such as the zero exchange current (i_0) or the symmetry factors (α) is often a difficult task ¹⁸⁷. These macroscopic parameters show strong dependence on the PEMFC operation parameters such as the temperature or the reactants relative humidity, the CL mesostructural properties and even the MEA or bipolar plates design ¹⁸⁸. Modeling-based optimization of CL reactants and water transport for enhanced performance and stability requires a good knowledge of these electrochemical parameters related to the chemical and nanostructural properties of the catalyst.

The possible impact of water and ionomer which are expected in the vicinity of the catalyst in realistic PEMFC environments on the ORR kinetics appears to be unexplored ^{189,190}. In fact, in these models the electrochemical double layer capacity is usually assumed to be constant (i.e. the electrochemical double layer structure been uncoupled from the elementary reactions) ¹⁹¹ (Figure 15). It is well known that this is an important assumption that can lead to contradictory interpretations of experimental data as the electrochemical interface is expected to evolve under transient conditions (such as in the case of aging nanoparticles, oxidation and corrosion mechanisms), and the structure of the electrochemical double layer influences in turn the electron transfer rate and thus the effective electroactivity

properties of the catalyst surface^{192,193,194}. More generally, experimental evidence has been reported by Adzic et al. that simple Butler-Volmer equations are inappropriate for describing HOR and ORR reactions on Pt microelectrodes¹⁹⁵. There is also some experimental evidence for nanosized electrodes, where pronounced nanoscale non-linear effects of charge transfer in the surrounding electrolytic environment are important and cannot be explained using conventional electrochemical theories¹⁹⁶.

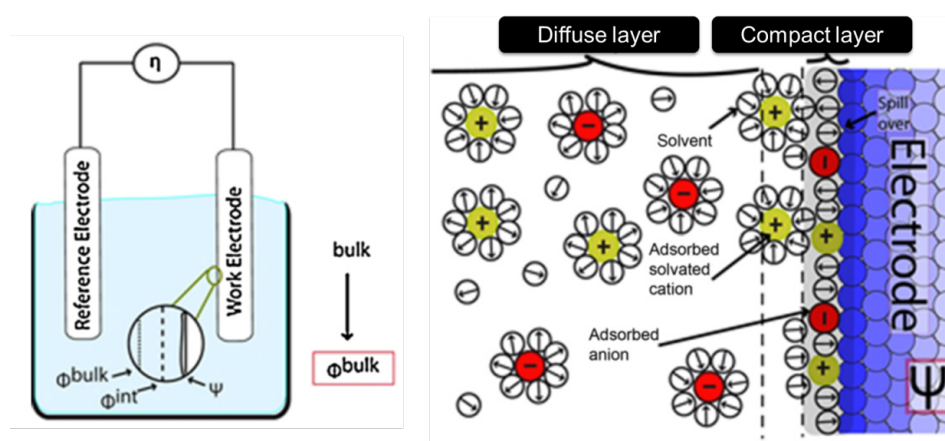


Figure 15. Schematics of an equilibrium electrochemical double layer in liquid/solid interfaces.

More refined Butler-Volmer models, splitting global reaction steps into a set of elementary steps, have been developed and mainly used to explore external contaminants impact on the PEMFC performance or MEA materials aging mechanisms^{197,198,199}. Kinetic parameters are usually estimated from experimental fitting, without checking the thermodynamic consistency of the proposed pathways at the atomistic level. Very few efforts have been reported to connect such elementary kinetic models with atomistic data obtained, for example, from *ab initio* calculations^{200,201}. The kind of electrochemical model used in a multiphysics model of a PEMFC can impact on estimated values of the other model parameters (if fitted from experimental data) such as the ones related to transport phenomena. It is thus important to develop appropriate elementary kinetic models for robust optimization of the other model parameters.

On the other side, the growing use of nanosciences are encouraging to understand and thus to control the fundamental structure and behavior of the PEMFC materials at the atomic and molecular level. First-principles or *ab initio* calculations (such as the DFT method) can be used to predict important quantities such as adsorbate atomic structures and bonding energies, and provide key information on the reaction mechanisms and pathways. This includes the determination of the controlling elementary reaction pathways and intrinsic kinetics involved in the ORR over Pt and Pt based alloys and their potential dependent behavior. This is also related to the understanding of the influence of the extrinsic reaction environment including the surface coverage, alloy composition, solution phase and electrochemical potential. New functionalities or lower Pt loadings (e.g. via the

development of multi-metallic catalysts with lower Pt loading²⁰²) are made available by manipulation of matter at this scale or through specificities of the nanodimensions, where the physical and chemical properties of materials differ from those of the bulk matter. In this context DFT has been largely used to explore different PEMFC reactions in the absence of interfacial electric field (e.g. ORR steps in^{203,204,205,206,207,208,209,210,211,212,213,214,215}). Generally these studies were performed using a few atoms/small clusters or extended surfaces to simulate the catalyst.

However, a complete description of ORR kinetics from first principles calculations for pure Pt and PtM surfaces is still missing and the influence of the nanoparticles morphology and the surrounding electrochemical double layer structure (strongly influenced in turn by the micro and mesoscopic transport phenomena of reactants, charges and water inside the electrodes) onto the effective ORR kinetics is not solved yet (Figure 16).

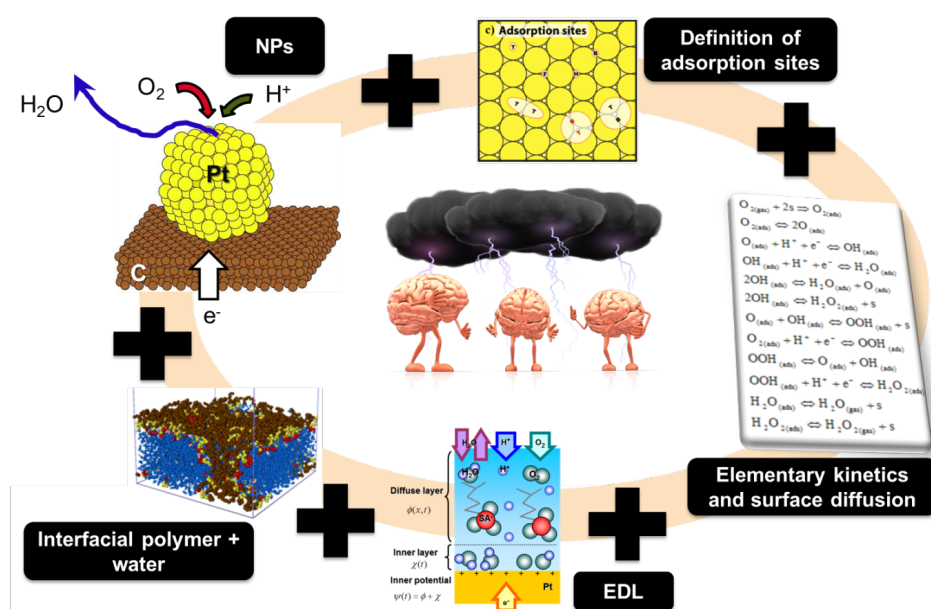


Figure 16. Aspects besides the challenge of modeling and simulation of electrochemical reactions and interfaces in PEMFC environments.

A receipt towards the building up a model describing electrochemical reactions in realistic PEMFC environments is presented in Figure 17.

First, by the use of DFT calculations thermodynamically favorable reactions steps are detected and the associated activation energies calculated: an example of a result for the ORR is reported in Figure 18²¹⁶ where the related chemical and electrochemical processes are modeled by series-parallel elementary kinetic steps (e.g. O₂ dissociation followed by the H₂O formation), and where Pt nanoparticles are modeled by a Pt(111) surface. For instance, this model neglects side effects (i.e. edge of sites on a nanoparticle, and kink sites). Although not perfect, this approach is still sufficient enough for predicting



Figure 17. Steps towards the development of an electrochemical model in PEMFC environments.

relevant CL potential evolution trends. Full elementary kinetic modeling of reactions on 3D nanoparticles still remains a great challenge, at least for atomistic theoretical studies ^{217,218}.

The calculated energy barriers are then used to estimate the kinetic rate constants for each single reaction step involved in the ORR. A Mean Field (MF) approach can be used to build the elementary kinetic model and describe the rate of the individual reactions in the CL.

The activation energy of each elementary ORR step can be coverage dependent. Only few published DFT results exploring coverage effect on the adsorption and activation energy are available ²¹⁹. So based on the DFT calculation at low coverage and the available literature on higher coverage for each ORR elementary step, we can estimate the maximum change that we can expect for E_{act} at high coverage. Then using a simple linear relationship between the energy barrier and the total coverage given by ²¹⁰

$$E_{act} = E_{act0} + I \sum_{i=0}^N \theta_i \quad (8)$$

it is possible to estimate E_{act} at each time step, where E_{act0} is the activation energy calculated by DFT and θ_i is the coverage of species i . In the result part we show a comparison of i-V curves with and without the dependence of E_{act} on the coverage. The reader is invited to refer to Ref. 216 for the details. It is however important to notice that the coverage effect can be quantitatively refined from Monte Carlo simulations as it will be illustrated in the oral presentation associated to this lecture.

Transition State (TS) theory formulation can be then systematically applied for the calculation of the kinetic rate parameters. The general formulation is given by:

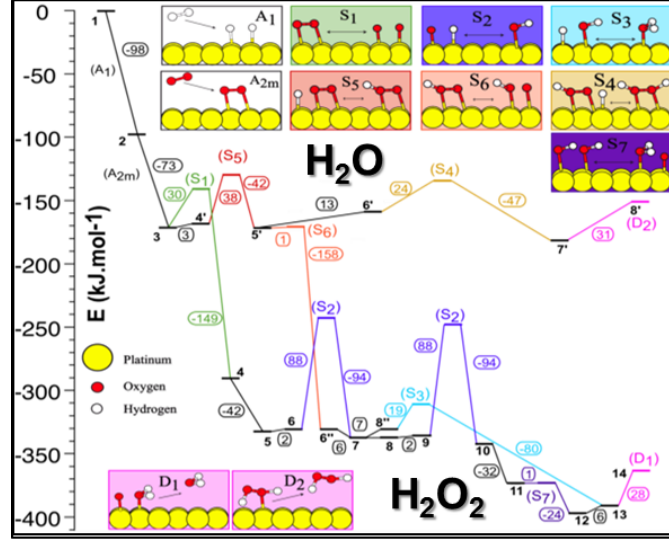
$$k_i = \frac{1}{N_A S} \frac{k_B T}{h} \frac{Q^{TS}}{\prod Q^{TS}} \exp\left(-\frac{E_{i,act}}{k_B T}\right) \quad (9)$$

where k_i and $E_{i,act}$ are the kinetic rate constant and activation energy of elementary step i respectively. S is the total surface area of catalyst, Q^{TS} and $\prod Q^{TS}$ are the partition functions identified to 1 here as a first approximation ^{220,221}.

It is important to note that the activation energies $E_{i,act}$ can be actually given by the addition of the DFT-calculated activation energy and an empirical parameter related to all the non-idealities not considered in the DFT calculations (e.g. presence of kinks, presence of solvent, etc.), thus

$$E_{i,act} = E_{i,act}^{DFT} + \delta_i \quad (10)$$

The empirical parameter can then be used for experimental fitting of the predicted observables (e.g. i-V curves) to guide further DFT calculations devoted to refine more the theoretical description of the reaction pathway. For instance, Reuter et al. proposed a



	E_{act}
$O_{2(s)} + s \leftrightarrow 2O_{(s)}$	30
$H_{(s)} + O_{2(s)} \leftrightarrow O_2H_{(s)} + s$	38
$O_{(s)} + H_{(s)} \leftrightarrow OH_{(s)} + s$	88
$OH_{(s)} + H_{(s)} \leftrightarrow H_2O_{(s)} + s$	19
$O_2H_{(s)} + H_{(s)} \leftrightarrow H_2O_{2(s)} + s$	24
$O_2H_{(s)} \leftrightarrow OH_{(s)} + O_{(s)}$	2
$OH_{(s)} + OH_{(s)} \leftrightarrow H_2O_{(s)} + O_{(s)}$	1

Figure 18. Elementary kinetic steps detected as the most favorable ones for the ORR on Pt(111) with the associated activation energies. Schematics built from Ref. 216.

systematic methodology for the development of error-controlled *ab initio* based kinetic models (Figure 19)^{222,223,224}. The methodology consists on refining iteratively the kinetic rates by starting from coarse kinetic models mixing DFT-based and empirical kinetic rate parameters. The parameter sensitivity analysis guides the further efforts still necessary to be done from *ab initio* calculations.²²⁵

The kinetic reaction rate v_i of a surface reaction involving two adsorbed species is calculated in the following way:

$$A_{ads} + B_{ads} \leftrightarrow AB_{ads} + s \quad (11)$$

$$v_i = k_i \theta_A \theta_B - k_{-i} \theta_{AB} \theta_s \quad (12)$$

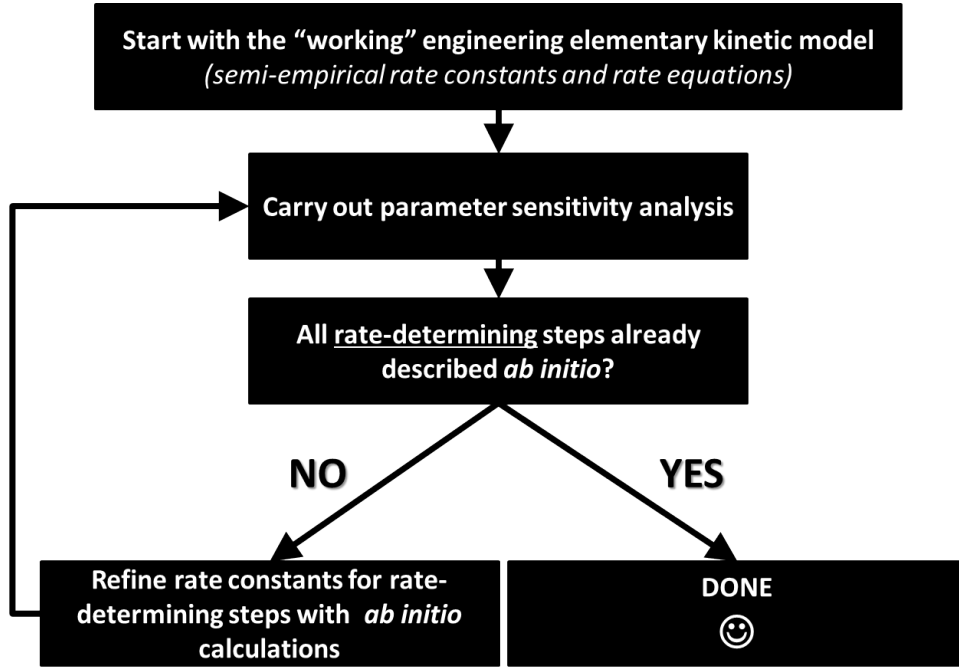
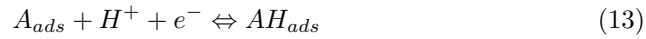


Figure 19. Methodology for the development of error-controlled *ab initio* based kinetic models (schematics built from the concepts described in Ref. 224 and inspired from Ref. 225).

where the kinetic rate constants are given by equation (9).

In the case of the electrochemical steps the kinetic rate constant and the rate of the elementary steps is written as follows:



$$v_i = k_i \theta_A C_{H^+} - k_{-i} \theta_{AH} \quad (14)$$

where C_{H^+} is the proton concentration at the catalyst surface and where the kinetic rate constants are given by

$$E_{i,act} = E_{i,act}^{DFT} + f[|\psi_M - \phi_{x=L}|] \quad (15)$$

where f is a function of the electrostatic potential difference across the adlayer (or surface potential) and corrects the DFT-calculated activation energy by the interfacial electric field effects. For instance, imagine an electron moving from the metal with an surface electronic charge density to the reaction plane situated at $x = L$ (Figure 20): the electrostatic potential difference between the metal and the reaction plane will contribute on increasing the activation energy of the reduction reaction (as for the electron is more difficult “to leave” the surface). In this case, the function f is a positive function. Other situations (e.g. oxidation, or reductions with negative charge densities) can be analyzed in an analogous way. It is important to note that this potential difference across the adlayer is a function of

- the metal charge density and the charge density associated to the specific electrolyte ionic adsorption (ad-ions surface concentration);
- the polarization of the adlayer, which is in particular function of the coverage of solvent molecules (e.g. water). The coverage by solvent molecules can be calculated through the mass action law as discussed in ²²⁶ and it is a function of the metal surface charge density.

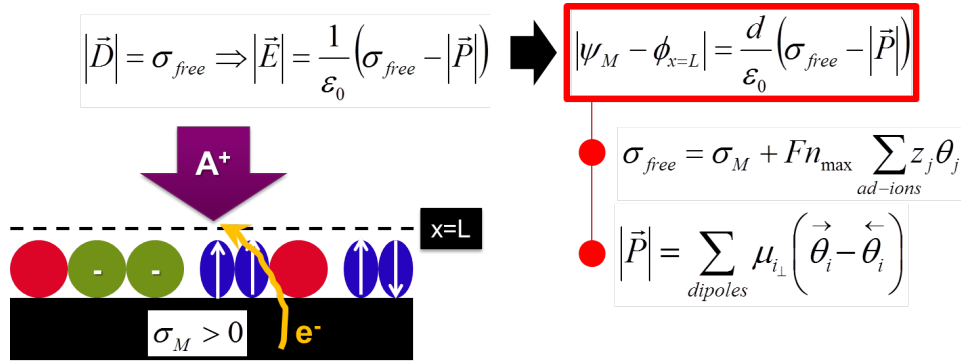


Figure 20. Schematics of the adlayer and of the calculation of the electrostatic potential difference from the surface displacement vector.

At first order f linearly depends on the surface potential, thus

$$f[|\psi_M - \phi_{x=L}|] = \alpha F |\psi_M - \phi_{x=L}| \quad (16)$$

where α is a constant parameter comprised between 0 and 1.

The electrostatic potential at the reaction plane can be calculated from the Gauss' law applied to the diffuse layer region (Figure 21), i.e.

$$\nabla \cdot \vec{D} = \rho_{\text{free}} \Rightarrow -\nabla^2 \phi = \frac{\rho_{\text{free}}}{\varepsilon} \quad (17)$$

where ε is the average electric permittivity of the electrolyte, where it was assumed that $\vec{D} = \varepsilon \vec{E}$ and that $\vec{E} = -\nabla \phi$ (i.e. neglecting magnetic fields) and where the charge volume density is given by

$$\rho_{\text{free}} = F \sum_i z_i C_i. \quad (18)$$

In the case of a 1D model, the boundary conditions of equation (17) are given by the value of the electrostatic potential at bulk and by the value of the electric field at the metal surface, i.e.

$$|\vec{E}| = \frac{\sigma_M}{\varepsilon} \quad (19)$$

which is a function of the metal charge density through the Gauss' theorem ²²⁶.

For the ORR in PEMFC environments, the concentrations in equation (18) are the proton concentration and the (Nafion[®]) sulfonic acid group concentration. The former, and in particular its value at the metal surface necessary for equation (14) can be calculated from the conservation equation:

$$\nabla \cdot J_i = -\frac{\partial C_i}{\partial t} \quad (20)$$

where J_i is the proton flux through the diffuse layer. The physics governing the proton transport is written as follows:

$$J_i = J_i (\nabla \tilde{\mu}_i) \quad (21)$$

where $\nabla \tilde{\mu}_i$ is the gradient of the proton electrochemical potential in the electrolyte. Under the assumption of diluted solutions, we can write

$$J_i = -D_i \nabla C_i - D_i \frac{F}{RT} C_i \nabla \phi \quad (22)$$

which jointly with equations (17) and (20) arises onto the so called Poisson-Nernst-Planck (PNP) system of equations.

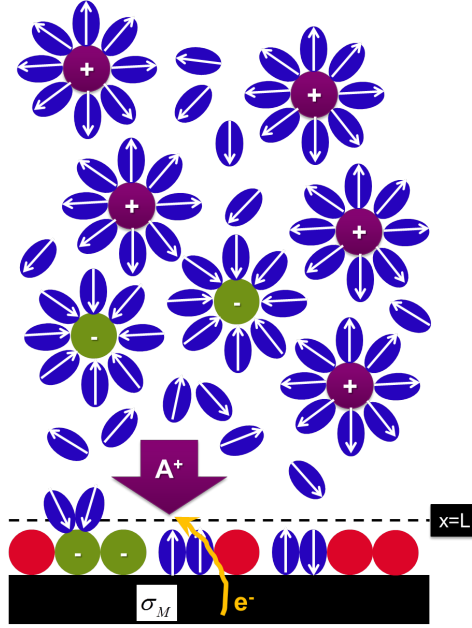


Figure 21. Schematics of the electrochemical double layer model.

Regarding the sulfonate group concentration the treatment is more complex. The simplest approach would be considering them as spatially fixed charges, as previously done by Franco et al., e.g. in Refs. 15, 144, 201. However, as recently demonstrated by Franco et al. on the basis of CGMD calculations^{227,228}, the hydrophilicity degree of the substrate can

strongly impact the interfacial morphology of Nafion[®] thin films and thus the sulfonic acid group concentration distribution over space (Figure 22). This is expected to impact the electrochemical double layer structure which will impact in turn the effectiveness of the ORR.

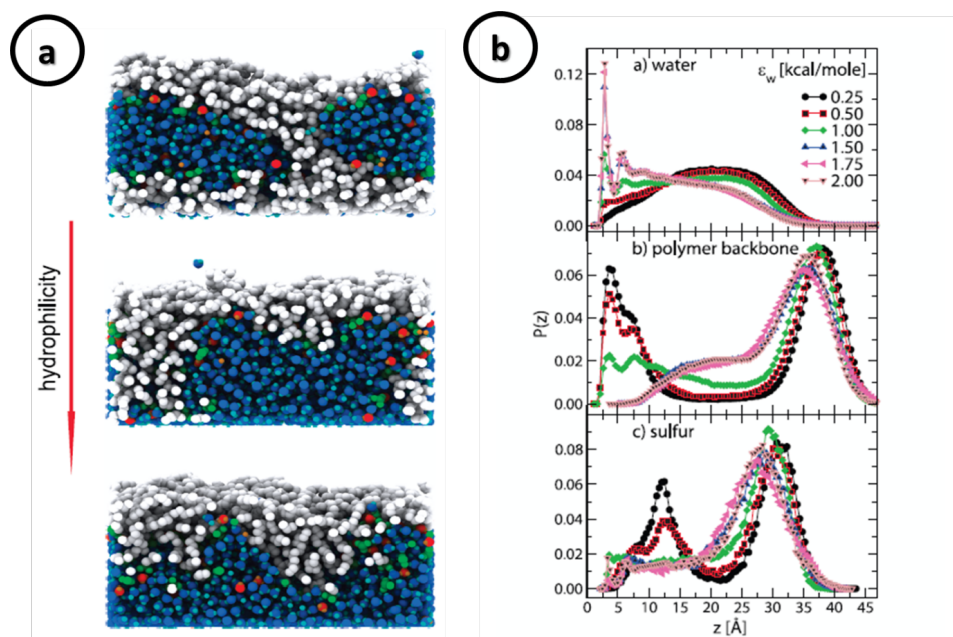


Figure 22. a) Snapshots of CGMD-calculated structures of hydrated Nafion ultra-thin films, at $T = 350$ K and number of water molecules per sulfonic acid group = 22, for an interaction with the support of increasing hydrophilic character ($\epsilon_w = 0.25, 1.0, 2.0$ kcal/mole, from top to bottom). We observe the formation of extended water pools (blue) which are separated from the confining polymer matrix (grey) by the charged sulfonic groups interface (green); hydronium complexes are also shown (red). For $\epsilon_w = 2.0$ kcal/mole the ionomer is completely desorbed from the substrate. Note the evaporated water molecules, on the top of the films. b) Mass probability distributions as a function of the distance from the support, z , at the indicated values for ϵ_w . We have considered a) water oxygens, b) polymer backbone units, and c) sulfur atoms. Source: Ref. 227.

For instance the following features are observed from the CGMD calculations:

- the ionomer density at the vicinity of the substrate decreases as the substrate hydrophilicity increases, and the opposite trend occurs at the top of the film;
- a compact water layer is formed in the hydrophilic cases;
- for the hydrophobic case, the side chains are pointing out from the surface. No presence of water is observed on the top as a hydrophobic film surface is formed which could prevent gas and water absorption;
- still in the most hydrophobic cases, the formation of hydrophilic water channels (inverted micelles) are more evident, and the formation of polymer layers are detected which would prevent the water and proton to diffuse through the film thickness;

- the agglomeration of adsorbing anions (SO_3^-) is observed when the hydrophilicity increases;
- for the most hydrophobic cases the polymer is adsorbed mainly via backbone, and for the most hydrophilic cases the presence of backbone is less evident at low hydration;
- the backbone can be adsorbed even in the hydrophilic surfaces.

The ionomer film structure will be impacted by the catalyst/carbon oxidation state (which determines its hydrophilicity). As the distribution of charge at the vicinity of the substrate is strongly affected by the ionomer structure, the surface hydrophilicity is expected to impact the proton concentration at the reaction plane, and non-uniform reaction rates are expected inside the CL.

It is important to note that the hydrophilicity of the Pt is expected to evolve during the PEMFC operation as its oxidation state changes (it becomes more oxidized when the ORR occurs at its surface). Thus, the structure of Nafion at the interface is also expected to evolve upon the PEMFC operation. All these structural features are expected to strongly impact the ORR kinetics through the polymer poisoning of the catalyst and the effective ionic transport and water uptake properties of the thin film.

Franco et al. ongoing efforts are aiming to develop electrochemical models accounting for these important features²²⁹: some examples will be provided in the oral presentation associated to this lecture, and some exercises in relation to this will be proposed in the hands-on part of the lecture.

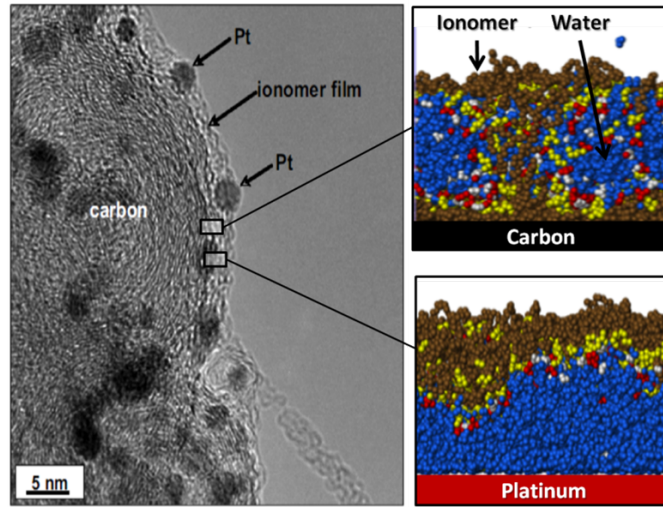


Figure 23. Schematics of the interfacial Nafion[®] thin film structure at the vicinity of Pt and C before PEMFC operation. Source: Ref. 3.

For desorption steps the following relations are used:

$$AH_{ads} \Rightarrow AH_g + s, \quad (23)$$

$$v_i = k_i \theta_{AH} . \quad (24)$$

Adsorption steps can be simulated using collision theory:



$$k_i = \frac{sc}{n_{\max}} \frac{P}{\sqrt{2\pi m k_B T}} \quad (26)$$

$$v_i = k_i \theta_s \quad (27)$$

where P and m are the partial pressure at $x = L$ and the atomic mass of reactant A respectively. sc is the sticking coefficient estimated from published values²³⁰. v_i is the rate of the elementary step, n_{\max} is the surface density of sites and θ_s the coverage of free sites.

Using the reaction rate of the set of elementary steps of each reaction mechanism in Figure 18 we can write a balance equation for calculating the coverage of each single ORR intermediate species by numerical integration, thus

$$\frac{n_{\max}}{N_A} \frac{d\theta_k}{dt} = \sum_l v_l - \sum_{l'} v_{l'} \quad (28)$$

where the balance equations are function of the reaction rates of creation and consummation of specie k. The calculation of the coverage is subject to the conservation of the total number of adsorption sites, which in the case of the cathode Pt catalyst surface can be written as follows

$$\theta_s + \sum_k \theta_k + \theta_{H_2O \rightarrow} + \theta_{H_2O \leftarrow} + \theta_{ionomer} + \theta_{j+} = 1 \quad (29)$$

where the coverage by water molecules pointing to/pointing out the surface and contributing onto the calculation of the surface potential appear. Equation (29) also depends on the coverage by the ionomer, where the contributions related to the side chains and backbone adsorption can be calculated with appropriate kinetic models to be exposed in the oral presentation associated to this lecture. θ_{j+} is the coverage by specifically adsorbed ions (relevant for liquid electrolytes or mixtures of liquid electrolytes with Nafion®) which will impact the surface potential magnitude (Figure 20).

A similar approach is used for describing the kinetics of C corrosion process²³¹. In order to take into account the effect of catalyst degradation (structure evolution) on the activity and stability properties, we extend classical elementary kinetic modeling by implementing time-dependent kinetic parameters (associated with each step) given by degradation-dependent activation energies calculated on catalyst surfaces with atomistic structures representative of different snapshots of the degradation process²³².

The metal surface charge density σ_M (Figure 20) is calculated by numerical integration of the conservation equation

$$J - J_{\text{Far}} = J - F \sum_l \tilde{v}_l = -\frac{\partial \sigma_M}{\partial t} \quad (30)$$

where $J(r, t)$ is the current density and $J_{\text{Far}}(r, t)$ is the faradic current density. This term is calculated inside the kinetic model by using the protonic reaction steps in the kinetic model. Finally, the cathode electronic potential is calculated from the scheme reported in Figure 20, $\phi_{x=L}$ given by the solution of equation (17).

Examples of calculated observables with this model are presented in Figures 24 and 25.

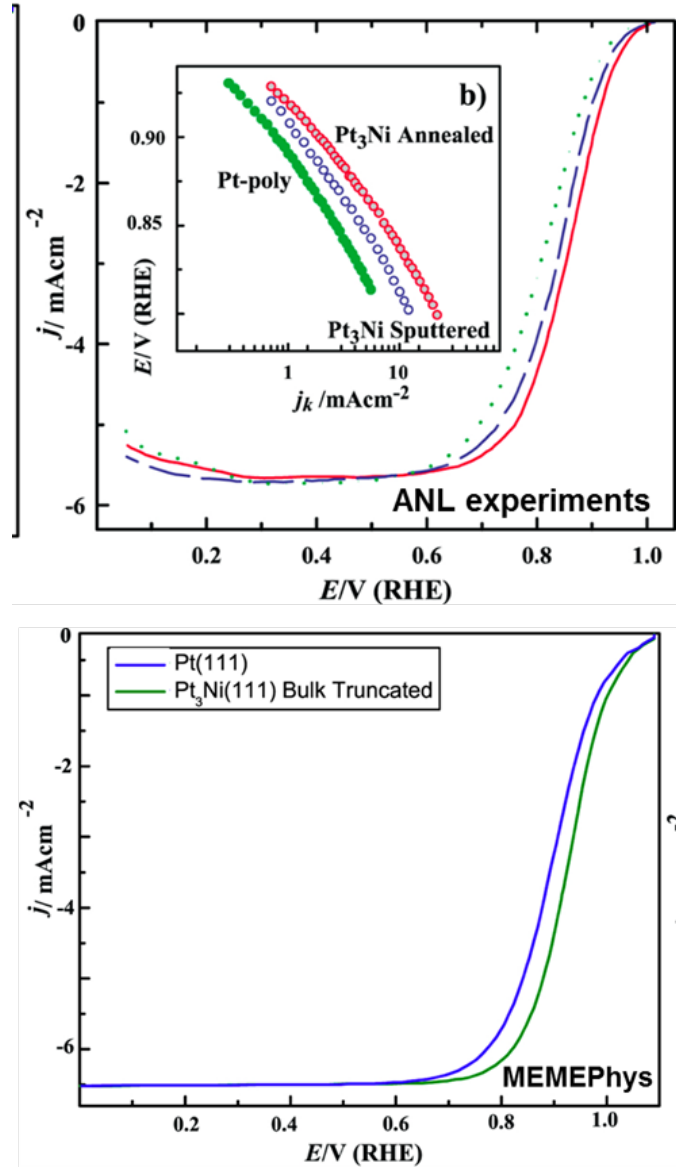


Figure 24. Calculated ORR activity for Pt(111) and Pt₃Ni(111) bulk-truncated catalysts and comparison with experimental data.

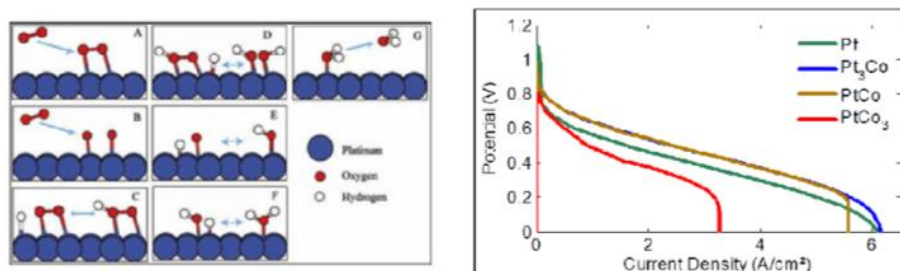


Figure 25. Representation of one of the DFT-calculated elementary kinetic models of the ORR mechanism on Pt(111) and examples of calculated i-V curves for Pt and Pt-Co catalysts by using DFT-based kinetic parameters.

In conclusion, the approach presented in this lecture allows relating the electrochemical kinetics with the chemistry and structure of the active material, through an elementary kinetic approach describing explicitly the electrochemical double layer structure, and thus having predictive capabilities that the classical empirical Butler-Volmer approach does not have (Figure 26).

	EMPIRICAL APPROACH	PHYSICAL APPROACH
CHEMISTRY	Empirical one-step kinetics	Multiple-steps kinetics
CHARGE TRANSFER	Butler-Volmer equation $i \cong \exp\left(\beta_a \frac{F}{RT} \eta_{act}\right) - \exp\left(-\beta_c \frac{F}{RT} \eta_{act}\right)$	Transition-State kinetic rates $v_i = v_i[E_{act}, \psi_M - \phi_{x=L} , \theta_i, C_i]$
CELL VOLTAGE	Substraction of overpotentials $E = E_0 - \eta_{an}[i] - \eta_{cat}[i] - iR_{el}$	Electrostatic potentials $\psi = f[i, C_i, \{\tilde{\theta}_i, \tilde{\theta}_i\}]$
DOUBLE LAYER	<u>Passive</u> constant EDL capacity $i_T = i[\eta] + C_{dc} \frac{d\eta}{dt}$	<u>Active</u> variable EDL capacity (not dissociated from REDOX) $J - J_{Far} = -\frac{\partial \sigma}{\partial t}$

Figure 26. Comparison between the classical Butler-Volmer modeling approach of electrochemistry and the approach presented in this lecture.

3.2 Modeling solid phases formation and evolution in Lithium Ion Batteries

Parallel to various experimental research programs, mathematical models that describe the behavior of LIBs and their interaction with other devices (e.g. vehicle electric engines) have received more and more attention for almost 30 years. These models range from those that are fitted to experimental data under various conditions (e.g. equivalent circuit ²³³ and neural network models ²³⁴) to the ones that describe the various physical mechanisms in the cell ²³⁵. Only very few reviews on LIB modeling have been reported in the last 10 years ^{236,237,238}. The majority of the published reviews provide a state of the art of LIBs modeling from a top-down engineering viewpoint, covering key issues such as the reusability of the governing equations for different battery systems, the aspects related to the coupling between mathematical descriptions of the electrochemical and thermal mechanisms in the cells, and the modeling at three different scales, namely electrode-level, cell-level and stack-level.

This lecture presents key techniques to develop bottom-up multiscale models for LIBs, i.e. spanning scales from atomistic mechanisms to the single cell level. Such types of models are also important for the prediction of the impact of the chemical and structural properties of the materials onto the overall LIB response. This paper does not intend to be exhaustive, but instead, it brings together general concepts and approaches (both methodological and numerical) as well as examples of applications. First, general aspects on physical modeling are revisited and some approaches for multiscale modeling are presented. Then, a critical review on ongoing efforts within the community is discussed. Finally, general conclusions, indications of the remaining challenges and suggested directions of further research are provided.

The so-called *phase-field modeling approach* is now receiving growing attention to understand phase separation, until now mainly on LiFeO_4 materials. Phase field models allow moving beyond traditional Fick's law in describing lithium diffusion in LIB electrodes. Phase field models are potentially more accurate and allow simpler tracking of phase boundaries than Fick's equation.

The phase field modeling approach, initially developed for describing phase separation and coarsening phenomena in a solid ²³⁹ and later for electrochemistry applications ^{240,241}, first consists in considering the total free energy of the intercalation (or conversion material), as follows:

$$F = \oint_V (f_{bulk} + f_{grad} + f_{app}) dV + \oint_V \oint_{V'} [f_{non\ local}] dV dV' \quad (31)$$

where f_{bulk} is the local chemical free energy density (function of the composition, e.g. Figure 27), f_{grad} is the gradient energy density (accounting for the heterogeneities penalties), f_{app} is the coupling potential energy between the applied fields and order parameters, and the second integral accounts for the long range interactions.

The chemical potential of each phase is given by

$$\mu_j = \frac{\partial F}{\partial c_j(\vec{r}, t)} \quad (32)$$

and the conservation equation governing the phases formation and displacement is given

by

$$\frac{\partial c_i}{\partial t} = -\nabla \cdot J = \nabla (M_{ij} \nabla \mu_j) \quad (33)$$

where M_{ij} refers to the mobility of each phase (could depend on the phases concentrations). Equation (33) is known as the *Cahn-Hilliard* equation²⁴². This is a fourth order equation, extremely sensitive to initial conditions and parameters values, which thus needs appropriate numerical schemes to solve them. This motivated a stronger and stronger interest for applied mathematics which brings onto the development of highly accurate but fast numerical methods such as the Chebyshev-spectral method, the generalized Newton's method, Fast Fourier Transform methods and multigrid methods, each method having pros and cons depending on the application problem^{243,244,245}. Furthermore, it should be noticed that phase field modeling is an elegant approach in which parameters can in principle be estimated from first principles calculations, e.g. as the interphase energies.

Han et al. reported one of the pioneering works on the application of a phase field model to describe phase separation in LiFePO_4 electrodes²⁴⁶. Using the phase field model the authors investigate to what extent non-Fickian behavior can affect results from experimental techniques for measuring diffusion coefficients, such as Galvanostatic Intermittent Titration Technique (GITT) and Potentiostatic Intermittent Titration Technique (PITT).

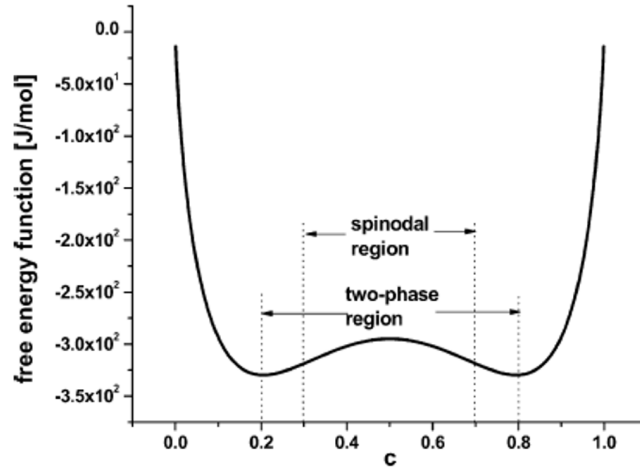


Figure 27. Example of free energy density functional used in the phase field modeling of LiFePO_4 electrodes. Source: 246.

Kao et al. compared phase field modeling results on LiFePO_4 to X-ray diffraction data and proposed the idea of overpotential-dependent phase transformation pathways²⁴⁷. From then, models developed account more and more for the strongly anisotropic transport in crystalline LiFePO_4 ^{248,249,250} as well as surface reaction kinetics^{251,252}. Within this sense, Bazant et al. introduced significant contributions on the application of anisotropic phase field modeling coupled with faradaic reactions to describe the intercalation kinetics

in LiFePO_4 (Figure 28)^{253,254,255}. For small currents, spinodal decomposition or nucleation leads to moving phase boundaries (Figure 29). Above a critical current density, the spinodal decomposition is found to disappear, and the particles start to fill homogeneously. This effect increases the active area for intercalation, and likely contributes to the high-rate capabilities and favorable cycle life of LiFePO_4 ^{255,256}. According to Bazant et al., this may explain the superior rate capability and long cycle life of nano- LiFePO_4 cathodes.

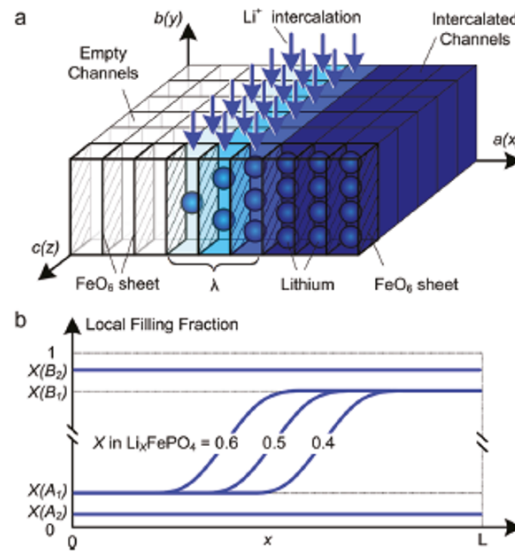


Figure 28. Schematic model of a Li_xFePO_4 nanoparticle at low overpotential (a) lithium ions are inserted into the particle from the active (010) facet with fast diffusion and no phase separation in the depth (y) direction, forming a phase boundary of thickness λ between full and empty channels (b) the resulting 1D concentration profile (local filling fraction) transverse to the FePO_4 planes for a particle of size L . Source: 254.

An alternative approach to the fourth-order Cahn-Hilliard equations is the so-called “Allen-Cahn approach” (Figure 30) which arises into non-conservative second-order equations of type “reaction-diffusion”. These equations are appropriate to describe conversion reactions in LIBs, such as CoO materials converting onto Co and Li_2O during the LIB discharge. As in the Cahn-Hilliard approach, the values of interface energies and diffusion coefficients in the Allen-Cahn approach can be estimated from DFT calculations^{257,258}. First ongoing efforts within this sense by Franco et al. will be detailed in the oral presentation associated to this lecture (Figure 31).

3.3 Modeling the relationship between the electrode structure and the performance and degradation of EPGs

Several models have been developed attempting to capture the influence of the electrodes structural properties (at the micro and mesoscales) onto the EPG performance and durability. For instance, in the case of LIBs for example, the anisotropic nature of ion diffusivity

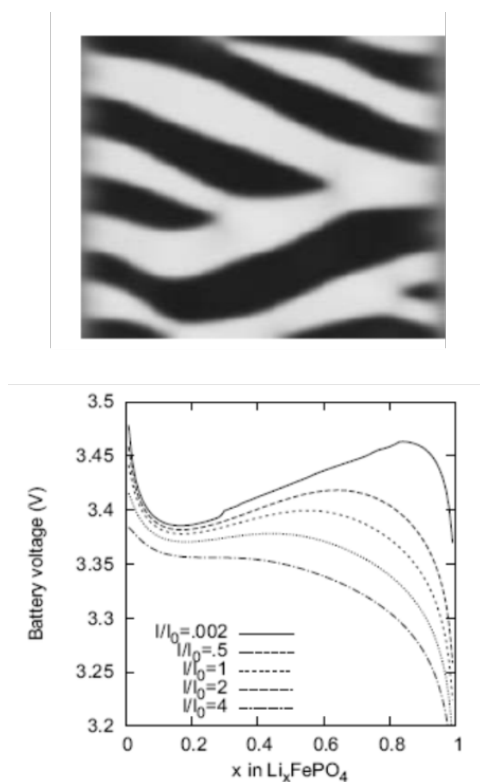


Figure 29. Top) calculated phase separation in a particle which has been allowed to relax from a homogeneous state at zero current with stress-free boundaries. White regions are lithium-rich. Down) Calculated LIB voltage at different applied currents. Source: Ref. 256.

in LiFePO_4 has motivated synthesis approaches that facilitate control of size and shape of LiFePO_4 agglomerates to maximize Li^+ transport. In general, it is important to develop modeling tools that can evaluate the relative impact of each single scale onto the overall efficiency of the LIB. A major problem in most modeling approaches is the reliable determination of model parameters. Especially for homogenized models, the microstructural parameters have a strong influence on the simulation results. A lack of knowledge of several parameters leads to a reduction of the models prediction capability.

To provide a detailed description of the mechanisms, a 3D representation is required for the morphology of composite materials used in EPGs.²⁵⁹ Nowadays, two ways of accounting for the detailed structure of the electrodes have been developed: one consisting in building up artificial structures capturing the main features of the real electrodes (e.g. length scales, particles shapes...), and another one based on computer-aided reconstruction of the real electrode structure. This is discussed in the following with particular focus on LIBs, but analogous discussions can be established for other EPGs.

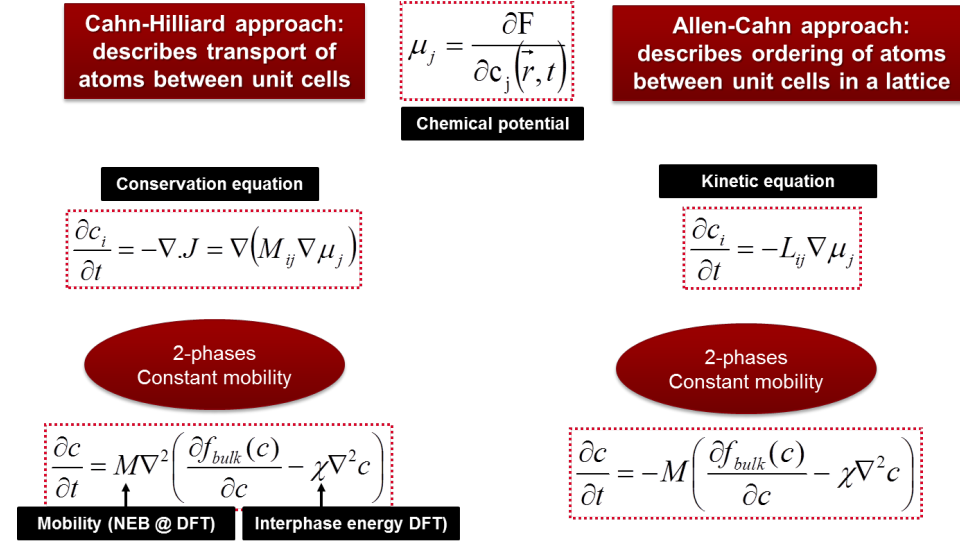


Figure 30. Cahn-Hilliard approach vs. Allen-Cahn approach.

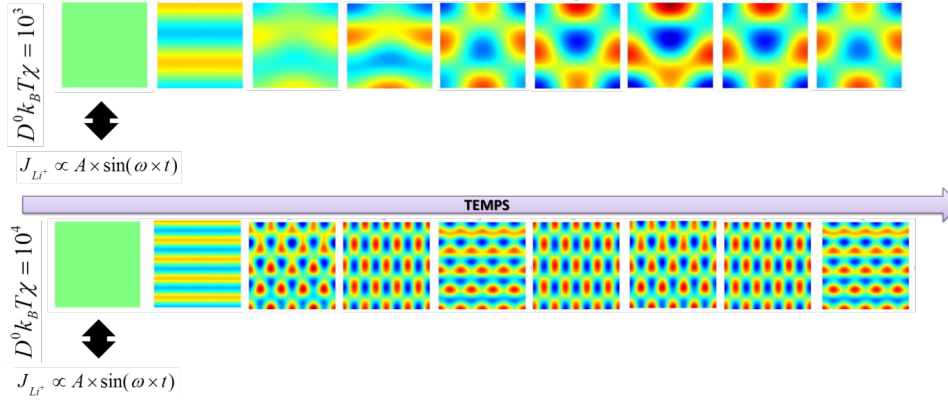


Figure 31. Simulated evolution of the microstructure of a conversion LIB electrode when cycling for two different values of interphase energy.

3.3.1 “Artificial” mesostructures

Dargaville and Farrell proposed a mathematical model to simulate the discharge of a LiFePO_4 positive electrode accounting for three size scales representing the multiscale nature of this material (Figure 32) ²⁶⁰. A shrinking core is used on the smallest scale to represent the phase transition of LiFePO_4 during discharge. The model is then validated against existing experimental data and is then used to investigate parameters that influence active material utilization. Specifically, the size and composition of agglomerates of LiFePO_4 crystals are studied by quantifying the relative effects of the ionic and electronic

conductivities onto the overall electrode capacity. The authors found that agglomerates of crystals can be tolerated under low discharge rates and that the electrolyte transport does limit performance at high discharge rates. For the former, the results from the particle scale show why minimizing the formation of agglomerates and shrinking the size of individual crystals is so successful at increasing the performance of a LiFePO_4 cell. For the latter, doubling the concentration of Li^+ in the electrolyte can increase capacity by up to 15%, though effort should be placed in seeking an electrolyte with better transport parameters, e.g., aqueous Li_2SO_4 . But aqueous electrolytes suffer from low electrochemical window.

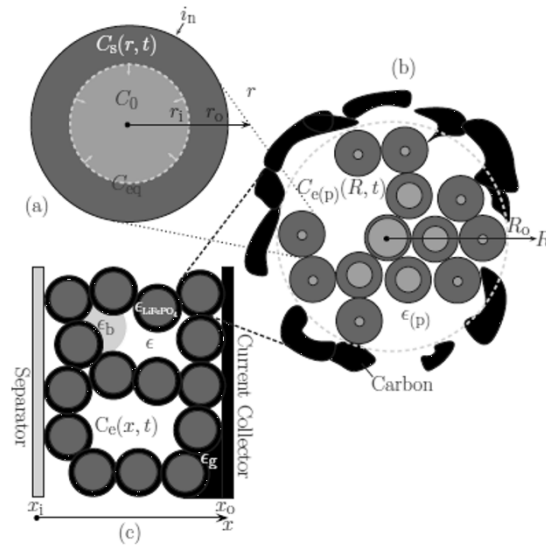


Figure 32. Schematic of the three size scales in the model of Dargaville and Farrell: a) crystal, b) particle, c) positive electrode.

A stochastic model consisting on energy-based structural optimization, has been developed by Smith et al.^{261,262} and allows the calculation of re-arranged equilibrium particle positions and orientations are calculated at given density (Figure 33). Resultant grain morphologies and assessment of the efficacy of each microstructure to enhance Li ion transport is quantified by the authors, who also report optimized grain morphologies for Li transport.

Du et al. recently reported a similar study in which a fixed number of monodisperse ellipsoidal particles are randomly packed based on a MD algorithm and then meshed using Cartesian voxels.²⁶³ The authors carried out 3-D finite element simulations on representative elementary volumes (REV) to estimate the parameters values in a cell model that vary with electrode microstructure, including the effective diffusivity, effective conductivity, and volumetric reaction rate. Results show lower effective diffusivity and conductivity in the electrode than predicted by the Bruggeman relation used in their cell model (cf. equation (4)), and a significant sensitivity in cell performance to this difference at high discharge rates.

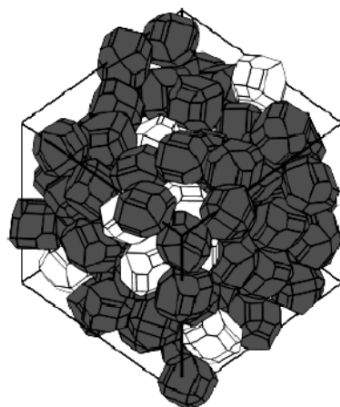


Figure 33. Example of stochastically simulated LiFePO_4 agglomerate morphology.

Goldin *et al.* present a three-dimensional model that can resolve electrode structure at the submicron scale.²⁶⁴ Although the three-dimensional model is capable of representing arbitrary electrode microstructure, the authors consider regular arrays of spherical particles. The model is applied to evaluate approximations in one-dimensional models and to assist in establishing empirical relationships that can be used in reduced-dimension models. General relationships for effective particle radius in one-dimensional models are derived from the three-dimensional simulations. The results also provide a basis for estimating the empirical Bruggeman exponents that affect Li-ion transport within electrolyte solutions. Three dimensional simulations of a dual-insertion Li-ion cell during galvanostatic discharge are compared with an equivalent one-dimensional model. The three-dimensional model fully resolves the electrode particles, which are assumed to be spherical but are packed into alternative lattice arrangements. The three dimensional model also fully resolves the porous electrolyte volume between the electrode particles. Under all conditions studied, intercalation diffusion appears to be the rate-limiting process that controls discharge characteristics.

More recently Song and Bazant proposed a simple but interesting model for the simulation of EIS as function of the morphology of the active particles (Figure 34).²⁶⁵ The model allows accounting for curved diffusion geometries as well as the diffusion length distribution. Using this model, the authors have investigated the ways these configurational aspects affect interpretation of diffusion impedance spectra. The model has been also applied to experimental impedance data of a Si nanowire electrode. Comparing the regression results of the different versions, we are able to show that including each of the cylindrical diffusion geometry and the heterogeneous radius distribution of the nanowires greatly improves the fit and leads to rather different, and presumably more accurate, values of the electrochemical parameters.

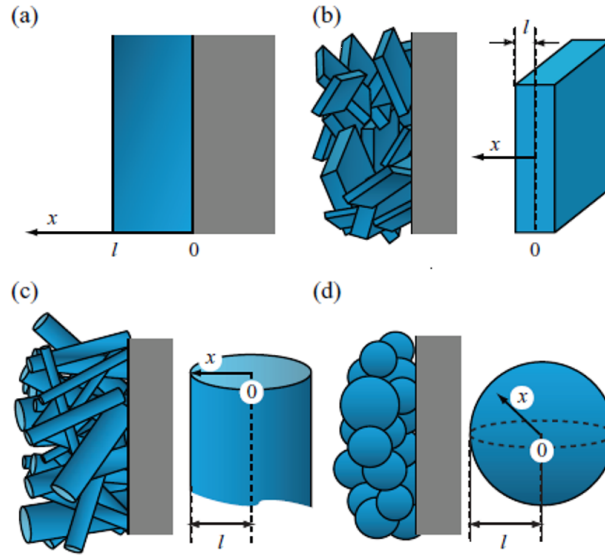


Figure 34. Song and Bazant's model electrode configurations, particle geometries, and corresponding coordinate systems, where the blue region and the gray region represent the active material and the current collector, respectively: (a) thin film electrode, (b) electrode with planar particles, (c) electrode with cylindrical particles, and (d) electrode with sphere particles.

3.3.2 “Real” mesostructures

Thiedmann et al. develop an impressive stochastic simulation model in 3D to reconstruct real and generate virtual electrode microstructures.²⁶⁶ For this purpose, a statistical technique to fit the model to 3D image data gained by X-ray tomography is developed. The detailed knowledge of the spatial distribution of the components of composite electrodes (e.g. LiFePO_4 electrodes with carbon additive) allows the authors to calculate macroscopic model parameters such as the active surface areas and the tortuosity, not directly accessible by other measurements, as well as physical parameters (e.g. diffusion constants, exchange parameters, conductivities ...). These spatially-resolved numerical representations are used by the authors to simulate the local and macroscopic electrochemical response of a LIB graphite electrode as a function of galvanostatic cycling (Figure 35).²⁶⁷ Through this analysis, the C-rate dependence on the dendrite formation and salt precipitation, a comparison against classical models based on artificial structures, and the well-known Newman models is determined. For high C-rates, the effect of tortuosity on salt precipitation, lithium accumulation and depletion is quantified.

Similarly, Ender et al. report 3D FIB tomography results of a complete LIB, including a positive LiFePO_4 -based electrode, a negative graphite-based electrode and a glass fiber separator (Figure 36).^{268,269,270} Macroscopic model parameters are also determined and their influence on the simulate overall LIB response is analyzed.

Bazant et al. report a very interesting microstructurally-resolved model of Li electrochemical intercalation and deintercalation processes (discharge and charge, respectively) in experimentally obtained 3D microstructures.²⁷¹ In their approach, an experimentally

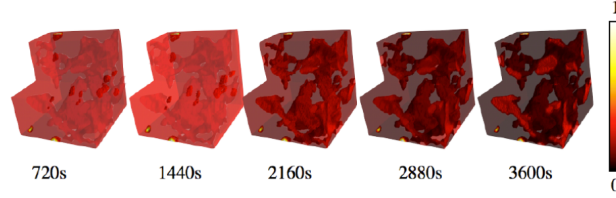


Figure 35. Calculated discharge sequence of a reconstructed graphite electrode.

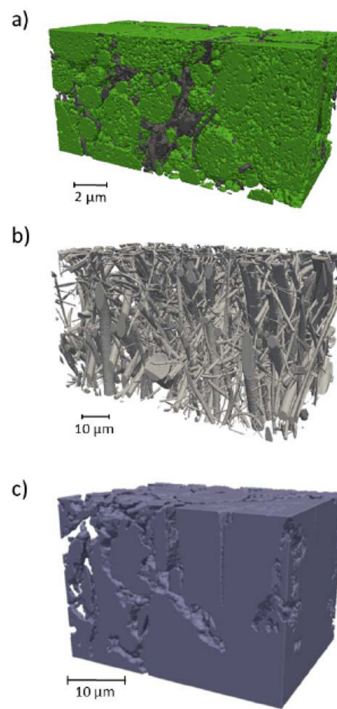


Figure 36. 3D reconstructions obtained by FIB tomography of (a) a LiFePO_4 composite positive electrode, (b) a glass fiber separator and (c) a graphite negative electrode.

obtained voxelated 3D microstructure array is converted to an input geometry described by a phase-field-like domain parameter. With such a parameter to distinguish the electrolyte, electrode and additive particles, the authors are able to solve the transport equations of Li^+ in the electrolyte coupled with the transport equations in the active material, without using complex structural meshing technique. The authors investigate several conditions of intercalation kinetics such as the effect of different voltage or current loadings on the electrode behavior as well as the role of the microstructure. In addition, different transport dynamics in the electrode, such as solid solution behavior (as observed in a portion of concentration range in LiCoO_2 , Figure 37) or phase-separation behavior (as observed in LiFePO_4), were

studied.

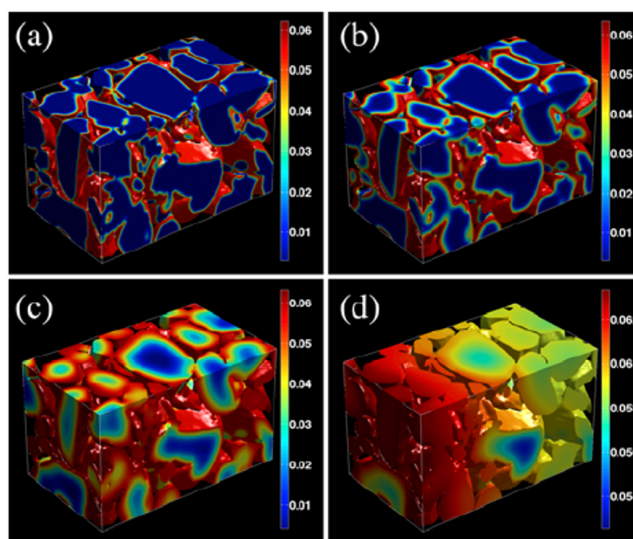


Figure 37. Li concentration (mol/cm^3) evolution during the discharge of a Li_xCoO_2 microstructure at a constant voltage loading. Through electrochemical reaction, Li ions are injected into cathode particles: (a) - (d) corresponding to time of 4.04, 20.1, 133.3 and 1226.5 sec (assuming the value of the diffusion coefficient of $10^{-10} \text{ cm}^2/\text{s}$).

3.3.3 Calculation of the electrode structure from the materials chemistry and the fabrication process

Alternatively, electrode micro-structures can be generated *in silico* by atomistic methods. To improve the understanding of the CL structure for example in PEMFCs, the effects of applicable solvent, particle sizes of primary carbon powders, wetting properties of carbon materials, and composition of the CL ink should be explored²⁷². These factors determine the complex interactions between Pt/C particles, ionomer molecules and solvent molecules and, therefore, control the catalyst layer formation process. Mixing the ionomer with dispersed Pt/C catalysts in the ink suspension prior to deposition will increase the interfacial area between ionomer and Pt/C nanoparticles. The choice of a dispersion medium determines whether ionomer is to be found in the solubilized, colloidal or precipitated forms.

Optimum performance of PEMFC in relation with H_2/O_2 transport limitations and water management for a number of parameters (type of agglomerate, CL thickness, CL porosity, distribution of Nafion[®] content, Pt loading, etc.) has been already largely investigated. Coarse Grained Molecular Dynamics (CGMD) models have been developed by Malek et al. to predict the self-organization of the CLs and to understand its impact on the effective transport and electrochemical properties²⁷³. CGMD is essentially a multiscale technique (parameters are directly extracted from classical atomistic MD) and account for

the conformational flexibility of ionomer molecules appropriately. CGMD simulations have been employed for characterizing microstructure of CL in view of effect of solvent, ionomer, and Pt particles.

In a recent work, Malek and Franco have proposed an approach to combine CGMD capabilities with kinetic modeling for the simulation of the feedback between detailed electrochemistry and transport with materials aging mechanisms: that means that at each numerical simulation time step, the model describes how the calculated local conditions impact local materials degradation kinetics, simultaneously to how the materials degradation affects, in the next time step, the local conditions (Figure 38).

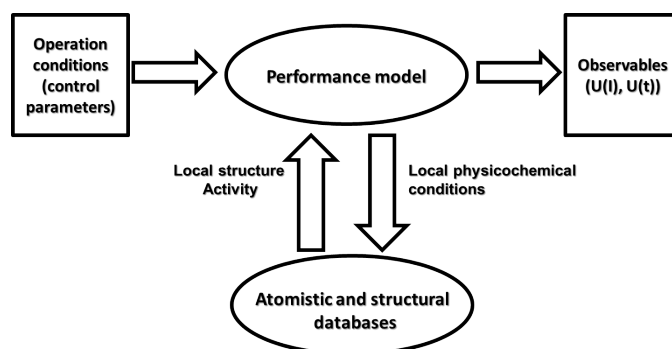


Figure 38. Multiscale modeling approach for the prediction of EPGs durability proposed by Franco et al.

CGMD simulations of a PEMFC electrode has been used to build a structural database for electrodes with different C contents in terms of interpolated mathematical functions describing the impact of the C mass loss (induced by corrosion) on the evolution of the ionomer coverage on Pt and C, the electronic conductivity of the CB, the C surface area and the Pt surface area (which re-organizes during the C corrosion process).²⁷⁴ These functions are then integrated into a cell model to simulate the impact of C corrosion on the Membrane-Electrodes Assembly performance decay (Figure 39). CGMD methods, which are actively researched in a large number of application areas, combine units of the material into larger fragments (called “beads”), which can be modeled efficiently using low-timescale methods, such as Brownian dynamics. Parameterization of the interactions of these units requires feedback from atomistic simulations. The details on this methodology for performing studies of self-organization in PEMFC electrodes mixtures have been described by Malek et al.,²⁷⁵ where they represent all atomistic and molecular species, i.e., Nafion[®] ionomer chains, solvent molecules, water, hydronium ions, carbon and Pt particles, by spherical metallic, polar, nonpolar, and charged beads with pre-defined sub-nanoscale length scale.

In these CGMD simulations, the corrosion process of carbon was simulated as follows: the carbon beads on the surface of carbon particles were first identified by characterizing all the water beads in contact with carbon beads. The equilibrated structure at I/C ratio of 0.9 and 1:1 Pt/C ratio was used as the starting point (Figure 40a). Thereafter, carbon beads are randomly removed from surface of carbon particles at different carbon loss percent-

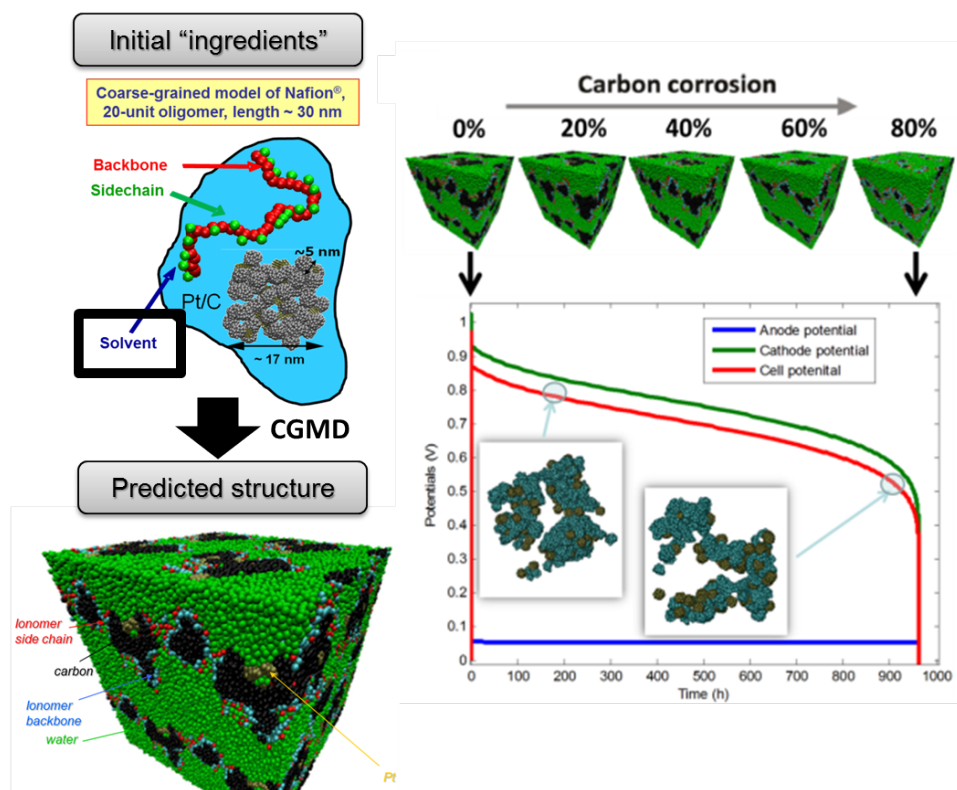


Figure 39. CGMD model of cathode carbon corrosion in PEMFCs. Figure reconstructed from K. Malek and A. A. Franco, *J. Phys. Chem. B* **115**, 8088 (2011).²⁷⁴

ages. The random procedure used a random number generator based on a given reaction probability distribution. Each of the structures was equilibrated after carbon removal. The gray beads in Figure 40b depict the eliminated carbon beads from the surface (i.e., corroded carbon which is proportional to carbon loss), whereas the black beads represent the remaining carbon beads.

Water coverage is an indirect tool to investigate the effect of ionomer and Pt content on microstructure of CL. Figures 41 and 42 illustrate water on carbon and Pt plotted vs. carbon losses. The inserts in Figure 41 show the microstructure of CL blends at various carbon losses. Water coverage, on the other hand, shows a steady increase (Figure 42) up to %40 carbon loss, maximizes at around %40, and decreases with further increasing of carbon loss (i.e., increasing I/C ratio). The calculated water coverage on Pt linearly increases by percentage of carbon loss. This suggests that a relatively high number of Pt particles are exposed to the Nafion ionomers on the surface of carbon, which causes a transition of the ionomer surface from predominantly hydrophobic to predominantly hydrophilic.

On the other hand, by increasing the percentage of carbon loss, the ionomer coverage (not shown here) drops from 0.5 at %0 carbon loss to slightly less than 0.4 at %5 carbon loss and stabilizes thereafter with a small variation between 0.4 to 0.35 %.

At a low carbon loss percentage, Pt nanoparticles on the carbon surface attract most of the water, while a relatively constant amount of water is adsorbed at the ionomer surface at different carbon losses. By increasing carbon loss, more Pt particles are exposed to water and the water coverage increases as a function of carbon loss, as depicted in Figure 42.

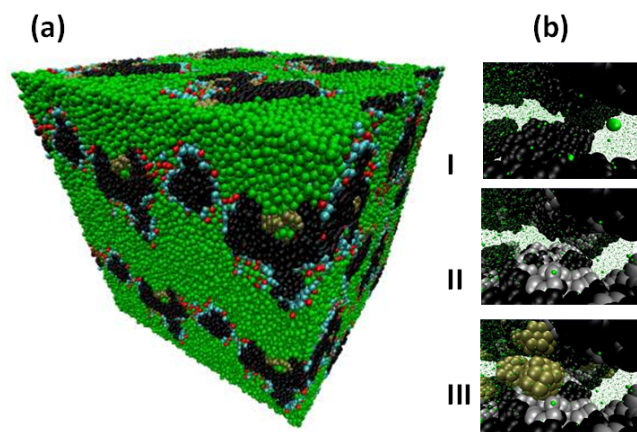


Figure 40. (a) The final structure of CL obtained from CGMD simulations. (b) Illustration of the algorithm used for modeling of carbon corrosion where carbon beads are randomly removed from surface of carbon particles at different carbon loss percentages. The gray beads depict the eliminated carbon beads from the surface (i.e., corroded carbon which is proportional to carbon loss), whereas the black beads represent the remaining carbon beads. Green: solvent beads; black: carbon beads; Gray: corroded carbon beads; blue: ionomer backbone; red: ionomer sidechain; gold: Pt. Source: Ref. 272.

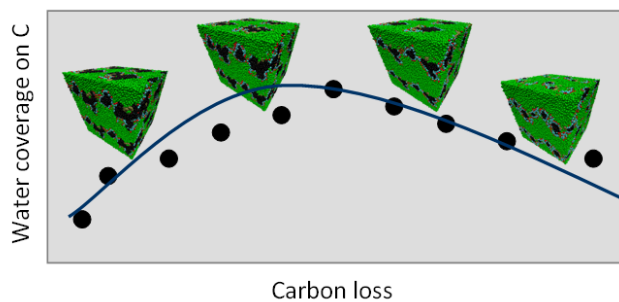


Figure 41. Calculated water coverage (including hydronium ions) on carbon as a function of carbon losses. The inserts show the microstructure of CL blends at various carbon losses. Source: Ref. 272.

Figure 43 shows the CL model which can be used for the performance decay calculations as the ones reported in Figure 39 by incorporating the CGMD data. Figure 44

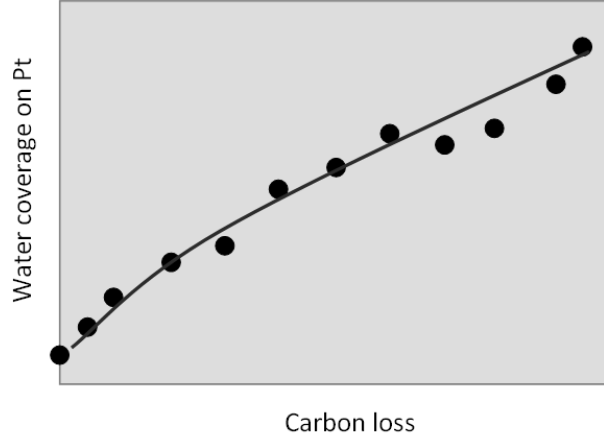


Figure 42. Calculated water coverage (including hydronium ions) on Pt as a function of carbon losses. Source: Ref. 272.

presents two kinds of micro-scale agglomerate models that can be implemented, one being representative of C support without primary pores, and the second one with primary pores the choice depending on the electrode design specifications.

For the case of a CL with only secondary pores the conservation equations for the individual gas species are written

$$\begin{aligned} \frac{\partial}{\partial t} ((1-s)\phi_{CL}^{SP} C_i) + \nabla_y \cdot (c_i v_g + j_i^d) &= S(y, z, t) + R(y, z, t) \\ &= S(y, z, t) + \gamma^{SP} J_i(r=0, y, z, t) \end{aligned} \quad (34)$$

where z is the spatial coordinate along the electrode thickness, x and y the coordinates on plane, S is a source term related to an elementary kinetic model as the one exposed in Section 3.1, R describes the rate of mass transfer between the ionomer film coating around the agglomerates and the secondary pores. γ^{SP} ($\text{m}^2 \cdot \text{m}^{-3}$) refers to the specific surface area of C secondary pores (“contact surface area” between the ionomer and the secondary pores per unit of CL volume). Because of the chemical-governed self-organization of the materials within the CL, ϕ_{CL}^{SP} , τ_{CL}^{SP} and γ^{SP} are functions of the catalyst, ionomer and C mass contents, and their values can be estimated from the CGMD-generated data. When aging of one or several of these materials occurs, the CL structure is expected to evolve. As a first approximation, we assume here that ϕ_{CL}^{SP} , τ_{CL}^{SP} and γ^{SP} only evolve with C corrosion -C mass loss- (ionomer degradation in the CL is not described yet within the model) following the CGMD databases. As C corrosion can be inhomogeneous within the CL (e.g. PEM side vs. GDL side differences or air inlet vs. air outlet differences induced by different local water contents), ϕ_{CL}^{SP} , τ_{CL}^{SP} and γ^{SP} are in fact functions of y and z coordinates. Catalyst degradation, inducing changes on its morphology, size and dispersion on the C support probably contributes on the CL meso-structure evolution.

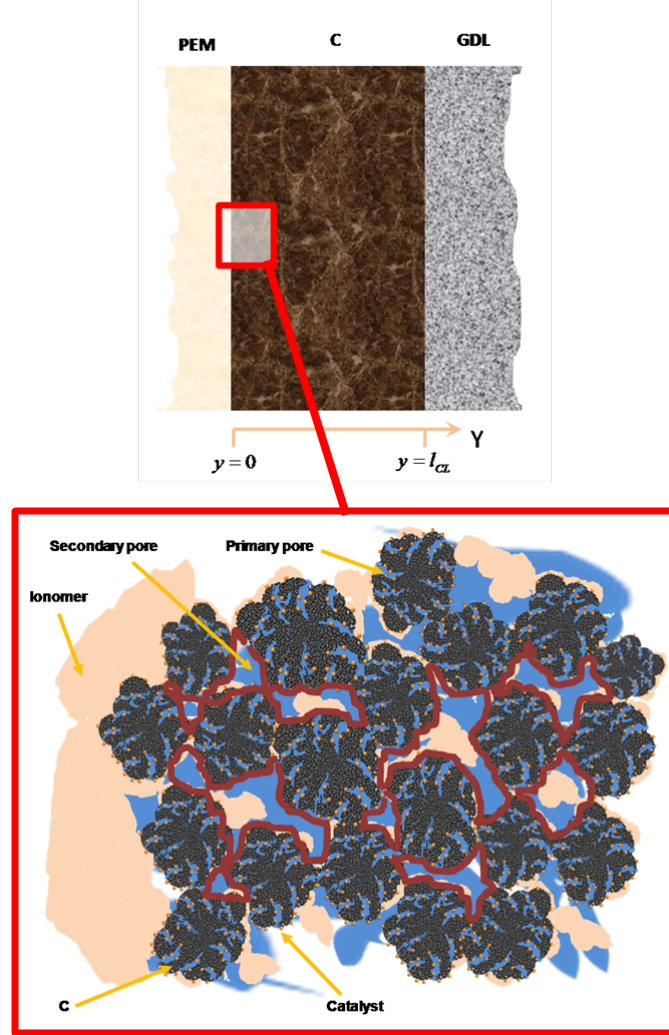


Figure 43. CL model at the macro and meso-scale where the boundaries defining the specific surface area of secondary pores γ^{SP} are indicated (in the figure, case of C support with primary pores).

The gas velocity and the species diffusion fluxes in equation (34) can be expressed based on Stefan-Maxwell-Knudsen approach with Knudsen coefficients $D_{i,Kn}$ and the absolute permeability K being also functions of the C mass (through the secondary pores mean radius), thus of time if the C degradation is included in the simulation ².

For the liquid water in the secondary pores $C_{H_2O,l}$, the conservation equation is given by

$$\frac{\partial}{\partial t}(sC_{H_2O,l}) + \nabla_y \cdot (sC_{H_2O,l}v_l) = -S(y,z,t) + W(y,z,t)$$

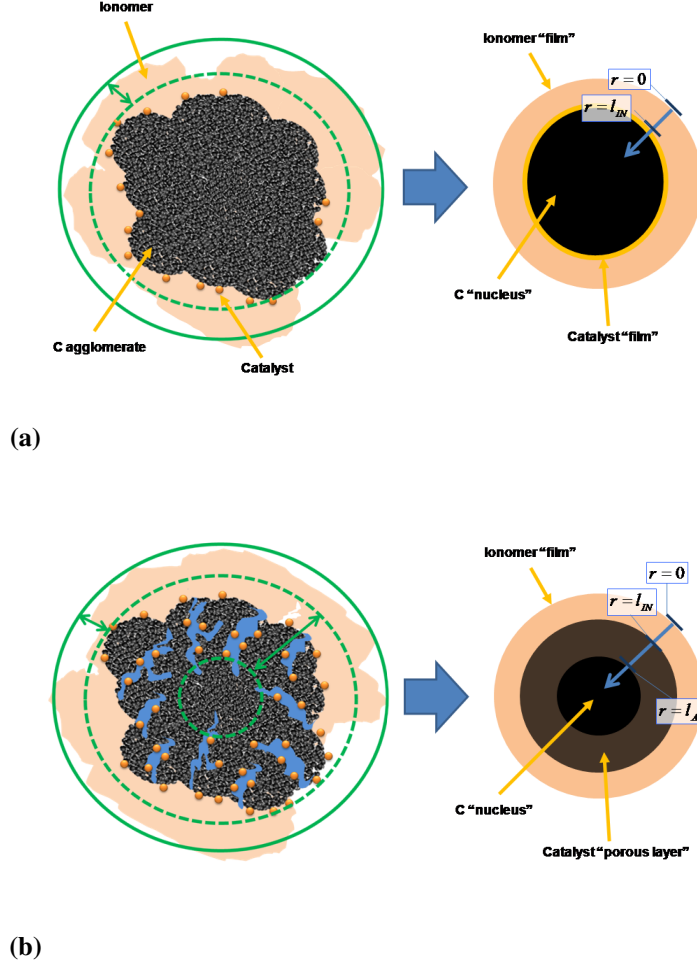


Figure 44. Modeled C agglomerates, under the spherical hypothesis, without (a) and with (b) primary pores. In particular, the ionomer phase is indicated.

$$= -S(y, z, t) + \gamma^{SP} k (C_{\text{H}_2\text{O}, \text{ionomer}}(y, z, t) - C_{\text{H}_2\text{O}, l}(y, z, t)) \quad (35)$$

where $W(y, z, t)$ is the rate of water transfer (desorption/absorption) from the ionomer film to the secondary pores (a linear kinetics is assumed).

For the electron transport across the C in the CL, we have

$$\begin{aligned} \nabla_y \cdot \left(-g_{e^-}^{CL, \text{eff}} \nabla_y \psi \right) &= \nabla_y \cdot \left(-(\wp_{CL}^C)^{\frac{\log_{10}(\wp_{CL}^C / \tau_{CL}^C)}{\log_{10}(\wp_{CL}^C)}} g_{e^-}^{CL} \nabla_y \psi \right) \\ &= S_{e^-}(y, z, t) = \pm \gamma_{\text{catalyst}} J(y, z, t) + \gamma^{SP} J_{\text{COR}}(y, z, t) \end{aligned} \quad (36)$$

where ϕ_{CL}^C is the C support volume fraction (function of time if C corrosion is included in the model), $\gamma_{catalyst}$ depends on time if a catalyst degradation mechanism and/or the C corrosion-driven catalyst coarsening are included in the simulation. Again, $\gamma_{catalyst}$ can also depend on space because of the degradation inhomogeneities induced, for example, by the local water content within the CL. $J_{COR}(y, z, t)$ is the local current density related to the C corrosion kinetics (Figure 45)

$$J_{COR}(y, z, t) = \sum_i v_i^{COR} \quad (37)$$

where v_i^{COR} are the elementary reactions depending on the local water content in the ionomer phase $C_{H_2O, ionomer}$.

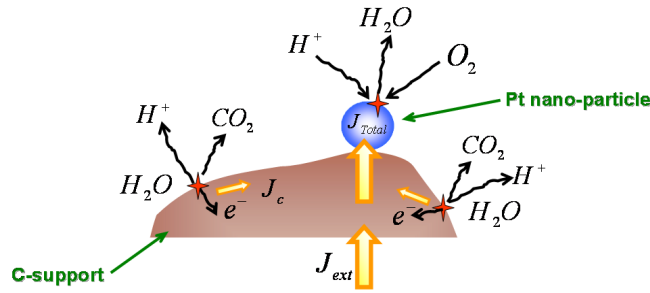


Figure 45. Nano-scale parasitic current related to C corrosion. Source: ²⁷⁶.

Finally, the following balance equation allows calculating the instantaneous C mass content in the CL level based on the elementary C corrosion kinetics ^{274,276}

$$m_C(y, z, t) = m_C(y, z, t = 0) - \int_0^t \left(\sum_j v_j^{COR} \right) M_C \gamma^{SP}(y, z, t) V_{CL@}(y, z) dt. \quad (38)$$

The carbon corrosion model for the case of CL with both primary and secondary pores will be discussed in the hands-on session associated to this tutorial.

This modeling approach has provided very interesting information on the competition of aging phenomena. Some experimental data suggests that external anode and cathode contaminants (e.g. CO in the anode, SO₂ in the cathode) can enhance the damage of the PEMFC materials. But according to some modeling work carried out by Franco et al. with this approach, the injection of these contaminants can mitigate, under appropriate current-cycled conditions, the intrinsic materials aging mechanisms as demonstrated based on an approach combining experiments and this multiscale numerical model ¹⁵; this work clearly illustrated the interest of treating the complex mechanisms interacting between them towards engineering optimization of the PEMFC operation. More arguments on the importance of modeling and simulating of mechanisms in interaction in EPGs will be provided in the oral presentation associated to this lecture.

4 Conclusions and Challenges

Numerical simulation and computer-aided engineering emerges nowadays as important tools to speed up the EPGs R&D and to reduce their time-to-market for numerous applications.

The development of such models must have several properties

- predictive capabilities of the relative contributions of the different scales and mechanisms into the macroscopic EPGs efficiency and durability;
- high flexibility towards its application to any type of chemical and structural properties of the used materials and components;
- easily adaptable to any type of operation condition and system.
and will enable
- reduction of the amount of experiments (and thus the cost) currently needed to build up classical empirical models with limited prediction capabilities;
- a better targeting of experimental characterizations in representative conditions of the end-user application;
- new operation strategies reducing the performance degradation and also strategies to improve the stability of the materials and components.

Numerous theoretical efforts to mathematically describe the EPGs operation have been reported worldwide since around 30 years. Physical models became more and more accurate and predictive, for example thanks to the widespread development of quantum mechanics and molecular dynamics models allowing capturing the impact of the materials chemistry onto some effective properties essentially related to the electrochemical reactions and lithium ion transport. Non-equilibrium thermodynamics phase field modeling approach appears to be a powerful modeling technique to understand phase formation and separation for example in LIB intercalation materials at the nano/microscale. Moreover, impressive electrode 3D reconstruction techniques have been developed allowing to capture for the first time the impact of the “real” mesostructure (e.g. binder distribution) onto the local lithium reactivity and transport properties and the global cell efficiency.

In particular, integrative multiphysics, multiscale and multiparadigm models spanning multiple scales and aiming to simulate competitions and synergies between electrochemical, transport, mechanical and thermal mechanisms become now available. One of the major interests of such a class of models is its capability to analyze the impact of the materials and components structural properties on the global cell performance. Some of the reported models have been already used to understand materials degradation phenomena and their impact on the EPGs efficiency and capacity fade: metallic dissolution (e.g. in PEMFCs, PEMWEs and LIBs), carbon corrosion (e.g. in PEMFCs), PEM degradation (e.g. in PEMFCs and PEMWEs), SEI formation (e.g. in LIBs and LABs) and graphite exfoliation (e.g. in LIBs) are some of the mechanisms currently studied.

Despite the tremendous progress achieved on developing multiscale models of EPGs with predictive capabilities, there are still major challenges to be overcome.

Demonstrating a durability of several thousands of hours for PEMFCs, up to 60000 hours for some applications, is thus now the prime requirement. The longer the durability, the higher the price customers will be ready to pay for their investment. Demonstrating such a long life-time in real operating conditions within an autonomous system is a real challenge. System failures often impair satisfactory demonstrations of the stack reliability. Statistical proof of a repeated success in achieving a long life time asks for a large number of units operating at customers facilities, which is very costly. Before launching such large scale demonstrations, necessary when approaching market maturity, fuel cell system providers need a reliable prediction of their products lifetime.

The main bottleneck now, if one wants to shorten the “time to market” for new PEM-FCs, is that one needs an efficient method to take into account durability targets in all R&D actions on components and unit operating management strategies. A simple “try and error” method is manageable to get system durability from a few hundreds to a few thousands of hours. It is practically impossible when one wants to get from a few thousands of hours up to several tens of thousands of hours.

This lecture aimed to demonstrate that modeling at multiple scales (from atomistic level to the simulation of processes at the cell level) can constitute a reliable method to predict the EPG system performance and lifetime and to benchmark components and improve operating strategies with respect to a durability target. Predictive modeling is a requisite to establish this methodology for EPG. Performance and durability of a EPG is the result of a very complex set of interrelated events, with competitive effects but also synergies between performance degradation processes. By accelerating one phenomenon, one usually creates conditions that are no more representative of the subtle balance between reactions in the real operating conditions. One can thus either overestimate performance losses degradation rates (leading to developing very highly resistant membranes, support carbons or catalysts for conditions never encountered in a real system), or underestimate degradation because some negative feedback loops (cancellation or synergetic effects) are not taken into account. Only a physically based, multiscale and multitemporal model can provide the tool to combine all possible degradation phenomena and analyze their global impact on durability in a given set of operating conditions.

As an input for this model, one has to understand the fundamentals of performance losses and degradation. Specific experimental data is needed, to correlate degradation and deterioration phenomena to operating conditions for stationary applications, and identify the paths leading to failure phenomena. This experimental part is an effort to obtain data on degradation quantitatively and reproducibly, since the understanding of kinetics of various degradation processes are the key of final performance and lifetime prediction.

Numerical simulation of competitive degradation phenomena on the basis of a bottom-up framework should also be achieved, in order to determine the most important mechanisms as function of the applied external operation conditions and to quantitatively predict the EPG durability. Complete models aiming to simulate and understand aging, contamination and performance mechanisms in competition will be of significant importance for the prediction of the durability of PEMFCs in automotive conditions.

For the case of the electrocatalysis, more efforts should be devoted to perform *ab initio* calculations with solvent and electric field, in relation to the catalyst dissolution and oxidation (how the activity is affected by the catalyst degradation, and conversely, how the catalyst degradation kinetics is affected by the intrinsic catalytic activity?).

Ab initio thermodynamics models have been largely developed to screen the activity, selectivity and stability of catalyst candidates and have allowed the selection of the most interesting ones. Those predicted catalysts can be then tested experimentally, and sometimes they really work and sometimes not. The catalysts that do not work in real conditions probably do not work because of the lack of consideration of electrochemical environment (electric field, solvent, etc.) in classical *ab initio* thermodynamics approach. Then complementary *ab initio* kinetic models can provide a “virtual simulator framework” to achieve the optimization of the operation conditions for which the predicted catalyst should work in the real environment (Figure 46). Thus, more efforts should be developed within this sense.

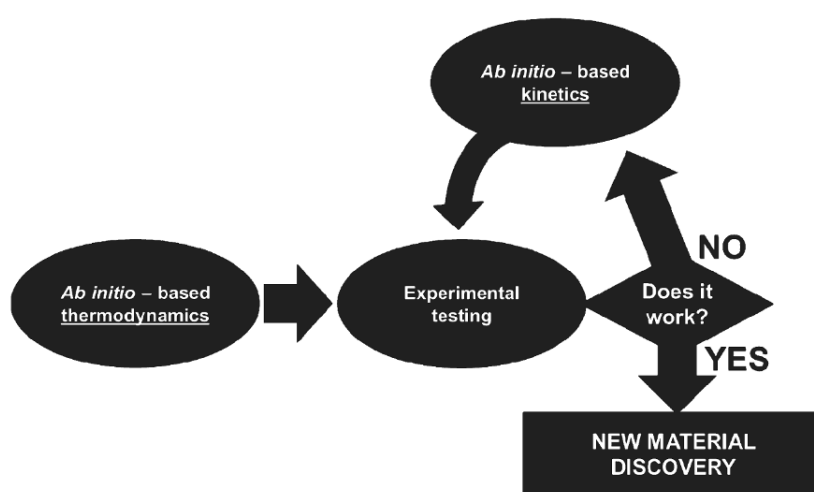


Figure 46. Towards a physical modeling assisted materials development through *ab initio* based kinetics and the virtual reactor approach. Reprinted from: A.A. Franco, RSC Advances, 3 (32) (2013) 13027-13058.

For the case of batteries the majority of the reported multiscale models focus on the understanding of the operation and the impact of the structural properties of LiFePO₄ or graphite electrodes onto the global cell efficiency. And in the other hand, quantum mechanics and molecular dynamics models focus on the understanding of the impact of the materials chemistry onto their storage or lithium transport properties at the nanoscale. It is now crucial to develop multiscale models that are able to incorporate both structure and chemical databases, in other words, that they are able to mimic the materials behavior in realistic electrochemical environments. Within this sense, other intercalation and conversion materials have to be also modeled. The development of such a model tackles the issues related to how to couple discrete with continuum models and will need to set up rigorous methods to integrate *ab initio* data into elementary kinetics models of the lithiation/delithiation reactions.

Moreover, accurate modeling of the interfacial electrochemical reactions that com-

bine chemistry with diffusion of radicals and formation of the heterogeneous crystalline or glassy SEI layer of anode or a passivation layer on the cathode is an intrinsically multiscale problem, which is largely unaddressed but of great technological relevance. Phase field models describing the lithium kinetics on the basis of parameters which can in principle be estimated from quantum mechanics calculations, would be the key to achieve these goals.

Secondly, computational tools for the analysis of performance and degradation of fuel cells and batteries are currently fragmented and developed independently at different groups. Ideally, a multiscale model useful in engineering practice should have the following characteristics:

- it should be flexible, i.e. it should allow the developers to “virtually test” different cell designs to decide which cell satisfies their technical needs, as well as quickly studying new cell designs;
- it should be portable, i.e. the model should be a computer (or computing platform) independent code;
- it should be scalable, i.e. the model code should allow developers to run it on single computers and multicore processors up to supercomputers;
- it should be easy to use, i.e. the model implementation details should be abstracted in a way that the developer interacts with an user-friendly interface;
- it should allow cloud computing and network development, i.e. simultaneous development by multiple researchers should be permitted as well as a performing exchange of information to benchmarking different model versions;
- web platforms should be developed aiming to share physics and mathematical modules to facilitate different models calibration and benchmarking.

Approaches synergistically combining both top down and bottom up modeling viewpoints should be further developed. Macroscopic equations in top-down models should be written in terms of parameters with values calculated from lower scale simulations. Implementation of such parameters into the macroscopic model should be done including empirical errors. Methodological evaluation of these parameters should be done systematically: for instance, coarse models should be developed first, with parameters sensitivity studies guiding further calculations at lower scales.

More from a materials engineering perspective, morphogenesis of the electrodes as function of the ink properties and manufacture process (e.g. solvent used, deposition time, etc.), should be further studied.

Furthermore, the generation and the integration of those chemical and structural databases in the phase field and cell models should be carried out in a fully integrated way by exploiting the recent progresses in data flows management.

More generally, combining multiscale modeling with the use of virtual reality could provide significant progress on providing virtual experimentation of EPGs.²⁷⁷

Finally, to get progress on the development of multiscale models, it is crucial to develop multidisciplinary between application domains. For example, computational scientists

working in cosmology, geology and climate science could bring interesting methodological concepts for the widespread use of multiscale modeling in electrochemistry.

It should be finally noticed that the analysis and discussions here above also applies for other electrochemical systems for energy storage and conversion, such as redox flow batteries, lithium sulfur batteries, Direct Alcohol Fuel Cells, etc.^{103,278,279,280}

Finally, since recently, Franco et al. at LRCS is developing a new multiscale computational framework of electrochemical devices for energy storage and conversion (www.modeling-electrochemistry.com) which aims to propose solutions to some of the challenges exposed in this Section. This new model, called MS LIBER-T (*Multiscale Simulator of Lithium Ion Batteries and Electrochemical Reactor Technologies*), constitutes a breakthrough compared to the previously developed MEMEPhys simulation package, also by Franco et al., penalized by its dependence on commercial software toolboxes and solvers such as Simulink. MS LIBER-T is coded on an independent C language basis, highly flexible and portable (it can be eventually coupled to commercial software such as Matlab/Simulink), and supports direct multiparadigm calculations, for instance, simulations coupling on the fly the numerical resolution of continuum models (e.g. describing reactants transport in a bulk) with the numerical resolution of discrete models (e.g. Kinetic Monte Carlo codes resolving detailed electrocatalytic reactions). Some demonstrations with this software will be carried out in the oral presentation associated to this lecture.

References

1. M. Mohseni, B. Ramezanzadeh, H. Yari, M.M. Gudarzi, The role of nanotechnology in automotive industries, book chapter in: *New Advances in Vehicular Technology and Automotive Engineering*, J.P. Carmo and J. E. Ribeiro Eds., InTech (2012).
2. A. A. Franco, "A multiscale modeling framework for the transient analysis of PEM-FCs - From theory to the engineering practice"; Habilitation to become Research Director (H.D.R.) manuscript, Université Claude Bernard Lyon-1 (France) (2010) (available for download in www.modeling-electrochemistry.com).
3. A.A. Franco, Multiscale modeling of electrochemical devices for energy conversion and storage, book chapter in: *Encyclopedia of Applied Electrochemistry*, edited by R. Savinell, K.I. Ota, G. Kreysa (publisher: Springer, UK) (2013).
4. D. Jollie, Fuel Cell Market Survey: Portable Applications, www.fuelcelltoday.com, 6 September 2005.
5. M. Mathias, H.A. Gasteiger, *Electrochemical Society Proceedings*, vol. PV 2002-31, in: M. Murthy, T.F. Fuller, J.W. Van Zee, S. - Gottesfeld (Eds.), *Third International Symposium on PEM Fuel Cells*, Salt Lake City, UT (2002).
6. X. Wang, R. Kumar, D. J. Myers, *Electrochemical and Solid-State Letters*, **9** (5) (2006) A225.
7. K.E. Swider-Lyons, M.E. Teliska, W.S. Baker, P.J. Bouwman, J.J. Pietron, *ECS Transactions*, **1** (6) (2005) 97.
8. H. Xu, R. Kunz, J. M. Fenton, *Electrochemical and Solid-State Letters*, **10** (1) (2007) B1.
9. K. Yazuda, A. Taniguchi, T. Akita, T. Ioroi, Z. Siroma; *Phys. Chem. Chem. Phys.*, **8** (6), (2006) 746.

10. P. J. Ferreira, Y. Shao-Horn, *Electrochemical and Solid-State Letters*, **10** (3) (2007) B60.
11. H. Tang, Z. Qi, M. Ramani, J. F. Elter, *J. Power Sources*, **158**, (2) (2006) 1306.
12. M Schulze, C. Christenn; *Appl. Surf. Sci.*, **252** (2005) 148.
13. V. O. Mittal, H. R. Kunz, J. M. Fenton, *Electrochemical and Solid-State Letters*, **9** (6) (2006) A299.
14. O. Lemaire, B. Barthe, L. Rouillon, A.A. Franco, "Mechanistic Investigation of NO₂ Impact on ORR in PEM Fuel Cells: A Coupled Experimental and Multi-Scale Modeling Approach", *ECS Trans.*, **25** (1) (2009) 1595.
15. A.A. Franco, M. Guinard, B. Barthe, O. Lemaire, "Impact of carbon monoxide on PEFC catalyst carbon support degradation under current-cycled operating conditions", *Electrochim. Acta*, **54** (22) (2009) 5267.
16. B. Pivovar et al, DOE FY 2005 Progress Report (2005).
17. F. Barbir, *Sol. Energy*, 2005, **78**, 661-669.
18. W. T. Grubb, *J. Electrochem. Soc.*, 1959, **106**, 275-281.
19. P. Millet, R. Ngameni, S. Grigoriev, N. Mbemba, F. Brisset, A. Ranjbari and C. Etivant, *Int. J. Hydrogen Energy*, 2010, **35**, 5043-5052
20. A. Marshall, B. Borresen, G. Hagen, M. Tsyppkin and R. Tunold, *Energy*, 2007, **32**, 431-436.
21. E. Rasten, G. Hagen and R. Tunold, *Electrochim. Acta*, 2003, **48**, 3945-3952.
22. S. Song, H. Zhang, X. Ma, Z. Shao, R. T. Baker and B. Yi, *Int. J. Hydrogen Energy*, 2008, **33**, 4955-4961
23. V. Baglio, A. D. Blasi, T. Denaro, V. Antonucci, A. S. Aric, R. Ornelas, F. Matteucci, G. Alonso, L. Morales, G. Orozco and L. G. Arriaga, *J. New Mater. Electrochem. Syst.*, 2008, **11**, 105-108
24. D. Čukman, M. Vuković and M. Milun, *J. Electroanal. Chem.*, 1995, **389**, 209-213
25. S. Fierro, T. Nagel, H. Baltruschat and C. Comninellis, *Electrochem. Commun.*, 2007, **9**, 1969-1974.
26. K. Macounova, M. Makarova and P. Krtil, *Electrochem. Commun.*, 2009, **11**, 1865-1869.
27. P. Millet, M. Pineri and R. Durand, *J. Appl. Electrochem.*, 1989, **19**, 162-166.
28. L. Ma, S. Sui and Y. Zhai, *Int. J. Hydrogen Energy*, 2009, **34**, 678-684.
29. H. Ito, T. Maeda, A. Nakano, Y. Hasegawa, N. Yokoi, C. Hwang, M. Ishida, A. Kato and T. Yoshida, *Int. J. Hydrogen Energy*, 2010, **35**, 9550-9560.
30. H. Ito, T. Maeda, A. Nakano and H. Takenaka, *Int. J. Hydrogen Energy*, 2011, 1-14.
31. L.F. Lopes Oliveira, S. Laref, E. Mayousse, C. Jallut, A.A. Franco, "A multi-scale physical model for the transient analysis of PEM Water Electrolyzer Anodes", *Phys. Chem. Chem. Phys.*, **14**(2012)10215.
32. M.S. Whittingham, *Chem. Rev.*, **104** (2004) 4271.
33. A. K. Padhi, K. S. Nanjundaswamy, J. B. Goodenough, *J. Electrochem. Soc.*, **144** (4) (1997) 1188.
34. M. Maccario, L. Croguennec, F. Weill, F. Le Cras, C. Delmas, *Solid State Ionics*, **179** (2008) 2383.
35. D. Choi, P. N. Kumta, *J. Power Sources*, **163** (2007) 1064.
36. H. Huang, S.-C. Yin, L. F. Nazar, *Electrochem. Solid-State Lett.*, **4** (2001) A170.
37. S.-T. Myung, S. Komaba, N. Hirosaki, H. Yashiro, N. Kumagai, *Electrochim. Acta*,

- 49** (2004) 4213.
38. A. Aimable, D. Aymes, F. Bernard, F. Le Cras, *Solid State Ionics*, **180** (2009) 861.
 39. T. Takeuchi, M. Tabuchi, A. Nakashima, T. Nakamura, Y. Miwa, H. Kageyama, K. Tatsumi, *J. Power Sources*, **146** (2005) 575.
 40. K. Xu, U. Lee, S.S. Zhang, M. Wood, T.R. Jow, *Electrochem. Solid State Lett.*, **6** (2003) A144.
 41. J. Cabana, Z. Stoeva, J. J. Titman, D. H. Gregory, M. R. Palacín, *Chem. Mater.*, **20** (2008) 1676.
 42. A.A. Franco, Multiscale modeling and numerical simulation of rechargeable lithium ion batteries: concepts, methods and challenges, *RSC Advances*, **3** (32) (2013) 13027-13058.
 43. T.F. Fuller, M. Doyle, J. Newman, *J. Electrochem. Soc.*, **141** (1994) 982.
 44. M.D. Levi, G. Salitra, B. Markovsky, H. Teller, D. Aurbach, U. Heider, L. Heider, *J. Electrochem. Soc.*, **146** (1999) 1279.
 45. S. Kobayashi, Y. Uchimoto, *J. Phys. Chem.*, **109** (2005) 13322.
 46. J. Goldstein, I. Brown, and B. Koretz, *J. Power Sources* **80**, 171 (1999).
 47. Q. Sun, Y. Yang, and Z.-W. Fu, *Electrochem. Commun.* **16**, 22 (2012).
 48. W. Li, C. Li, C. Zhou, H. Ma, and J. Chen, *Angew. Chem.* **118**, 6155 (2006).
 49. M. L. Doche, F. Novel-Cattin, R. Durand, and J. J. Rameau, *J. Power Sources* **65**, 197 (1997).
 50. P. Hartmann, C. L. Bender, M. Vračar, A. K. Dürr, A. Garsuch, J. Janek, and P. Adelhelm, *Nat. Mater.* **12**, 228 (2013).
 51. K. M. Abraham and Z. Jiang, *J. Electrochem. Soc.* **143**, 1 (1996).
 52. K. M. Abraham, *ECS Trans.* **3(42)**, 67 (2008).
 53. G. Girishkumar, B. McCloskey, A. C. Luntz, S. Swanson, and W. Wilcke, *J. Phys. Chem. Lett.* **1**, 2193 (2010).
 54. J. Read, K. Mutolo, M. Ervin, W. Behl, J. Wolfenstine, A. Driedger, and D. Foster, *J. Electrochem. Soc.* **150**, A1351 (2003).
 55. A. Débart, A. J. Paterson, J. Bao, and P. G. Bruce, *Angew. Chem. Int. Edit.* **47**, 4521 (2008).
 56. W. Xu, J. Xiao, J. Zhang, D. Wang, and J.-G. Zhang, *J. Electrochem. Soc.* **156**, A773 (2009).
 57. J. Xiao, D. Wang, W. Xu, D. Wang, R. E. Williford, J. Liu, and J.-G. Zhang, *J. Electrochem. Soc.* **157**, A487 (2010).
 58. T. Kuboki, T. Okuyama, T. Ohsaki, and N. Takami, *J. Power Sources* **146**, 766 (2005).
 59. J. Xiao, W. Xu, D. Wang, and J.-G. Zhang, *J. Electrochem. Soc.* **157**, A294 (2010).
 60. A. Débart, J. Bao, G. Armstrong, and P. G. Bruce, *J. Power Sources* **147**, 1177 (2007).
 61. A. K. Thapa and T. Ishihara, *J. Power Sources* **196**, 7016 (2011).
 62. A. K. Thapa, K. Saimen, and T. Ishihara, *Electrochem. Solid-State Lett.* **13**, A165 (2010).
 63. R. R. Mitchell, B. M. Gallant, C. V. Thompson, and Y. Shao-Horn, *Energ. Environ. Sci.* **4**, 2952 (2011).
 64. F. Mizuno, S. Nakanishi, Y. Kotani, S. Yokoishi, and H. Iba, *Electrochemistry* **78**, 403 (2010).

65. B. D. McCloskey, D. S. Bethune, R. M. Shelby, G. Girishkumar, and A. C. Luntz, *J. Phys. Chem. Lett.* **2**, 1161 (2011).
66. J. Xiao, J. Hu, D. Wang, D. Hu, W. Xu, G. L. Graff, Z. Nie, J. Liu, and J.-G. Zhang, *J. Power Sources* **196**, 5674 (2011).
67. G. Girishkumar, B. McCloskey, A. C. Luntz, S. Swanson, and W. Wilcke, *J. Phys. Chem. Lett.* **1**, 2193 (2010).
68. C. Tran, X.-Q. Yang, and D. Qu, *J. Power Sources* **195**, 2057 (2010).
69. D. Zheng, H.-S. Lee, X.-Q. Yang, and D. Qu, *Electrochem. Commun.* **28**, 17 (2013).
70. M. Eikerling, *J. Electrochem. Soc.* **153**, E58 (2006).
71. B. E. Conway, *Electrochemical Supercapacitors*, Kluwer-Plenum, New York (1999).
72. A. Burke, *J. Power Sources*, **91** (2000) 37.
73. F. Lufrano, P. Staiti, M. Minutoli, *J. Electrochem. Soc.*, **151** (1) (2004) A64-A68.
74. Yong Zhang, Hui Feng, Xingbing Wu, Lizhen Wang, Aiqin Zhang, Tongchi Xia, Huichao Dong, Xiaofeng Li, Linsen Zhang, *International journal of hydrogen energy* **34** (2009) 4889-4899.
75. Mingjia Zhi, Chengcheng Xiang, Jiangtian Li, Ming Li, and Nianqiang Wu, *Nanoscale*, 2013, **5**, 72-88.
76. B. Babakhani and D. G. Ivey, *Electrochim. Acta*, 2010, **55**, 4014.
77. S. Sarangapani, B. V. Tilak and C. P. Chen, *J. Electrochem. Soc.*, 1996, **143**, 3791.
78. A. Chu and P. Braatz, *J. Power Sources*, **112**, 236 2002.
79. R. Kotz and M. Carlen, *Electrochim. Acta*, **45**, 2483 2000.
80. S. Nomoto, H. Nakata, K. Yoshioka, A. Yoshida, and H. Yoneda, *J. Power Sources*, **97-98**, 907 2001.
81. G. Wang, L. Zhang, J. Zhang, *Chem. Soc. Rev.*, **41** (2012) 797.
82. K. Gödel, *Monatshefte für Mathematik und Physik*, **38** (1931) 173.
83. G.A. Holzapfel, T.C. Gasser, *Computer Methods in Applied Mechanics and Engineering*, **190** (2001) 4379.
84. F. Couenne, D. Eberard, L. Lefèvre, C. Jallut, B. Maschke, European Symposium on Computer Aided Process Engineering - 15, L. Puigjaner and A. Espuña, Eds..
85. R. Marchand, T. Ackerman, *J. of Geophysical Research*, **115** (2010) D16207.
86. M. Stan, *Nuclear Engineering and Technology*, **41** (1) (2009).
87. A.A. Baklanov, B. Grisogono, R. Bornstein, L. Mahrt, S.S. Zilitinkevich, P. Taylor, S.E. Larsen, M.W. Rotach, H.J.S. Fernando,, *Bulletin of the American Meteorological Society*, **92** (2011) 123.
88. J.D. Cole, *SIAM J. Applied Maths*, **55** (1995) 410.
89. M. Ptashnyk, T. Roose, *SIAM J. Applied Maths*, **70** (2010) 2097.
90. M. Ptashnyk, T. Roose, G. Kirk, *Eur. J. Soil Sci.*, **61** (2010) 108.
91. G. Botte, V. R. Subramanian, R.E. White, *Electrochim. Acta*, **45** (2000) 2595.
92. M. Doyle, J. Newman, *J. Appl. Electrochem.*, **27** (1996) 846.
93. M. Doyle, J. Newman, *J. Electrochem. Soc.*, **143** (1996) 1890.
94. T.F. Fuller, M. Doyle, J. Newman, *J. Electrochem. Soc.*, **141** (1994) 1.
95. T.F. Fuller, M. Doyle, J. Newman, *J. Electrochem. Soc.*, **141** (1994) 982.
96. C.Y. Wang, W.B. Gu, B.Y. Liaw, *J. Electrochem. Soc.*, **145** (1998) 3407.
97. L. Madec, L. Falk, E. Plasari, *Chemical Engineering Science*, **56** (2001) 1731.
98. P. Mandin, J.M. Cense, F. Cesimiro, C. Gbado, D. Lincot, *Computers & Chemical Engineering*, **31** (8) (2007) 980.

99. K. Reuter, O. Deutschmann (Ed.), Wiley-VCH, Weinberg (2009).
100. D. Sheppard, R. Terrell, G. Henkelman, *J. Chem. Phys.* **128** (2008) 134106.
101. R. Ferreira de Moraes, D. Loffreda, P. Sautet, A. A. Franco, *Electrochim. Acta*, **56** (28) (2011) 10842.
102. K. Malek, A.A. Franco, *J. Phys. Chem. B*, **115** (25) (2011) 8088.
103. A.A. Franco, Towards a bottom-up multiscale modeling framework for the transient analysis of PEM Fuel Cells operation, book chapter in: Polymer Electrolyte Fuel Cells: Science, Applications and Challenges, edited by A.A. Franco (publisher: Pan Stanford, distributor: Francis & Taylor) (2013).
104. A.A. Franco, "A multiscale modeling framework for the transient analysis of PEM-FCs - From theory to the engineering practice"; Habilitation to become Research Director (H.D.R.) manuscript, Université Claude Bernard Lyon-1 (France) (2010) (available online).
105. L. Cai, R. E. White, *Journal of Power Sources*, **196** (2011) 5985.
106. <http://www.mcs.anl.gov/petsc/>
107. <http://www.zib.de/ehrig/software.html>
108. <http://www.ctcms.nist.gov/fipy/>
109. G. Hager, *Introduction to High Performance Computing for Scientists and Engineers*, Chapman & Hall/CRC Computational Science (2010).
110. <https://www.vasp.at/>
111. <http://www.crystal.unito.it/>
112. <http://www.scm.com/>
113. <http://www.gaussian.com/>
114. http://bigdft.org/Wiki/index.php?title=BigDFT_website
115. M. A. Gabriel, T. Deustch, L. Genovese, G. Krosnicki, O. Lemaire, A. A. Franco, *Phys. Chem. Chem. Phys.* **12** (2010) 9406.
116. M. A. Gabriel, T. Deustch, A. A. Franco, *ECS Trans.*, **25** (22) (2010) 1.
117. <http://www.gromacs.org/>
118. <http://lammps.sandia.gov/>
119. <http://ambermd.org/>
120. <http://www.charmm.org/>
121. D. Frenkel, B. Smit, *Understanding Molecular Simulation: from algorithms to applications*, Academic Press (2002).
122. S. Plimpton, *Journal of Computer Physics*, **117** (1995) 1.
123. D.C. Rapaport The art of molecular dynamics simulation Cambridge University Press, 2009
124. M.P.Allen, D.J. Tildesley, *Computer simulation of liquids*, Clarendon Press (1989).
125. S. Bozic, I. Kondov, *Dataflow Management: A Grand Challenge in Multiscale Materials Modelling*, P. Cunningham and M. Cunningham, (Eds.), *eChallenges e-2012 Conference Proceedings*, IIMC International Information Management Corporation, 2012.
126. <http://www.knime.org/>
127. <http://www.unicore.eu/index.php>
128. W. R. Elwasif, D. E. Bernholdt, S. Pannala, S. Allu, S. S. Foley, cse, pp.102-110, 2012 IEEE 15th International Conference on Computational Science and Engineering, 2012.

129. M.C. Georgiadis, S. Myrian, N. Efstratios, R. Gani, *Computers and Chemical Engineering*, **26** (2002) 735.
130. M. Mangold, S. Motz, E.D. Gilles, *Chemical Engineering Science*, **57** (2002) 4099.
131. P. Breedveld, F. Couenne, C. Jallut, B. Maschke, M. Tayakout - Fayolle, Proc. 4th European Congress on Chemical Engineering, Grenada, Spain (2003).
132. F. Couenne, C. Jallut, B. Maschke, P. Breedveld, M. Tayakout, *Mathematical and Computer Modeling of Dynamical Systems*, **12** (2-3) (2006) 159.
133. P.C. Breedveld, *Physical systems theory in terms of bond graphs*, PhD thesis, University of Twente, Enschede, Netherlands (1984).
134. P.C. Breedveld, G. Dauphin-Tanguy (eds.), *Current topics in bond graph related research*, Journal of The Franklin Institute, Special issue on bond-graph modeling, **328** (5-6), Pergamon Press (1991).
135. H.M. Paynter, *Analysis and design of engineering systems*, MIT Press, Cambridge, MA (1961).
136. J.U. Thoma, *Introduction to bond graphs and their applications*, Pergamon Press, Oxford (1975).
137. J.U. Thoma, *Simulation by bond graphs - Introduction to a graphical method*, Springer Verlag (1989).
138. <http://www.20sim.com/>
139. <http://www.lmsintl.com/amesim-platform>
140. <http://www.3ds.com/products/catia/portfolio/dymola>
141. <https://www.openmodelica.org/>
142. L. Menard, G. Fontes, S. Astier, *Journal of Mathematics and Computer Simulation*, **81** (2) (2010) 327.
143. J.J. Esperilla, J. Felez, G. Romero, A. Carretero, *Simulation Modeling Practice and Theory*, **15** (1) (2007) 82.
144. A.A. Franco, PhD thesis, Université Claude Bernard Lyon 1 (2005) (available online).
145. A. A. Franco, C. Jallut, B. Maschke, "Multi-scale Bond Graph Model of the Electrochemical Dynamics in a Fuel Cell", in Proc. 5th MATHMOD Conference, (Eds. I Troch and F. Breitenecker), Vienna, Austria (2006) P103.
146. A. A. Franco, M. Tembely, C. Jallut, B. Maschke, in Proc. World Hydrogen Conference (WHEC) 16, Lyon, France (2006) 572.
147. A.A. Franco, C. Jallut, B. Maschke, in preparation (2013).
148. L. Ljung, *System Identification - Theory for the User*, Prentice-Hall, Upper Saddle River, New Jersey, 2nd edition (1999).
149. L. Ljung, T. Glad, *Automatica*, **30** (2) (1994) 265.
150. E. Walter, *Identification of State Space Models*, Springer Verlag, Berlin (1982).
151. E. Walter, editor, *Identifiability of Parametric Models*, Pergamon Press, Oxford (1987).
152. P. M.J. Van den Hof, J. F.M. Van Doren, S. G. Douma, *Identification of parameters in large-scale physical model structures, for the purpose of model-based operations*. In: P.M.J. Van den Hof et al. (Eds.), *Model-Based Control: Bridging Rigorous Theory and Advanced Technology*, Springer-Verlag, New York, USA (2009).
153. S. Santhanagopalan, Q. Guo, R.E. White, *J. Electrochem. Soc.* **154** (2007) A198.
154. V.R. Subramanian, S. Devan, R.E. White, *J. Power Sources* **135** (2004) 361.

155. V.R. Subramanian, R.E. White, *Comput. Chem. Eng.* **24** (2000) 2405.
156. V.R. Subramanian, D. Tapriyal, R.E. White, *Electrochem. Solid-State Lett.* **7** (2004) A259.
157. V. Boovaragavan, C.A. Basha, *J. Appl. Electrochem.* **36** (2006) 745.
158. V. Boovaragavan, C.A. Basha, *J. Power Sources* **158** (2006) 710.
159. Q. Guo, V.R. Subramanian, J.W. Weidner, R.E. White, *J. Electrochem. Soc.* **149** (2002) A307.
160. V.R. Subramanian, V. Boovaragavan, K. Potukuchi, V.D. Diwakar, A. Guduru, *Electrochem. Solid-State Lett.* **10** (2007) A25.
161. V.R. Subramanian, V. Boovaragavan, V.D. Diwakar, *Electrochem. Solid-State Lett.* **10** (2007) A225.
162. V.R. Subramanian, V.D. Diwakar, D. Tapriyal, *J. Electrochem. Soc.* **152** (2005) A2002.
163. V. Boovaragavan, S. Harinipriya, Venkat R. Subramanian *Journal of Power Sources* **183** (2008) 361.
164. D.M. Bernardi, M.W. Verbrugge, *J. Electrochem. Soc.* **139** (1992) 2477.
165. D.M. Bernardi, M.W. Verbrugge, *AIChE J.* **37** (1991) 1151.
166. T.E. Springer, T.A. Zawodzinski, S. Gottesfeld, *J. Electrochem. Soc.* **138** (1991) 2334.
167. A.Briyikoglu, *Int. J. Hydrogen Energ.* **30** (2005) 1181.
168. C.Y. Wang, *Chem. Rev.* **104** (2004) 4727.
169. U. Pasaogullari, C.Y. Wang, *Electrochim. Acta* **49** (2004) 4359.
170. U. Pasaogullari, C.Y. Wang, *J. Electrochem. Soc.* **152** (2005) A380.
171. H. Meng, C.-Y. Wang, *J. Electrochem. Soc.* **152** (2005) A1733.
172. Y. Wang, C.-Y. Wang, *J. Electrochem. Soc.* **153** (2006) A1193.
173. Y.W. Rho, S. Srinivasan, Y.T. Kho, *J. Electrochem. Soc.* **141** (1994) 2089.
174. K. Broka, P. Ekdunge, *J. Appl. Electrochem.* **27** (1997) 281.
175. F. Jaouen, G. Lindbergh, G. Sundholm, *J. Electrochem. Soc.* **149** (2002) A437.
176. N.P. Siegel, M.W. Ellis, D.J. Nelson, M.R. von Spakovsky, *J. Power Sources* **115** (2003) 81.
177. S.-M. Chang, H.-S. Chu, *J. Power Sources* **161** (2006) 1161.
178. G. Lin, T. Van Nguyen, *J. Electrochem. Soc.* **153** (2006) A372.
179. T. Navessin, S. Holdcroft, Q.P. Wang, D.T. Song, Z.S. Liu, M. Eikerling, J. Horsfall, K.V. Lovell, *J. Electroanal. Chem.* **567** (2004) 111.
180. D.T. Song, Q.P. Wang, Z.S. Liu, M. Eikerling, Z. Xie, T. Navessin, S. Holdcroft, *Electrochim. Acta* **50** (2005) 3347.
181. D.T. Song, Q.P. Wang, Z.S. Liu, T. Navessin, M. Eioerling, S. Holdcroft, *J. Power Sources* **126** (2004) 104.
182. D.T. Song, Q.P. Wang, Z.S. Liu, T. Navessin, S. Holdcroft, *Electrochim. Acta* **50** (2004) 731.
183. Q.P. Wang, M. Eikerling, D.T. Song, Z.S. Liu, *J. Electroanal. Chem.* **573** (2004) 61.
184. Q.P. Wang, M. Eikerling, D.T. Song, Z.S. Liu, T. Navessin, Z. Xie, S. Holdcroft, *J. Electrochem. Soc.* **151** (2004) A950.
185. Q.P. Wang, D.T. Song, T. Navessin, S. Holdcroft, Z.S. Liu, *Electrochim. Acta* **50** (2004) 725.

186. P. Mocoteguy, F. Druart, Y. Bultel, S. Besse, A. Rakotondrainibe, *J. Power Sources* **167** (2007) 349.
187. M. Eikerling, A. A. Kornyshev, A. A. Kulikovsky, Physical modeling of fuel cells and their components. In: A.J. Bard, M. Stratmann, D. Macdonald, P. Schmuki, editors. Encyclopedia of electrochemistry, volume 5, electrochemical engineering. Weinheim: VCH-Wiley, 2007: 447.
188. C. Song, Y. Tang, J. L. Zhang, J. Zhang, H. Wang, J. Shena, S. McDermid, J. Li, P. Kozak, *Electrochim. Acta*, **52** (7) (2007) 2552.
189. S. S. Jang, B. V. Merinov, T. Jacob, W. A. Goddard III, in International Conference on Solid State Ionics-15, Abstract P498, International Solid State Ionics (2005).
190. R. Subbaraman, D. Strmcnik, V. Stamenkovic, N. M. Markovic, *J. Phys. Chem. C*, **114** (18) (2010) 8414
191. O. Antoine, Y. Bultel, R. Durand, *J. Electroanal. Chem.* **499** (1) (2001) 85.
192. P. M. Biesheuvel, A. A. Franco, M. Z. Bazant, *J. Electrochem. Soc.* **156** (2) (2009) B225-B233.
193. A. Frumkin, *Z. Phys. Chem. Abt. A* **164** (1933) 121.
194. J. O. Bockris and A. K. N. Reddy, Modern Electrochemistry, Plenum, New York (1970).
195. J. X. Wang, T. E. Springer, R. R. Adzic, *J. Electrochem. Soc.* **153** (9) (2006) A1732.
196. D. Krapf, B. M. Quinn, M.-Y. Wu, H. W. Zandbergen, C. Dekker, S. G. Lemay, *Nano lett.* **6** (11) (2006) 2531.
197. G. A. Camara, E. A. Ticianelli, S. Mukerjee, S. J. Lee, and J. McBreen, *J. Electrochem. Soc.* **149** (6) (2002) A748-A753.
198. R. M. Darling, J. P. Meyers, *J. Electrochem. Soc.* **150** (2003) A1523.
199. S.G. Rinaldo, J. Stumper, M. Eikerling, *J. Phys. Chem. C* **114** (2010) 5773.
200. E. Christoffersen, P. Liu, A. Ruban, H. L. Skriver, J. K. Nørskov, *J. Cat.* **199** (2001) 123.
201. A.A. Franco, M. Tembely, *J. Electrochem. Soc.* **154** (7) (2007) B712.
202. B. Fang et al., Nano-engineered PtVFe catalysts in proton exchange membrane fuel cells: Electrocatalytic performance, doi:10.1016/j.electacta.2010.02.048~
203. J.K. Nørskov, J. Rossmeisl, A. Logadottir, L. Lindqvist, J.R. Kitchin, T. Bligaard, H.J. Jonsson, *Phys. Chem. B* **108** (2004) 17886.
204. J. Greeley, J. Rossmeisl, A. Hellmann, J. K. Nørskov, *Z. Phys. Chem.* **221** (9-10) (2007) 1209.
205. T. Jacob, W. A. Goddard, *Chem. Phys. Chem.* **7** (2006) 992.
206. T. Jacob, *Fuel Cells* **6** (2006) 159.
207. T. Jacob, W. A. Goddard III, *J. Am. Chem. Soc.* **126** (2004) 9360.
208. A. Eichler, F. Mittendorfer, *J. Hafner, Phys. Rev. B* **62** (7) (2000) 4744.
209. A. Eichler, J. Hafner, *Phys. Rev. Lett.* **79** (1997) 4481.
210. S. Kandoi, A.A. Gokhale, L.C. Grabow, J.A. Dumesic, M. Mavrikakis, *Catal. Lett.* **93** (2004) 93.
211. T. Jacob, W. A. Goddard III, *J. Phys. Chem. B* **108** (2004) 8311.
212. Y. Xu, A.V. Ruban, M. Mavrikakis, *J. Am. Chem. Soc.* **126** (2004) 4717.
213. Yuguang Ma, Perla B. Balbuena, *Chem. Phys. Lett.* **447** (2007) 289-294.

214. Yixuan Wang and Perla B. Balbuena, *J. Chem. Theory Comput.* **1** (2005) 935-943.
215. Yixuan Wang and Perla B Balbuena, *J. Phys. Chem. B* **109** (2005) 18902-18906.
216. R. Ferreira de Moraes, D. Loffreda, P. Sautet, A. A. Franco, *Electrochim. Acta*, **56** (28) (2011) 10842.
217. L. Wang, A. Roudgar, M. Eikerling, *J. Phys. Chem. C* **113** (2009) 17989.
218. M. Neurock, *J. Cat.* **216** (1-2) (2003) 73.
219. A. Michaelides, P. Hu, *J. Chem. Phys.* **114** 1 (2001).
220. K. J. Laidler, *Theories of Chemical Reaction Rates*, McGraw-Hill, New York, 1969.
221. K. J. Laidler, *Chemical Kinetics*, 3rd ed, Harper Collins, New York, 1987.
222. B. Temel *et al.*, *J. Chem. Phys.*, **126** (2007) 204711.
223. M. Maestri, K. Reuter, *Angew. Chemie*, **123** (2011) 1226.
224. M.K. Sabbe, M.-F. Reyniers, K. Reuter, *Cat. Sci. Technol.*, **2** (2012) 2010.
225. http://th.fhi-berlin.mpg.de/th/Meetings/DFT-workshop-Berlin2011/presentations/2011-07-21_Reuter_Karsten.pdf
226. A. A. Franco, P. Schott, C. Jallut, B. Maschke, *J. Electrochem. Soc.*, **153** A1053 (2006).
227. D. Damasceno Borges, A.A. Franco, K. Malek, G. Gebel, S. Mossa, "Inhomogeneous transport in model hydrated polymer electrolyte supported ultrathin films", *ACS Nano*, article published ASAP (2013).
228. D. Damasceno Borges, K. Malek, S. Mossa, G. Gebel, A.A. Franco, *ECS Trans.*, **45** (2013) 101.
229. A. A. Franco, M. Quiroga, in preparation (2013).
230. M. D. Williams, D. S. Bethune and A. C. Luntz, *J. Chem. Phys.* **88** (4) (1988).
231. K. Malek, A. A. Franco, *J. Phys. Chem. B* **115**, 8088 (2011).
232. A. A. Franco, S. Passot, P.Fugier, E. Billy, N. Guillet, L. Guetaz, E. De Vito, S. Mailley, *J. Electrochem. Soc.*, **156** (2009) B410.
233. V. Johnson, A. Pesaran, T. Sack, in: *Proceedings of the 17th Electric Vehicle Symposium*, Montreal, Canada, 2000.
234. C. O'Gorman, D. Ingersoll, R. Jungst, T. Paez, in: W.R. Cieslak, *et al.* (Eds.), *Selected Battery Topics*, Vol. PV 98-15, The Electrochemical Society Proceedings Series, Pennington, NJ, 1998, p. 248.
235. J. Newman, *Electrochemical Systems*, 2nd Edition, Prentice-Hall, Englewood Cliffs, NJ, 1973.
236. R. Spotnitz, *Interface*, Winter (2005) 39.
237. C.Y. Wang, V. Srinivasan, *Journal of Power Sources*, **110** (2002) 364.
238. V. Ramadesigan, P. W. C. Northrop, S. De, S. Santhanagopalan, R. D. Braatz, V. R. Subramanian, *J. Electrochem. Soc.*, **159** (3) (2012) R31.
239. J. Cahn, J. Hilliard, *The Journal of Chemical Physics*, **28** (2) (1958) 258.
240. J.E. Guyer, W. J. Boettinger, J.A. Warren, G.B. McFadden, *Phys. Rev. E*, **69** (2004) 021603.
241. J.E. Guyer, W. J. Boettinger, J.A. Warren, G.B. McFadden, *Phys. Rev. E*, **69** (2004) 021604.
242. A.J. Bray, *Adv. Phys.*, **43** (1994) 357.
243. J. Kim, *Communications in Nonlinear Science and Numerical Simulation*, **12** (2007) 1560.

244. S.M. Choo, S.K. Chung, K.I. Kim, *Comput Math Appl.*, **39** (2000) 229.
245. M. Copetti, C.M. Elliott, *Materials Science and Technology* **6** (1990) 273.
246. B. C. Han, A. Van der Ven, D. Morgan, G. Ceder, *Electrochimica Acta*, **49** (2004) 4691.
247. Y. H. Kao, M. Tang, N. Meethong, J. M. Bai, W. C. Carter, Y. M. Chiang, *Chem. Mater.*, **22** (21) (2010) 5845.
248. B. Kang, G. Ceder, *Nature*, **458** (7235) (2009) 190.
249. D. Morgan, A. Van der Ven, G. Ceder, *Electrochem. Solid-State Lett.*, **7** (2) (2004) A30.
250. M. S. Islam,; Driscoll, D. J.; Fisher, C. A. J.; Slater, P. R. *Chem. Mater.* 2005, 17 (20), 5085-5092.
251. N. Dupre, J. Oliveri, J. Degryse, J.F. Martin, D. Guyomard, *Ionics*, **14** (3) (2008) 203.
252. H. J. Tan, J. L. Dodd, B. Fultz, *J. Phys. Chem. C*, **113** (48) (2009) 20527.
253. G. K. Singh, G. Ceder, M.Z. Bazant, *Electrochim. Acta*, **53** (26) (2008) 7599.
254. P. Bai, D. A. Cogswell, M. Z. Bazant, *Nano Letters*, **11** (2011) 4890.
255. D.A. Cogswell, M.Z. Bazant, *ACS Nano*, **6** (3) (2012) 2215.
256. D.A. Cogswell, M. Z. Bazant, Abstract #734, 220th ECS Meeting (2011).
257. R. Khatib, A.-L. Dalverny, M. Saubanère, M. Gaberscek, M.-L. Doublet, *The Journal of Physical Chemistry C*, **117** (2013) 837.
258. A.-L. Dalverny, J.-S. Filhol, M.L. Doublet, *Journal of Materials Chemistry*, **21**, (2011) 10134.
259. G. B. Less, J. H. Seo, S. Han, A. M. Sastry, J. Zausch, A. Latz, S. Schmidt, C. Wieser, D. Kehrwald, S. Fell, *J. Electrochem. Soc.*, **159** (6) (2012) A697.
260. S. Dargaville, T. W. Farrell, *J. Electrochem. Soc.*, **157** (2010) A830.
261. K. C. Smith, P. P. Mukherjee, T. S. Fisher, Abstract #748, 220th ECS Meeting (2011).
262. K.C. Smith, P.P. Mukherjee, T.S. Fisher, *Phys. Chem. Chem. Phys.*, **14** (2012) 7040.
263. W. Du, N. Xue, A. M. Sastry, J.R.R.A. Martins, W. Shyy, Abstract #883, Honolulu, ECS Meeting PRiME 2012.
264. G. M. Goldin, A. M. Colclasure, A. H. Wiedemann, R. J. Kee, *Electrochimica Acta*, **64** (2012) 118.
265. J. Song, M.Z. Bazant, *J. Electrochem. Soc.*, **160** (1) (2013) A15.
266. R. Thiedmann, O. Stenzel, A. Spetl, P. R. Shearing, S.J. Harris, N. P. Brandon, V. Schmidt, *Computational Materials Science*, **50** (2011) 3365.
267. B. Vijaraghavan, D.-W. Chung, P. Shearingz, N. Brandon, S. Harris, R. E. García Abstract #347, 221st ECS Meeting (2012).
268. M. Ender, J. Illig, E. Ivers-Tiffée, Abstract #1069, Honolulu, ECS Meeting PRiME 2012.
269. M. Ender, J. Joos, T. Carraro and E. Ivers-Tiffée, *J. Electrochem. Soc.*, in press (2012).
270. M. Ender, J. Joos, T. Carraro, E. Ivers-Tiffée, *Electrochem. Comm.*, **13** (2) (2011) 166.
271. B. Orvananos, H.-C. Yu, M. Bazant, K. Thornton, Abstract #750, 220th ECS Meeting (2011).

272. K. Malek, T. Mashio, Atomistic and Molecular Modeling of Degradation Mechanisms in PEMFCs, book Chapter in: Polymer Electrolyte Fuel Cells: Science, Applications and Challenges, A.A. Franco Ed., Pan Stanford, Singapore (2013).
273. K. Malek, M. Eikerling, Q. Wang, T. Navessin, Z. Liu, *J. Phys. Chem. C*, **111** (2007) 13627.
274. K. Malek and A. A. Franco, *J. Phys. Chem. B* **115**, 8088 (2011).
275. K. Malek, T. Mashio, and M. Eikerling, *Electrocatalysis* **2**, 141 (2011).
276. A.A. Franco, M. Gerard, *J. Electrochem. Soc.*, **155** (4) (2008) B367.
277. M. Combes, B. Buin, M. Parenthoën, J. Tisseau, *Procedia Computer Science*, **1** (2012) 761.
278. A. A. Franco, W.G. Bessler, M.L. Doublet, Eds., Multiscale modeling and numerical simulation of electrochemical devices for energy conversion and storage, Springer (2013) in preparation.
279. A.A. Franco Ed., Rechargeable lithium batteries: from fundamentals to applications, Woodhead (UK) (2013), in preparation.
280. A. A. Franco, Modeling of Direct Alcohol Fuel Cells, book chapter in: Direct Alcohol Fuel Cells Technologies: research and development (publisher: Springer), edited by E. Gonzalez and H. Corti, in press (2013).

Multiscale Transport Methods for Exploring Nanomaterials and Nanodevices

Frank Ortmann^{1,2} and Stephan Roche^{2,3}

¹ Institute for Materials Science and Max Bergmann Center of Biomaterials, Technische Universität Dresden, 01062 Dresden, Germany

² ICN2 - Institut Català de Nanociència i Nanotecnologia, Campus UAB, 08193 Bellaterra (Barcelona), Spain
E-mail: {frank.ortmann,stephan.roche}@icn.cat

³ ICREA, Institució Catalana de Recerca i Estudis Avançats, 08070 Barcelona, Spain

Novel materials or compounds are key for future technological innovations and hold great potential for new applications for which the functionality of such materials has to be assessed in experimental devices. Simulations can bring deep insight in fundamental properties but should include both realistic sample sizes and all relevant interactions. However, including both at the same time is a formidable task. In this paper, we address concepts and methods used to describe charge transport phenomena in condensed matter by multiscale methods including interaction of electrons with other electrons or phonons but also with disorder, as present in experiment. We illustrate the interface between atomistic properties, which are determined from first principles, and large-scale transport simulations. Recent results from efficient order-N electronic transport simulations serve as examples to discuss the influence of disorder on transport coefficients including electron and polaron transport, which are observed experimentally in inorganic and/or organic materials.

1 Introduction: State of the Art of Computational Approaches for Nanodevice Simulation

Novel materials are a strong driving force for economic power as was steel and plastics in the last centuries or as evidenced in the notion of Bronze age in archaeology. In human history, the knowledge about processing of such materials is an essential strategic advantage, while nowadays we have means to study also their theoretical properties which is equally important and complementary for designing or researching new ones.¹

Novel technologies emerge often from new functional materials that become manageable in praxis such as in novel smart mobile devices which rely heavily on battery technology, high quality displays, and fast and energy saving chips. This was also emphasized by the European Commission stating that technology developments are largely driven by such advancements: “Alternative paths to components and systems development - including nanoelectronics, more integration of functionalities on chips, the use of new materials and progress in photonics - will drive a large part of technology developments.”

Many fundamental questions are open in the field of nanoelectronics and new materials, which, due to the complexity of both, cannot be answered by conventional simplified approaches. The research of novel functional materials is therefore highly interdisciplinary covering the domains of chemistry, material science, physics, and engineering with their methods and scope of length scales. Advanced knowledge of such fields has necessarily to be combined. In addition, the complexity of quantum laws in nanoelectronics complicates

upscaling attempts such that, at the cross-road of new materials and nanoelectronics (especially for *beyond-CMOS* applications), only multiscale modeling approaches can take into account the mutual interaction between structure and materials and can advance knowledge sufficiently fast in the near future. In particular the understanding of charge transport is a central goal in semiconductor research, a field which has been the basis of the increasing reasoning power of mankind during the last 50 years and continues to be thanks to the continuous miniaturization of microelectronics accompanied by the increasing speed and power of computing devices. Although semiconductor electronics is the technological basis for the “information era” it becomes clear that many kinds of electronic devices beyond those for information processing and flow are demanded in all kinds of applications. These span over a wide range from sensing, monitoring and controlling to lighting and energy harvesting.

1.1 Challenges below 1 nm

Improving such devices means often improving the materials in terms of their specific functionality which in itself breaks down to the chemistry of the materials. Such chemistry is governed by the laws of quantum mechanics which include all kinds of complicated physics that arises from the interaction of the electrons, their nature as indistinguishable particles, correlations and quantum fluctuations which occur on a variety of length and energy scales (from weak van der Waals bonding to strong exchange interactions). The smallest characteristic length scale for this is on the order of 0.01-0.1 nm. However not all of these interactions are relevant on larger length scales and sometimes it is not known which one ‘survives’ at a typical device scale of 0.1-1 μm .

Important approaches to bridge the smallest length scales up to 1 nm by using a proper description of the quantum laws are the wide class of *ab initio* methods. In particular density functional theory (DFT) allows to describe small quantum systems of up to a few hundreds of atoms by mapping the many-body problem to an effective single particle picture. This is made possible by the functional description of interactions and such functionals exist nowadays in many flavors (see Sect. 4 for more details).

Similarly a manifold of codes exist, such as AB-INIT,² SIESTA,³ or VASP,⁴ to mention a few, each of which is adapted to the specific needs of their users. Such simulation software provides an atomistic viewpoint by simulating the atomic structure from monomers to larger clusters of atoms and molecules with the tendency that larger systems are treated with lower accuracy (which is commonly accepted practice). *Ab initio* methods were very successful in recent years for the description of bulk materials, nanostructures but also molecules and are therefore well established. Often, due to the ease of use, experimental groups use them to support their measurements by complementing simulations. In recent years the versatility of DFT implementations has been extended towards computational spectroscopic tools to broaden their application spectrum in the field of transport or optoelectronic properties of materials and structures.

Unfortunately, the computational cost of *ab initio* methods is still very expensive. Although they run on parallel-computing architectures there is a practical limit of few hundreds of atoms. This originates from the complexity of electrons represented by wavefunctions and, hence, possessing an inner structure with widely variable properties compared to simple particles used in classical molecular dynamics. The requirement of self-consistency

is only one consequence of the quantum nature which slows down such methodology. Although many concepts exist to weaken this impact, there is a practical limitation to sizes of systems, at present of about 1 nm, which are treatable by *ab initio* methods.

1.2 Challenges above 1 nm

Despite the success of DFT methods below 1nm they cannot reach the length scale on which one discusses functional materials (which is at least one-two orders of magnitude above). Indeed, the characteristics and fundamental properties of interest of functional materials are co-defined on a larger length scale beyond the *ab initio* scope. This is because they are additionally influenced by other properties emerging on larger length scales, which are governed by low concentration dopants, impurities or structural defects for instance, or simply because the relevant structures may reach these dimensions themselves. The value of 1 nm seems therefore a critical length and an upper bound for sophisticated modeling of quantum laws.

Multiscale modeling is capable of overcoming this barrier. The concept is based on the observation that not all interactions must necessarily be treated within the first-principles framework. This observation allows one to introduce a hierarchy of interactions, which might be based either on very general considerations or just adapted and valid for the presently studied properties. Based on this, a hierarchy of levels of treatment may be introduced. The lowest (microscopic) level deals with the smallest objects at the highest accuracy. It can be identified with the full *ab initio* level. Multiscale modeling defines first the *models* on each level and second the *interfaces* for transferring relevant information to the respective upper level (or even lower level for feedback loop) where they are further processed. The advantage is that not all information available on the (computationally heavy) lower level enters the upper-level modeling. The exchange across the interface is restricted to relevant information which is precisely where multiscale modeling is benefiting from. In addition, the modeling of interactions on the upper macroscopic level replaces respective couplings on the more refined lower level. This allows one to reduce the work at the lower level by treating smaller parts (non-interacting subsystems) there.

For instance a finite-range impact on electrostatics and on electronic properties can be expected from impurities or dopants depending on the local surrounding of a host crystal. Additional long-range parts such as arising from the Coulomb interaction might be separable and can be treated on the upper level. The information on local electronic properties can be obtained with *ab initio* methods using large supercells. On the other hand, the evolution of a system as a whole composed of millions of atoms including a certain distribution of such dopants is unpredictable by *ab initio* methods when the entire system is included at the same theory level. Taking advantage of the above observation and of a simplified but yet realistic modeling on the upper level describing a macroscopic part of the system, a solution can be found by treating only a subsystem fully *ab initio* to extract relevant information. The information together with separated long-range interactions completes the upper-level model to treat the full system.

In case of transport properties of novel materials, interconnections of such subsystems are explored by traveling quantum particles. Consequently the interaction between these derives from a complicated way of electron motion itself which is implied by the quantum nature of the charge carriers sometimes evading ones intuition. Consequently working out the numerics is necessary.

To design a realistic scenario of the influence of disorder on transport properties of materials and devices one has to consider different length scales and modeling strategies simultaneously. First, a microscopic picture of the atomic structure is necessary to access electronic properties. This can only be provided with state-of-the-art first principles simulations. These calculations can be carried out using simple unit cells in the case of clean systems. When considering crystal imperfections or dopants, larger supercells with few impurity atoms or defect sites are necessary to simulate.

The interface to the macroscopic level is an essential ingredient of the modeling. It defines which information is exchanged, i.e. which features of the *ab initio*-simulation part are relevant enough to be important on the macroscopic length scale. From these simulations one extracts electronic structure parameters which represent at best the interactions at this level. For the efficiency of the multiscale approach it is very advantageous if the extracted parameters are generic and transferable. This should be considered when setting up the modeling strategy to reduce or, at best, avoid feedback effects.

Once the macroscopic model is well defined and its parameters are provided through the interface, a variety of situations can be investigated keeping the interface parameter fixed but changing the arrangement or interconnection of such subsystems and/or environmental conditions. We see that flexibility is one of the central advantages of multiscale approaches. This will help to gain much more knowledge for complex systems which at present is otherwise not accessible.

Finally, the material properties with all their dependencies on external parameters can be used to define another superior level of simulation such as the simulation of a whole device. On such a level both specific device characteristics such as geometry etc. and material properties (intrinsic or specifically tailored) enter the final results. In these lecture notes we demonstrate some examples from recent research that show which questions can be addressed by multiscale approaches. In Sect. 2 and 3 we start by displaying the general transport frameworks used for the simulations on the macroscopic level, i.e. the Kubo and Landauer formalisms. Some computational details are included for illustration. We then introduce the DFT framework in Sect. 4. In Sects. 5, 6, and 7 we present the results for selected examples illustrating the power of multiscale approaches. Finally we conclude in Sect. 8.

2 Kubo-Transport Methodology

2.1 General Description of Kubo's Approach

Charge-carrier transport can be described within the Kubo transport framework,⁵ where the Kubo formula relates the carrier conductivity to the current-current correlation function. This formula is a result of an expansion of the response of a system to the perturbation (applied electric field) up to linear terms in the field, for which the term *Linear Response Theory* has been coined.⁶ The resulting current is given by the expectation value with the current operator \hat{j}_α ,

$$J_\alpha = \text{Tr}[\hat{\rho}(t)\hat{j}_\alpha], \quad (1)$$

where $\hat{\rho}(t)$ is the density matrix of the system including the electric field at time t and $\text{Tr}[\dots]$ means the usual trace operation. As a result of the linear expansion one can finally

write for the dc conductivity

$$\sigma_{\alpha\beta} = \frac{1}{V} \int_0^\infty dt \int_0^{1/k_B T} d\lambda \text{Tr}[\hat{\rho}(0) \hat{j}_\beta \hat{j}_\alpha(t + i\hbar\lambda)] \quad (2)$$

for finite temperature T . The indices α, β in Eq. (2) denote the Cartesian components of the tensor $\sigma_{\alpha\beta}$ and k_B is the Boltzmann constant. For the specific cases we are considering here, we can further assume a simplified form for the diagonal conductivity ($\alpha = \beta$)

$$\sigma_{\text{dc}} = \sigma_{\alpha\alpha} = \frac{1}{2k_B T \Omega} \int_{-\infty}^\infty dt \text{Tr}[\hat{\rho}(0) \hat{j}_\alpha \hat{j}_\alpha(t)] \quad (3)$$

where Ω is the system volume. By introducing position operators \hat{x}_α and $\hat{x}_\alpha(t) = \hat{U}^\dagger(t) \hat{x}_\alpha \hat{U}(t)$ [with $\hat{U}(t)$ the time evolution operator] we can write with $\Delta\hat{X}(t) = [\hat{x}_\alpha(t) - \hat{x}_\alpha(0)]$

$$\sigma_{\text{dc}} = \frac{e_0^2}{2k_B T \Omega} \lim_{t \rightarrow \infty} \frac{d}{dt} \text{Tr}[\hat{\rho}(0) \Delta\hat{X}^2(t)] \quad (4)$$

which relates the conductivity to the time dependent spread of wave functions.

At $T=0$ this corresponds to the standard result of the Kubo-Greenwood approach

$$\sigma_{\text{dc}}(E) = \frac{e_0^2}{2} \lim_{t \rightarrow \infty} \frac{d}{dt} \Delta X^2(E, t) \quad (5)$$

where

$$\Delta X^2(E, t) = \text{Tr}[\delta(E - \hat{H}) \Delta\hat{X}^2(t)] \quad (6)$$

and $\delta(E - \hat{H})$ the Dirac delta distribution.

2.2 Computational approaches

To explore carrier transport in disordered systems we use an efficient implementation based on a real-space computational approach to calculate the Kubo-Greenwood conductivity (for details see Refs. 7–10). The efficiency is witnessed in the linear (order N) scaling with system size N which allows to explore the relevant length scales even in 3D for standard sample sizes that contain tens of millions of atoms reaching the micron scale. This method solves the time-dependent Schrödinger equation and computes the diffusion coefficient

$$D(E_F, t) = \frac{1}{\rho(E_F)} \frac{d}{dt} \Delta X^2(E_F, t) \quad (7)$$

(with $\rho(E) = \text{Tr}[\delta(E - \hat{H})]$ the density of states) and the Kubo conductivity (Eq. (5)). Thereby one uses an expansion of $\hat{U}(t)$ in a basis of Chebyshev polynomials and Lanczos' recursion procedure.

The basic Chebyshev expansion of the time-evolution operator reads

$$\hat{U}(\Delta t) \equiv e^{-\frac{i\hat{H}\Delta t}{\hbar}} = \sum_{n=0}^{\infty} c_n(\Delta t) P_n(\hat{H}) \quad (8)$$

which allows for an efficient time propagation even for large systems with 10^8 atoms. This is because the expansion can be truncated rapidly at finite n for energy spectra of \hat{H} with

finite support $[a-b, a+b]$ since the expansion coefficients $c_n(\Delta t)$ decay according to the Bessel functions $J_n(\frac{-b\Delta t}{\hbar})$.¹¹

Another important approach in the numerical evaluation of the conductivity at a given energy E_F is based on the Lanczos method. This method is used to calculate the trace in Eq. (6) and in the density of states. It is based on the replacement $\text{Tr}[\dots] \rightarrow N \langle \Psi | \dots | \Psi \rangle$ with random-phase wave packets $|\Psi\rangle$. The method consists of a recursive way of calculating the traces according to the following recipe that starts by tridiagonalizing the system Hamiltonian \hat{H} . For the first recursion step we take a starting vector of the recursion $|\Psi_1\rangle$ and calculate

$$a_1 = \langle \Psi_1 | \hat{H} | \Psi_1 \rangle \quad (9)$$

$$|\Psi'_2\rangle = \hat{H}|\Psi_1\rangle - a_1|\Psi_1\rangle \quad (10)$$

$$b_1 = \sqrt{\langle \Psi'_2 | \Psi'_2 \rangle} \quad (11)$$

$$|\Psi_2\rangle = \frac{1}{b_1} |\Psi'_2\rangle \quad (12)$$

and for all following steps ($n > 1$) we use the relations

$$a_n = \langle \Psi_n | \hat{H} | \Psi_n \rangle \quad (13)$$

$$|\Psi'_{n+1}\rangle = \hat{H}|\Psi_n\rangle - a_n|\Psi_n\rangle - b_{n-1}|\Psi_{n-1}\rangle \quad (14)$$

$$b_n = \sqrt{\langle \Psi'_{n+1} | \Psi'_{n+1} \rangle} \quad (15)$$

$$|\Psi_{n+1}\rangle = \frac{1}{b_n} |\Psi'_{n+1}\rangle. \quad (16)$$

The obtained recursion coefficients a_i and b_i are the matrix elements of the tridiagonal Hamiltonian (\hat{H}') in the Lanczos basis where the initial state $|\Psi_1\rangle$ can be conveniently chosen as a random-phase state.

This representation allows for a simple evaluation of the spectral quantity $\rho_{\Psi_1}(E) = \langle \Psi_1 | \delta(E - \hat{H}) | \Psi_1 \rangle$ (density of states) which occurs frequently in the calculation of transport coefficients. It can be evaluated in a simple algebraic way by means of a continued fraction expansion

$$\langle \Psi_1 | \delta(E - \hat{H}) | \Psi_1 \rangle = - \lim_{\eta \rightarrow 0} \frac{1}{\pi} \text{Im} \left[\langle \Psi'_1 | \frac{1}{E + i\eta - \hat{H}'} | \Psi'_1 \rangle \right] \quad (17)$$

$$\langle \Psi'_1 | \frac{1}{E + i\eta - \hat{H}'} | \Psi'_1 \rangle = \frac{1}{E + i\eta - a_1 - \frac{b_1^2}{E + i\eta - a_2 - \frac{b_2^2}{E + i\eta - a_3 - \frac{b_3^2}{\ddots}}}}. \quad (18)$$

The implementation and evaluation of Eq. (17) is straight forward once the matrix elements a_i and b_i have been determined. The artificial broadening parameter η is introduced to ensure convergence and which should be taken as small as possible while standard terminations for the expansion can be used.

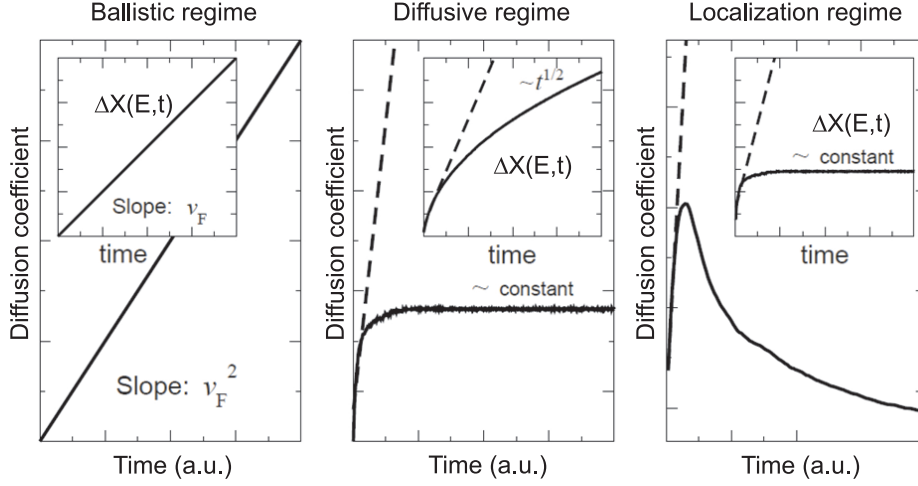


Figure 1. Diffusion coefficients (main frames) and mean square displacement (insets) for different disorder strengths: no disorder (left), medium disorder (middle) and strong disorder (right). The figure is a courtesy of Dinh Van Tuan.

2.3 Semiclassical Conductivity and Mean Free Path

To illustrate the generic behavior that can occur for the conductivity in Eq. (5) we plot in Figure 1 different transport regimes. In the absence of disorder, wave packets expand unlimited and ballistically and the diffusion coefficient D does not converge to a constant but increases linearly. For sufficient disorder one observes a departure from a linear increase of D . In case the disorder is strong enough $D(t)$ exhibits a maximum value D_{\max} and the corresponding regime can be identified with the semiclassical diffusion and constant diffusion coefficient. Correspondingly one defines a semi-classical conductivity $\sigma_{\text{sc}} = \frac{D_{\max}}{\rho}$ and the mean free path $\ell_e = \frac{D_{\max}}{2v_F}$.

For even stronger disorder the diffusion coefficient eventually starts to decay beyond its maximum. This regime is the localization regime where one distinguishes different classes such as weak localization or strong localization, the latter is also known as Anderson localization. Figure 1 (c) shows that the spread ΔX seemingly assumes a constant value which is related to the localization length. For strong disorder still one can define the maximum of $D(t)$ and semiclassical quantities such as ℓ_e but their interpretation becomes difficult if the mean free path is close to a lattice constant.

In real systems the discussed transport regimes can switch from one energy to the other. In this context, the case of a transition from diffusive to a localization regime is known as mobility edge.

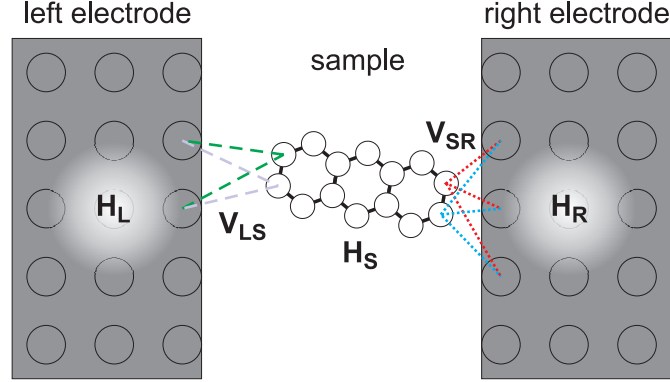


Figure 2. Setup for Landauer transport simulations. A molecule is sandwiched between the contacts (leads).

3 Landauer Transport Approach

While the Kubo method which has been introduced in the above section is particularly well suited to calculate the properties of large systems including disorder or any structural defect or structural features and finite temperature (as we demonstrate in Sect. 7.2), the Landauer-type of approaches considered in this section find frequent applications in quantum-transport studies on nanostructures such as one-dimensional transport geometries, junctions etc. at zero temperature by assuming that electron transport is fully coherent, i.e. no phase-breaking phenomena occur on the transport time scale. Landauer's approach has the advantage to include the crucial role of the contacts in geometries such as depicted in Figure 2 while the contacted sample has typically small length (but often also small widths). Sometimes it is simply a single atom or molecule.¹²

This division into subsystems is reflected in the blockwise definition of the Hamiltonian

$$\hat{H} = \begin{pmatrix} \hat{H}_L & \hat{V}_{LS} & 0 \\ \hat{V}_{LS}^\dagger & \hat{H}_S & \hat{V}_{SR} \\ 0 & \hat{V}_{SR}^\dagger & \hat{H}_R \end{pmatrix} \quad (19)$$

where $\hat{H}_L, \hat{H}_S, \hat{H}_R$ are the (partial) Hamiltonians of the left (L) electrode, the sample (central part), and the right (R) electrode, respectively. The off-diagonal terms \hat{V} describe the coupling between these subsystems and are related to overlapping wave functions in the contact region.

The corresponding matrix Green function fulfills the equation

$$\begin{pmatrix} \hat{E} - \hat{H}_L & -\hat{V}_{LS} & 0 \\ -\hat{V}_{LS}^\dagger & \hat{E} - \hat{H}_S & -\hat{V}_{SR} \\ 0 & -\hat{V}_{SR}^\dagger & \hat{E} - \hat{H}_R \end{pmatrix} \begin{pmatrix} \hat{G}_L & \hat{G}_{LS} & 0 \\ \hat{G}_{SL} & \hat{G}_S & \hat{G}_{SR} \\ 0 & \hat{G}_{RS} & \hat{G}_R \end{pmatrix} = \hat{\mathbb{1}} \quad (20)$$

which is just a system of linear equations and $\hat{E} = \hat{\mathbb{1}}(E + i\eta)$ is a shorthand notation with unity matrix $\hat{\mathbb{1}}$. One is mainly interested in the sample Green function \hat{G}_S which

propagates electronic states in the sample. By substitution we can arrive at

$$\left(\hat{E} - \hat{H}_S - \hat{\Sigma}\right) \hat{G}_S = \hat{\mathbb{1}} \quad (21)$$

where we introduce the so-called self energy

$$\hat{\Sigma} \equiv \hat{\Sigma}_L + \hat{\Sigma}_R = \hat{V}_{LS}^\dagger (\hat{E} - \hat{H}_L)^{-1} \hat{V}_{LS} + \hat{V}_{RS}^\dagger (\hat{E} - \hat{H}_R)^{-1} \hat{V}_{RS}. \quad (22)$$

As we see from Eq. (22) the self energy is composed of lead quantities and lead-sample couplings only. It can be understood as a correction term to the ordinary energy term contained in \hat{E} caused by the coupling to the leads (given by the quantities \hat{V}). We easily see that if $\hat{V} \rightarrow 0$ the self-energy vanishes and \hat{G}_S solves the free equation

$$\left(\hat{E} - \hat{H}_S\right) \hat{G}_S = \hat{\mathbb{1}}. \quad (23)$$

It is an important observation that in the general case (of $\hat{\Sigma} \neq 0$) this propagation of electronic states depends on the electrodes. This is different to the Kubo transport methodology.

While the methods presented briefly in this section easily fill books,¹² we focus only on central aspects related to carrier transport. In geometries such as depicted in Figure 2 one usually investigates the conductance of a sample (in contrast to the average conductivity σ). The conductance \mathcal{G} can be written as a trace over corresponding operators¹³

$$\mathcal{G} = \frac{2e^2}{h} \text{Tr} \left[\hat{\Gamma}_L \hat{G}_S \hat{\Gamma}_R \hat{G}_S^\dagger \right] \quad (24)$$

where

$$\hat{\Gamma}_{L,R} = i(\hat{\Sigma}_{L,R} - \hat{\Sigma}_{L,R}^\dagger). \quad (25)$$

To solve the transport problem in the Landauer framework one starts with the leads evaluating Eq. (22) and proceeds towards Eq. (24). Given a tight-binding representation of the system these equations become simple matrix equations.

4 Ab initio Methods for Material Parameters

4.1 Hohenberg-Kohn Theory

Here we briefly display the methods related to *ab initio* part of the simulations as discussed in Sect. 1.1. In solid-state physics one uses density functional theory (DFT) as the established method to describe electronic properties of solids and their surfaces. Thereby the electron density $n(x)$ plays a central role¹⁴ beyond being merely an expectation value of the ground state $|g\rangle$. Formally it is used as the basic variable of the problem. As such it should be connected one-to-one to the external potential $V(x)$ which is indeed the case for so-called V-representable densities (at least apart from an unimportant constant).¹⁵ It follows that the total ground state energy E_g of the system is a functional of the density of the system $E[n(x)]$. A second theorem states that the functional is extremal¹⁴ leading to a minimal ground state energy.

However this functional $E[n(x)]$ is very difficult to obtain and no general solution is found apart from the special case of the homogeneous electron gas where $n(x)$ is constant

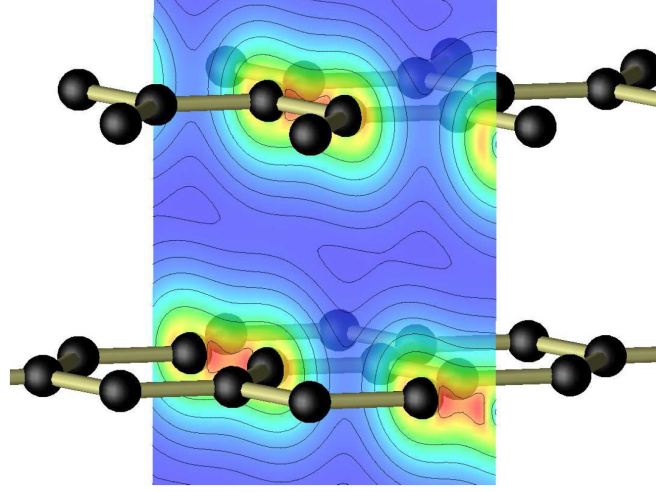


Figure 3. Valence charge density distribution in graphite after Ref. 16. The strong inhomogeneity between graphene layers is evident by iso-lines that indicate doubled values for the charge density from one to the next line.

throughout the system. As a matter of fact it turns out to be a challenge to describe inhomogeneous systems such as the one in Figure 3. Usually the density along a chemical bond is orders of magnitude above the density in between particularly for van der Waals-bonded systems such as graphite or organic solids.

4.2 Kohn-Sham Equations

Notwithstanding these complications, after clarification of its existence and minimum properties one has to determine the electron density $n(x)$ for a given external potential $V(x)$. This amounts to solving an equation like $\hat{H}|\psi\rangle = (\hat{T} + \hat{U} + \hat{V})|\psi\rangle$ for the many-body wave function $|\psi\rangle$ while knowing the electron-electron interaction \hat{U} (\hat{T} is the kinetic energy operator). Kohn and Sham proposed a simplification based on a single particle-picture by introducing an effective potential V_{eff} acting on the non-interaction particles^{17,18}

$$\hat{H}|\varphi_i\rangle = \left(\frac{-\Delta}{2} + \hat{V}_{eff} \right) |\varphi_i\rangle = \varepsilon_i |\varphi_i\rangle, \quad (26)$$

which is known as Kohn-Sham equation and where the first term is the kinetic energy. The effective potential consists of $V_{eff} = V(x) + V_H(x) + V_{XC}(x)$ with the classical Hartree potential V_H , the external potential $V(x)$ and the exchange-correlation potential $V_{XC}(x)$ which is supposed to include all remaining many-particle effects and is undoubtedly the complicated part. Given the knowledge of this potential, Eq. (26) has to be solved under the constraints that $n(x) = \sum_i^N |\varphi_i(x)|^2$ and $N = \int d^3x n(x)$ where N is the number of particles in the system.

Clearly there is a huge amount of literature how to describe the exchange-correlation potential in a suitable yet efficient way. The representation of this research, however, goes far beyond the present purpose. We only mention here that standard approximations exist which are well assessed including the local density approximation (LDA) and the generalized gradient approximation (GGA). In particular for the latter many different flavors exist.

4.3 Electronic Structure

One central goal when performing DFT simulations is to obtain material parameters. This can include the total energy E that we mentioned above or vibrational frequencies etc. But also the electronic structure is very important. Fortunately the eigenenergies ε_i in Eq. (26) can be interpreted as electronic energies characterizing the band structure. While this is merely an empirical finding and counterexamples exhibit problems with this interpretation, we will adopt this wide-spread interpretation to calculate the electronic structure in order to extract transport-relevant quantities.

5 New Electronics Features of Chemically-Modified Graphene-Based Materials: Mobility Gaps

5.1 Introduction

An illustration of the general multiscale approach is the exploration of new type of device principles, based on the concept of mobility gaps, and based on chemical doping of graphene-based materials and devices. We provide here below a more detailed discussion about that phenomenon, being an interesting example of emerging device functionalities from quantum transport effects.

Undoped single layer graphene behaves as a zero-gap semiconductor, and thus it turns out to be an unsuitable material for achieving efficient field-effect functionality in logic circuits. Indeed, experimental measurements reported ratios between the current in the ON state and the current in the OFF state not higher than one order of magnitude and therefore too low to meet technical requirements. A possibility to increase the (zero) gap of two-dimensional graphene single layers is to shrink their lateral dimensions. Using e-beam lithographic techniques and oxygen plasma etching, graphene nanoribbons can be fabricated with ribbon widths of a few tens of nanometers down to say 10nm. This confinement effects trigger electronic bandgaps^{19,20} with a decreasing gap magnitude with increasing nanoribbon width. However, theoretical predictions and experimental results have reported energy bandgaps far too small or very unstable in regards to edge reconstruction and defects, thus preventing to envision outperforming ultimate CMOS-FETs (Complementary Metal Oxide Semiconductor Field-Effect Transistors) with graphene-based devices.

To circumvent such an effect, one should instead recourse to larger width graphene nanoribbons (above 10 nm in lateral sizes) and one should compensate the loss of gain due to the bandgap shrinking by triggering the mobility gaps through chemical doping (such as substitutional boron or nitrogen). These mobility gaps are unique consequences of a wide distribution of quasi-bound states over the entire valence band (for acceptor-type impurities) in the first conductance plateau when dopants are randomly distributed across

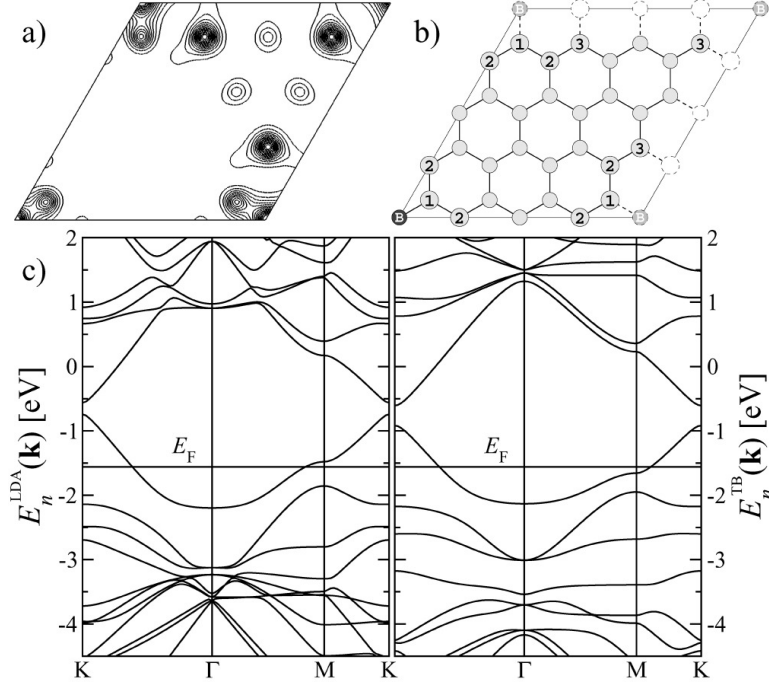


Figure 4. Boron doped graphene. (a) Charge density of the band crossed by the Fermi energy at the Γ point. (b) Structure indicating the Boron substitution and labeling of neighboring sites. (c) Comparison between DFT band structure (left) and tight-binding model (right) shows good agreement. The renormalized on-site energy for the boron site is extracted and shifted upwards by 4.3 eV compared to carbon sites.²¹ Figures are reproduced after Ref. 21.

the ribbon width, due to the strong dependence of the scattering potential on the dopant position with respect to the ribbon edges.

5.2 Material parameters

In Figure 4 we show a DFT calculation in supercell geometry which is used to extract effective microscopic parameters of boron doped graphene. These parameters have been used to feed the transport simulations as explained below. Thereby the nearest neighbor electronic coupling term is set constantly to $\gamma_0 = -2.7$ eV. The right panel of Figure 4 shows the band structure described by the tight-binding model based on the extracted parameters. The agreement around the Fermi energy is excellent. We next create large samples with random distributions of dopants to simulate a realistic situation and describe the physics arising on the macroscopic scale.

5.3 Results

Figure 5 shows the conductance (computed with the Landauer-Büttiker method of Sect. 3) of a 10 nm wide armchair nanoribbon with low-concentration boron doping.^{22,23} For

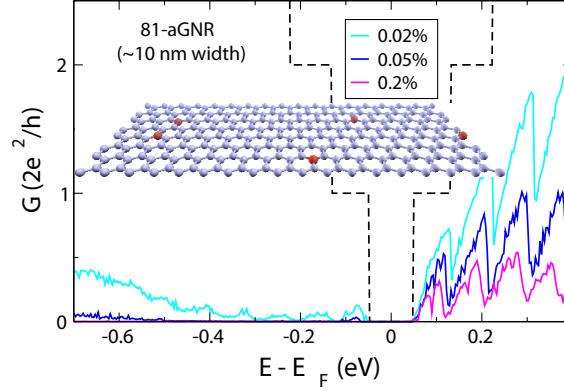


Figure 5. Main panel: average conductance as a function of energy for the semiconducting 81-armchair graphene nano-ribbon (aGNR) and three selected doping rates (0.02%, 0.05% and 0.2%, from top to bottom). Inset: schematic plot of a randomly doped 34-aGNR.

a doping density of about 0.2%, the system presents a mobility gap of the order of 1 eV. When lowering the doping level to 0.05%, the mobility gap reduces to about 0.5 eV and finally becomes less than 0.1 eV for lower density.

The final values for mobility gaps depend on the nanoribbon width and length, so that adjustment can be performed by upscaling either lateral or longitudinal sizes to achieve desired ON/OFF current characteristics, but the recipe is straightforward once the transport length scales (mean free paths, localization length) have been computed.

One notes however that the existence of mobility gaps (with conductance several orders magnitude lower than the quantum conductance) cannot yield a straightforward quantitative estimation of resulting ON/OFF current ratio, since this will require computing the charge flow in a self-consistent manner (using a Schrödinger-Poisson solver). This is essential since accumulated charges inside the ribbon channel are further screening the impurity potential, altering the final strength of mobility gaps obtained in equilibrium conditions. Some efforts have been made in that direction,²⁴ but this needs definitely further specific consideration and stand as an important challenge of multiscale modeling in the ICT domain.

6 Limits of Ballistic Transport in Silicon Nanowires

6.1 Introduction

Semiconducting nanowires with diameter down to the nanometer scale can also be fabricated by catalytic growth techniques. These Bottom-up nanostructures have become the subject of intense study and are considered as potential building blocks for nanoscale electronics due to their promising electronic and optical properties. Compared to classical planar technology, Silicon-based semiconducting nanowires (SiNWs) are able to better accommodate “all-around” gates, which improves field effect efficiency and device perfor-

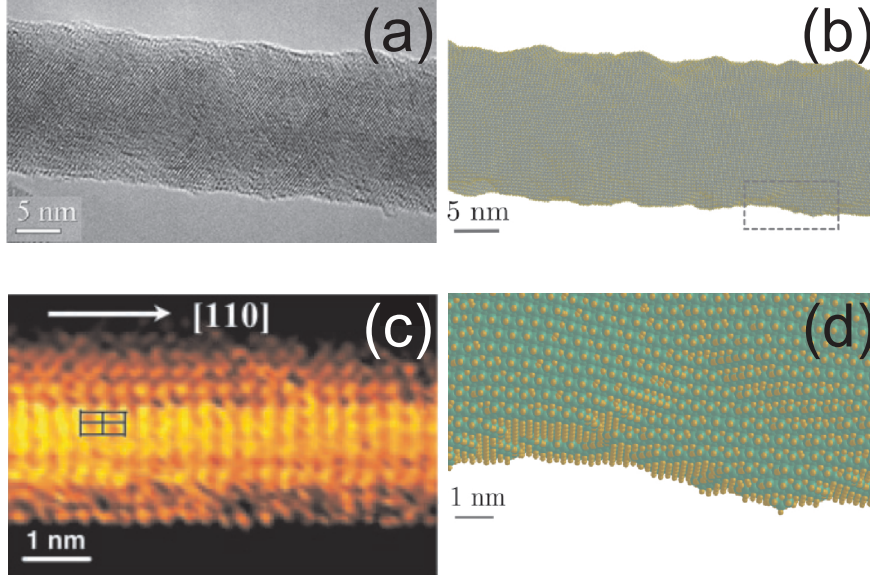


Figure 6. Illustrations of surface roughness in large (a) and small diameter (c) silicon nanowire, together with description of roughness profiles at the atomistic scale (b) and (d). Reproduced from Ref. 25.

mance. Also, in contrast to many other nanowire materials, structurally stable and electrically active SiNWs can be manufactured with small diameters $d < 5$ nm.

However, as the lateral size of the nanowires becomes smaller the impact of structural imperfections such as surface disorder and defects becomes increasingly important due to the high surface to volume ratio. In the case of lithographic SiNW-FETs surface roughness disorder (SRD) is known to be a limiting factor. Moreover, due to the indirect band gap of silicon, SiNWs can be expected to exhibit fundamentally different electronic properties depending on the nanowire crystal orientation. For engineering performant SiNW-based transistors it is thus imperative to find out how sensitive the transport properties are to SRD and which nanowire orientation is best suited for engineering highly performant transistors.

The understanding of charge transport in silicon nanowires demands for an extensive use of atomistic models (*ab initio* or tight-binding models). Indeed, in situations of strong geometrical and electrical confinement, electronic band structures and transport mechanisms are severely modified. The limits for ballistic transport depend on several factors owing to the fluctuations of microscopic scattering sources. For instance, scattering from impurity charges continuously and varies with downsizing device features as a consequence of size-dependent screening phenomena. One of the important limiting phenomena of ballistic transport is the surface roughness (SR) which is unavoidable at the atomistic scale. We have been investigating SR effects in SiNWs with diameter in the range of a few nanometers.

The electronic structure of the SiNWs is described by an accurate third nearest neighbor sp^3 tight-binding (TB) Hamiltonian, previously validated by *ab initio* calculations and comparison with experimental data.^{26,27} Based on such reparametrized Hamiltonian, one

$$L_r = 2.17 \text{ nm}$$

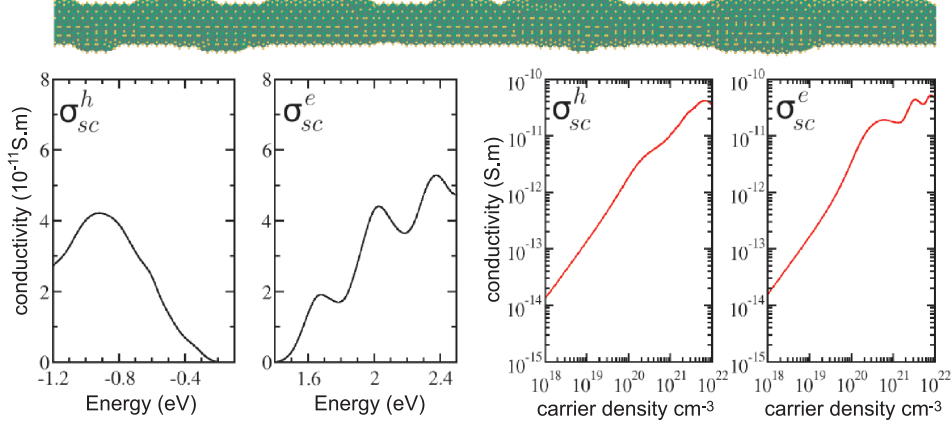


Figure 7. Top: Illustration of a small diameter SiNWs with a given roughness profile. The length scale of recurrent radius fluctuations is given by L_r (see text). Bottom: Computed charge conductivity for hole and electron as a function of charge energy or charge carrier density. Reproduced after Ref. 30.

proceeds in a real space implementation of the Kubo conductivity which allows exploring quantum transport in micron long and disordered nanowires (chemical dopants, surface roughness).

6.2 Results

Figure 6 shows a typical SR profile of our simulated nanowires, together with typical profiles observed in experiments. The surface roughness profile is characterized by the rms of the radius variations and by a correlation length L_r (the typical length scale of these fluctuations). The analysis of the roughness effect on ballistic transport has been achieved by using two complementary approaches: an order N Kubo-Greenwood method, which gives a straightforward access to the intrinsic elastic mean free paths and charge mobilities^{8,9,28} and a Landauer-Büttiker approach²⁹ which is particularly well suited to the quasi-ballistic regime, where contact effects start to prevail over intrinsic phenomena. Both methods have been implemented numerically and extensive use of supercomputing facilities has allowed extracting quantitatively the elastic mean free path that fixes the limit for ballistic conduction.

Figure 7 shows the computed charge conductivity for holes and electrons as a function of Fermi energy or charge carrier density, and a given roughness profile defined by L_r (the typical length scale of fluctuations of the radius).³⁰ The resulting room temperature mobility is plotted as a function of the carrier concentration in Figure 8 for ultimate SiNWs.

The room-temperature mobility is plotted as a function of the carrier concentration in Figure 8 for ultimate SiNWs with radius $R = 1 \text{ nm}$ and three different orientations ([001], [110] and [111]). The rms of the radius fluctuations is $\langle R^2 \rangle = 1 \text{ \AA}$ and the typical length scale of these fluctuations is $L_r = 2.17 \text{ nm}$. As evidenced in this figure, the

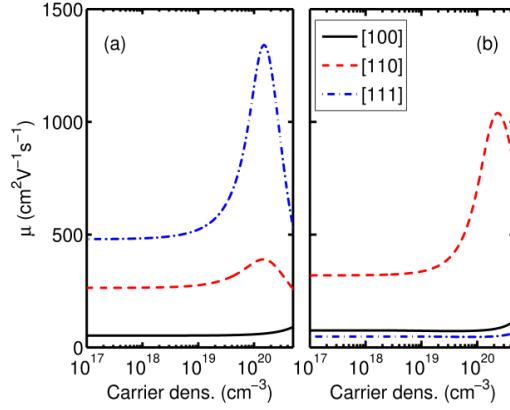


Figure 8. Computed charge mobility for hole (left) and electron (right) as a function of charge carrier density. Reproduced after.³⁰

roughness-limited mobility is highly dependent on the nanowire orientation. Indeed, such small nanowires are in the quantum regime where only one or a few subbands are occupied and available for charge transport at room temperature. Due to the anisotropy of the band structure of bulk silicon, the electronic properties (effective masses of the electron and holes, subbands degeneracies and splittings) of the nanowires are strongly dependent on their orientation. For example, the 6-fold degeneracy between the conduction band valleys of bulk silicon is completely lifted in [110]-oriented nanowires, which suppresses inter-valley scattering at low electron energies. Moreover, the lowest subbands of these [110]-oriented nanowires exhibit a rather light ($0.15 m_0$) effective mass compared to [001]- and [111]-oriented nanowires. This explains why the [110] orientation is found to be the best for electron transport. Likewise, the [111] direction is found to be the best for hole transport, because the hole mass is light in these nanowires, and because the splitting between the highest two valence subbands is the largest (150 meV), therefore inhibiting inter-band scattering at low carrier concentration. The above trends are expected to hold as long as the inter-valley splitting in the conduction band, or the splittings between the highest two valence bands is somewhat greater than $k_B T$, i.e. for radius $R < 3 \text{ nm}$ at room temperature. Quantum effects should average out beyond this radius.

7 Organic Semiconductors

7.1 Introduction

The last example shows the case of organic semiconductors which find applications in a variety of devices including organic light emitting diodes,^{31–34} organic field effect transistors^{35–40} organic thin film transistors,^{41–43} organic solar cells^{44,45} and organic spintronic devices.^{46–48} Such electronic and opto-electronic devices depend critically on the charge-

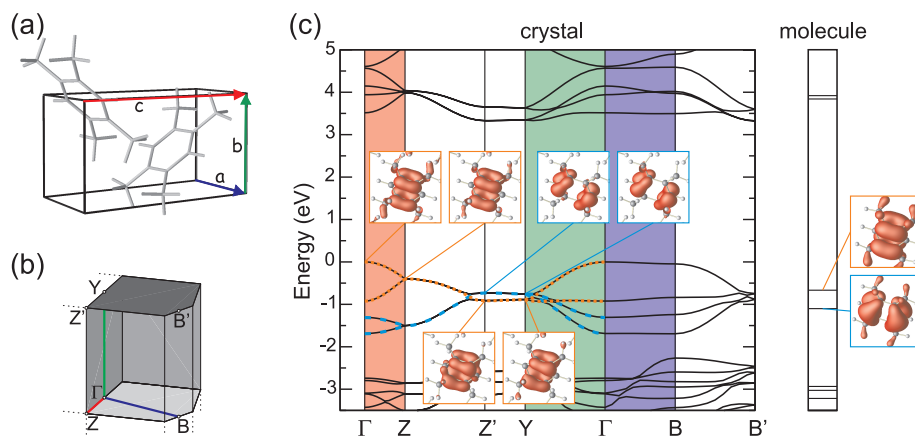


Figure 9. Structure (a), Brillouin zone (b) and band structure (c) of durene (tetramethylbenzene) crystals. (a) The unit cell consists of two molecules in a herringbone arrangement. The b lattice constant is 5.5\AA .⁵⁰ (c) Different background colors indicate directions in the BZ (b) and corresponding real-space directions in the crystal (a). Insets show wave functions for the highest occupied molecular orbital (HOMO) and HOMO-1 (right) compared to crystalline states derived from the HOMO and HOMO-1 (middle). Reproduced after Ref. 51.

carrier mobility in these materials. Unfortunately, the theoretical description of charge-transport processes is significantly complicated with respect to the above-described approach because of the complex materials but also because of the finite temperatures at which such devices operate. This will be explained here below.

Organic semiconductors even in crystalline phase are different from their inorganic counterparts that dominate the field of traditional semiconductor physics such as silicon or GaAs. In fact, the electronic bandwidth of organic crystals is relatively small and rarely reaches 0.5 eV (one exception is shown in Figure 9), while for example the graphene π band has a bandwidth of about 15 eV . Similar values are present for silicon. The reason for the order-of-magnitude difference is related to the fact that molecular orbitals (instead of atomic orbitals) are the basic electronic elements and couple weaker to each other. This is related to the longer distances between molecules (typically larger than 3.5 \AA) compared to a C-C bond length of 1.4 \AA in graphene (cf. Figure 3) which explains the weaker overlap given an exponential decay of the wave functions. Consequently the electronic coupling is well one order of magnitude below. Additionally the complex nodal structure of molecular orbitals (cf. Figure 9 (c)) leads eventually to a further reduction of their mutual interaction.⁴⁹

Another characteristic of organic semiconductors which is of equal importance for charge transport is that vibrational frequencies in organic materials are typically very low. Weak intermolecular forces, which are of the van der Waals type or weak hydrogen bonds, and large molecular masses lead to intermolecular modes with wavenumbers below 300 cm^{-1} . This corresponds to an energy scale which is easily accessible at room temperature. These low-frequency modes have a strong impact on the electronic structure as they trigger dynamical changes in the transfer integrals due to changes in the mutual orientation and distance of molecular orbitals. Said differently the softness of organic materials leads to a

continuously fluctuating potential landscape for the traveling electrons, an effect that can be captured conceptually by interaction terms between electrons and phonons. A common model for such interaction goes back to Holstein^{52,53} and introduces electron-phonon coupling as follows. For a given Hamiltonian that describes a completely frozen lattice $\{R_{ks}^0\}$ one writes (in second quantization notation)

$$\hat{H} = \sum_{MN} \varepsilon_{MN}^{(0)} \hat{a}_M^\dagger \hat{a}_N. \quad (27)$$

Hereby $\varepsilon_{MN}^{(0)}$ are the transfer integrals between molecular orbitals M and N . When taking possible geometric changes $R_{ks}^0 \rightarrow R_{ks} = R_{ks}^0 + u_{ks}$ (k labels the unit cell and s the atomic basis) into account the transfer integrals change. This change is captured by a Taylor series expansion

$$\hat{H} = \sum_{MN} \left[\varepsilon_{MN}^{(0)} + \sum_{ks} u_{ks} \cdot \nabla_{R_{ks}} \varepsilon_{MN}(\{R_{ks}\})|_{R_{ks}=R_{ks}^0} \right] \hat{a}_M^\dagger \hat{a}_N. \quad (28)$$

An equivalent expression can be obtained by replacing the real-space deflection coordinates u_{ks} for vibrational mode coordinates where we use the harmonic approximation. This leads, after quantization of the vibrational degrees of freedom to the Holstein-Peierls Hamiltonian (including the phonon energy as the last term)⁵⁴

$$\hat{H} = \sum_{MN} \left[\varepsilon_{MN}^{(0)} + \sum_Q \hbar\omega_Q g_{MN}^Q (\hat{b}_Q^\dagger + \hat{b}_{-Q}) \right] \hat{a}_M^\dagger \hat{a}_N + \sum_Q \hbar\omega_Q \left(\hat{b}_Q^\dagger \hat{b}_Q + \frac{1}{2} \right) \quad (29)$$

where g_{MN}^Q is the electron-phonon coupling constant associated to the mode Q and the transfer integral ε_{MN} and ω_Q is the phonon frequency. A simpler model of the electron-phonon interaction restricts to local coupling g_{MM}^Q only. Non-local terms g_{MN}^Q are set zero for $M \neq N$ and only the onsite-energy ε_{MN} is effectively coupled to vibrations.

$$\hat{H} = \sum_{MN} \varepsilon_{MN}^{(0)} \hat{a}_M^\dagger \hat{a}_N + \sum_M \sum_Q \hbar\omega_Q g_{MM}^Q (\hat{b}_Q^\dagger + \hat{b}_{-Q}) \hat{a}_M^\dagger \hat{a}_M + \sum_Q \hbar\omega_Q \left(\hat{b}_Q^\dagger \hat{b}_Q + \frac{1}{2} \right) \quad (30)$$

This form of the Hamiltonian is known als Holstein-Hamiltonian.⁵²

In order to determine this Hamiltonian for the particular system under study one has to fix the parameters $\varepsilon_{MN}^{(0)}$, ω_Q , and g_{MM}^Q . In the spirit of a multiscale approach they can be determined from DFT simulations. Figure 9 shows an example where such DFT simulations have been performed for durene crystals. Figure 9 (c) compares the crystal band structure (black solid lines) to an effective tight-binding model based on a set of $\varepsilon_{MN}^{(0)}$ for the HOMO bands (red dotted lines). The largest transfer integral is found in b direction with $\varepsilon_b = 116$ meV. Together with the second largest transfer integral of similar size this finally is responsible for the huge band width of almost 1 eV. Note that in Figure 9 (c) the HOMO set of bands and the bands derived from the next lowest molecular orbital, the HOMO-1 bands, overlap in a certain energy window. This however, does not imply a larger accessible bandwidth for the charge carriers nor a higher mobility.

The electron-phonon coupling constants are obtained in a similar way from the band structure. First one has to modify the atomic coordinates of the crystal according to a considered phonon eigenvector. For such a geometry a DFT calculation has to be performed

which is known as *frozen phonon method*. The implied changes in the electronic structure can be measured with respect to the fixed ground state in the spirit of the Taylor expansion mentioned above. If one does this procedure for a few amplitudes (positive and negative) one can extract the electron-phonon coupling constants from the linear changes in the transfer integrals upon geometry change. This has to be done for all the phonons which are considered relevant.

Transport modeling consists of the simulation of charge transport in organic matter using a Kubo approach similar to the one introduced above. This methodology evaluates the macroscopic current response to an applied electric field such as probed in standard measurements on carrier mobilities in organic semiconductors (see, e.g. Figure 10) and will be detailed below.

7.2 Theory and Modeling of Transport Processes

Like in Sect. 2 we also start with Eq. (3) for the conductivity but now the Hamiltonian to calculate the current-current correlation function is the Holstein Hamiltonian (30). The additional phonon-related terms in (30) give rise to additional contributions to the conductivity in comparison to what we have discussed for graphene. One qualitative and important difference is that phonon-assisted transport plays a role at ambient temperatures.⁵¹ Such contributions are visible in Figure 10 which shows theoretical carrier mobilities of holes in naphthalene compared to experimental measurements on highly purified single crystals. The effect of disorder is assumed to be minor in the theoretical study while the strong temperature-dependence is governed by phonon scattering of charge carriers. Phonon-assisted transport is denoted $\mu^{(\text{inc})}$ (left panel). As becomes clear from the left panel, phonon-promoted carrier transport is an important and dominant contribution for elevated temperatures simply because of the increasing number of phonons at high T . The dual role of the phonons (i) acting as scatterers for the charge carriers thus hindering transport and (ii) promoting transport through thermally assisted contributions is reflected in a crossover from dominant coherent to incoherent transport at a certain temperature (cf. Figure 10).

Conceptually this is described in a polaron picture where the charge carriers do not simply propagate as bare particles but are accompanied by a lattice deformation. One can imagine this as a cloud of phonons that dresses the carriers. It is clear that this composite particle (quasiparticle) is usually more heavy than the bare electron or hole which is reflected in reduced transfer integrals in the polaron picture $\varepsilon \rightarrow \tilde{\varepsilon}(g) < \varepsilon$.

In this picture the carrier mobility can be approximated as⁵⁷

$$\begin{aligned} \mu_{\alpha\beta} = & -\frac{1}{e_0 N_c 2k_B T} \left(\frac{e_0}{\hbar}\right)^2 \sum_{LMN} R_{L\alpha} \tilde{\varepsilon}_L R_{N\beta} \tilde{\varepsilon}_N \\ & \times \frac{1}{N_\Omega} \sum_{\mathbf{k}_1 \mathbf{k}_2} e^{-i\mathbf{k}_1(\mathbf{R}_M + \mathbf{R}_N)} e^{i\mathbf{k}_2(\mathbf{R}_M - \mathbf{R}_L)} n_{\mathbf{k}_1} (1 - n_{\mathbf{k}_2}) \\ & \times \int_{-\infty}^{\infty} dt e^{\frac{it}{\hbar} [\tilde{\varepsilon}(\mathbf{k}_1) - \tilde{\varepsilon}(\mathbf{k}_2)]} e^{-[\sum_{\mathbf{Q}} \Phi_{\mathbf{Q}}(t) G_{0L0N}^{\mathbf{Q}} e^{-i\mathbf{Q}\mathbf{R}_M}] e^{(\frac{-t}{\tau})^2}} \end{aligned} \quad (31)$$

where all the microscopic parameters from Hamiltonian (30), which are computed in the *ab initio* framework^{58,59} (such as transfer integrals ε_{MN} etc.), but also position vec-

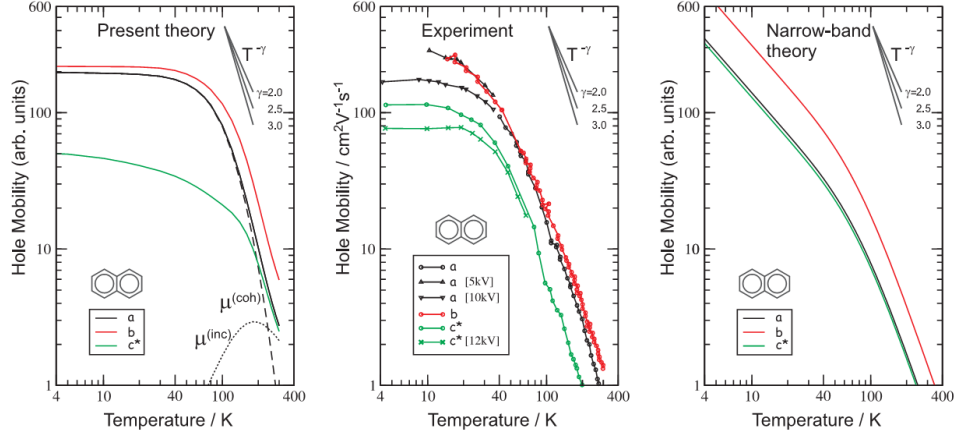


Figure 10. Charge carrier mobility (holes) in ultrapure naphthalene organic single crystals. Comparison between experimental results (middle), new theoretical prediction (left) as well as previous theory (right) for mobility anisotropy and temperature dependence. Adapted from Ref. 55. Experiment from Ref. 56. The molecular structure of Naphthalene is shown as insets.

tors \mathbf{R} , enter. In Eq. (31) the phonon occupation number is described by the Bose-Einstein statistics $N_{\mathbf{Q}} = \left(e^{\frac{\hbar\omega_{\mathbf{Q}}}{k_B T}} - 1 \right)^{-1}$ and impacts through the auxiliary function $\Phi_{\mathbf{Q}}(t) = N_{\mathbf{Q}} e^{i\omega_{\mathbf{Q}} t} + (1 + N_{\mathbf{Q}}) e^{-i\omega_{\mathbf{Q}} t}$ on phonon-absorption and phonon-emission events during transport. Finally the electron-phonon coupling enters in the polaron transfer integrals $\tilde{\epsilon}$, the polaron band structure $\tilde{\epsilon}(\mathbf{k})$ as well as in the effective coupling constant $G_{0L0N}^{\mathbf{Q}} = (g_{00}^{\mathbf{Q}} - g_{LL}^{\mathbf{Q}}) (g_{00}^{-\mathbf{Q}} - g_{NN}^{-\mathbf{Q}})$. The above-discussed phonon-assisted contributions to transport $\mu^{(\text{inc})}$ can be obtained from subtracting the coherent ones from the total mobility $\mu - \mu^{(\text{coh})}$ while $\mu^{(\text{coh})}$ is directly obtained from setting $G_{0L0N}^{\mathbf{Q}} = 0$ for all modes \mathbf{Q} in Eq. (31).

Recent advances of multiscale modeling include the simulation of polaron transport in disordered systems⁶⁰ which is an extension with respect to previous conventional approaches for ultrapure systems⁵⁵ such as displayed in Figure 10. A significant step forward in our understanding of transport in organic matter can be achieved when the impact of disorder is further clarified on a microscopic scale. In particular the interplay of impurity scatterers with phonons (dynamic scatterers) is poorly understood so far. One of the fundamental open questions is still the transport regime as a function of temperature (coherent, or phonon-assisted hopping) as well as the direction and dimensionality dependence. Consequently, besides the study of intrinsic properties presented in Figure 10, the impact of structural disorder and dopants on transport is equally important. The way in which disorder effects are described in the novel approach is essentially in a multiscale fashion by simulating a macroscopic sample of a size of few hundred nanometers in a 3D structure.⁶⁰ It is based on assumed material parameters such as known from previous studies and complemented with a given disorder potential $H_W = \sum_i w_i a_i^\dagger a_i$ in real space ($w_i \in [-W/2, W/2]$ and random).

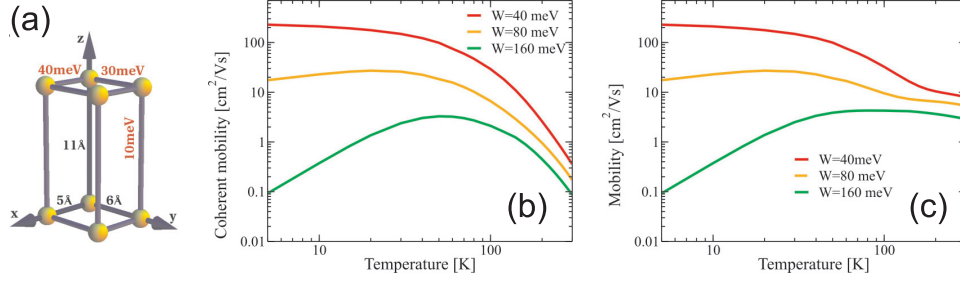


Figure 11. (a) 3D crystal structure with transfer integrals and lattice constants as indicated are used as input parameters for the model crystal. (b) Disorder dependent coherent transport and (c) disorder dependent total carrier mobility for this model. A carrier concentration of 10^{-3} , phonon mode of $\hbar\omega = 10$ meV and electron-phonon coupling of $g = 0.7$ have been used.

It can be shown that in the disordered situation $\mu^{(\text{coh})}(T)$ can be calculated as follows

$$\mu^{(\text{coh})}(T) = \frac{\Omega_{\text{at}}}{e_0 c k_B T} \int dE \sigma(E) n(E) [1 - n(E)] \quad (32)$$

while phonon-assisted contributions read

$$\begin{aligned} \mu^{(\text{inc})}(T) = & \frac{e_0 \Omega_{\text{at}}^2}{2 c \hbar^2 k_B T} \sum_M \tilde{\varepsilon}_M^2 R_M^2 \int dE_1 \int dE_2 \rho(E_1) \rho(E_2) n(E_1) [1 - n(E_2)] \\ & \times \int_{-\infty}^{\infty} dt e^{it(E_1 - E_2)} \{ \exp[2\Phi_\lambda(t) g_\lambda^2] - 1 \} \end{aligned} \quad (33)$$

where we pick out a certain transport direction. In Eq. (33) we use the disordered density of states $\rho(E)$. We find here an important connection of Eq. (32) to the approach displayed in Sect. 2 namely that $\sigma(E)$ has to be calculated in essentially the same way as in Eq. (5) only with electrons replaced by polarons.

The model introduced in Figure 11 gives rise to a temperature dependence of the mobility in Figure 11 (c). At low temperatures $\mu = \mu^{\text{coh}}$. The low- T mobility decay is due to defect scattering where phonons are not important. At high T μ decays with T or is relatively T independent for a certain range of disorder strength (expressed in the Anderson model by parameter W).

8 Conclusion and Perspective

In these lectures, we presented state of the art of multiscale transport modeling at the crossroad of material science and nanotechnology. We demonstrate the applicability of the concepts and methods for a variety of materials and structures including metals and semiconductors and ranging from 1D nanowires over 2D graphene to 3D organic crystals.

Many fields such as organic electronics, spintronics, beyond CMOS nanoelectronics, nanoelectromechanical devices, nanosensors, nanophotonics and nanophononics devices genuinely lack standardized and enabling tools, that are however mandatory to assess the potential of new concepts, or to adapt processes and architectures to achieve the desire

functionalities. The multiscale computational methodologies have to be versatile enough to explore those novel physical phenomena that require advanced quantum mechanics, while at the same time strong efforts are devoted to reach high level of predictability efficiency, therefore providing guidance for experiments and technology.

9 Acknowledgements

We gratefully acknowledge support from the Spanish Ministry of Economy and Competitiveness for national project funding (MAT2012-33911) and from SAMSUNG within the Global Innovation Program. Computing time has been granted by the Barcelona Supercomputing Center–Centro Nacional de Supercomputación, the Spanish Supercomputing Network.

References

1. F. Ortman, S. Roche, J. C. Greer, G. Huhs, X. Oriols, T. Shulthess, T. Deutsch, P. Weinberger, M. Payne, J. M. Sellier, J. Sprekels, J. Weinbub, K. Rupp, M. Nedjalkov, D. Vasilev, E. Alfinito, L. Reggiani, D. Guerra, D. K. Ferry, M. Saraniti, S. M. Goodnick, A. Kloe, L. Colombo, K. Lilja, J. Mateos, T. Gonzales, E. Velazquez, P. Palestri, A. Schenk, and M. Macucci, “Multi-scale modelling for devices and circuits”, Phantoms Foundation, April 2012.
2. <http://www.abinit.org/>
3. J. M. Soler, E. Artacho, J. D. Gale, A. García, J. Junquera, P. Ordejón, and D. Sánchez-Portal, *The SIESTA method for ab-initio order-N materials simulation*, J. Phys. Cond. Mat., **14**, 2745, 2002.
4. G. Kresse and J. Furthmüller, *Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set*, Comput. Mater. Sci., **6**, 15, 1996.
5. R. Kubo, *Statistical-Mechanical Theory of Irreversible Processes. I.*, J. Phys. Soc. Jap., **12**, 570, 1957.
6. G. D. Mahan, *Many-Particle Physics*, Kluwer Academic Publishers, New York, 2000.
7. S. Roche and D. Mayou, *Conductivity of quasiperiodic systems: A numerical study*, Phys. Rev. Lett., **79**, no. 13, 2518–2521, SEP 29 1997.
8. S. Roche, *Quantum transport by means of $O(N)$ real-space methods*, Phys. Rev. B, **59**, no. 3, 2284–2291, 1999.
9. S. Roche and R. Saito, *Magnetoresistance of Carbon Nanotubes: From Molecular to Mesoscopic Fingerprints*, Phys. Rev. Lett., **87**, 246803, 2001.
10. F. Ortman, A. Cresti, G. Montambaux, and S. Roche, *Magnetoresistance in disordered graphene: The role of pseudospin and dimensionality effects unraveled*, EPL, **94**, 47006, 2011.
11. Hiroyuki Ishii, Francois Triozon, Nobuhiko Kobayashi, Kenji Hirose, and Stephan Roche, *Charge transport in carbon nanotubes based materials: a Kubo-Greenwood computational approach*, C. R. Phys., **10**, no. 4, 283–296, MAY 2009.
12. G. Cuniberti, G. Fagas, and K. Richter, (Eds.), *Introducing Molecular Electronics*, Springer, Berlin, 2005.

13. S. Datta, *Electronic Transport in Mesoscopic Systems*, Cambridge University Press, Cambridge, 1995.
14. P. Hohenberg and W. Kohn, *Inhomogeneous Electron Gas*, Phys. Rev., **136**, B864, 1964.
15. R. M. Dreizler and E. K. U. Gross, *Density Functional Theory*, Springer Verlag, Berlin, 1990.
16. Frank Ortmann, “Vergleichende Simulationen von van der Waals-gebundenen Systemen in Dichtefunktionaltheorie mit einer semiempirischen Erweiterung”, Master’s thesis, Friedrich-Schiller Universität Jena, Germany, 2005.
17. W. Kohn and L. J. Sham, *Self-consistent equations including exchange and correlation effects*, Phys. Rev., **140**, A1133, 1965.
18. L. J. Sham and W. Kohn, *One-Particle Properties of an Inhomogeneous Interacting Electron Gas*, Phys. Rev., **145**, 561, 1966.
19. A. Cresti, N. Nemec, B. Biel, G. Niebler, F. Triozon, G. Cuniberti, and S. Roche, Nano Res., **1**, 361, 2008.
20. U. Treske, F. Ortmann, B. Oetzel, K. Hannewald, and F. Bechstedt, *Electronic and Transport Properties of Graphene Nanoribbons*, Phys. Stat. Sol. B, **207**, 304, 2010.
21. Sylvain Latil, Stephan Roche, Didier Mayou, and Jean-Christophe Charlier, *Mesoscopic Transport in Chemically Doped Carbon Nanotubes*, Phys. Rev. Lett., **92**, 256805, Jun 2004.
22. Blanca Biel, Francois Triozon, X. Blase, and Stephan Roche, *Chemically Induced Mobility Gaps in Graphene Nanoribbons: A Route for Upscaling Device Performances*, Nano Lett., **9**, no. 7, 2725–2729, JUL 2009.
23. Blanca Biel, X. Blase, Francois Triozon, and Stephan Roche, *anomalous Doping Effects on Charge Transport in Graphene Nanoribbons*, Phys. Rev. Lett., **102**, no. 9, MAR 6 2009.
24. P. Marconcini, A. Cresti, F. Triozon, G. Fiori, B. Biel, Y.-M. Niquet, M. Macucci, and S. Roche, ACS Nano, **6**, 7942, 2012.
25. S. Roche, T. Poiroux, G. Lecarval, S. Barraud, F. Triozon, M. Persson, and Y.-M. Niquet, *Simulation, modelling and characterisation of quasi-ballistic transport in nanometer sized field effect transistors: from TCAD to atomistic simulation*, Int. J. Nanotechnol., **7**, 348, 2010.
26. Y. M. Niquet, C. Delerue, G. Allan, and M. Lannoo, *Method for tight-binding parametrization: Application to silicon nanostructures*, Phys. Rev. B, **62**, 5109–5116, Aug 2000.
27. Y. M. Niquet, A. Lherbier, N. H. Quang, M. V. Fernández-Serra, X. Blase, and C. Delerue, *Electronic structure of semiconductor nanowires*, Phys. Rev. B, **73**, 165319, Apr 2006.
28. François Triozon, Stephan Roche, Angel Rubio, and Didier Mayou, *Electrical transport in carbon nanotubes: Role of disorder and helical symmetries*, Phys. Rev. B, **69**, 121410, Mar 2004.
29. Rémi Avriller, Sylvain Latil, François Triozon, X. Blase, and Stephan Roche, *Chemical disorder strength in carbon nanotubes: Magnetic tuning of quantum transport regimes*, Phys. Rev. B, **74**, 121406, Sep 2006.
30. Aurélien Lherbier, Martin P. Persson, Yann-Michel Niquet, François Triozon, and Stephan Roche, *Quantum transport length scales in silicon-based semiconducting*

nanowires: Surface roughness effects, Phys. Rev. B, **77**, 085301, Feb 2008.

31. M. Berggren, O. Inganäs, G. Gustafsson, J. Rasmusson, M. R. Andersson, T. Hjertberg, and O. Wennerstrom, *Light-emitting-diodes with variable colors from polymer blends*, Nature, **372**, 444, 1994.
32. A. J. Heeger, *Light emission from semiconducting polymers: Light-emitting diodes, light-emitting electrochemical cells, lasers and white light for the future*, Solid State Commun., **107**, 673, 1998.
33. S. R. Forrest, *The road to high efficiency organic light emitting devices*, Org. Electron., **4**, 45, 2003.
34. S. Reineke, F. Lindner, G. Schwartz, N. Seidler, K. Walzer, B. Lüssem, and K. Leo, *White organic light-emitting diodes with fluorescent tube efficiency*, Nature, **459**, 234–U116, 2009.
35. A. R. Brown, A. Pomp, C. M. Hart, and D. M. de Leeuw, *Logic gates made from polymer transistors and their use in ring oscillators*, Science, **270**, 972, 1995.
36. A. Dodabalapur, L. Torsi, and H. E. Katz, *Organic transistors - 2-dimensional transport and improved electrical characteristics*, Science, **268**, 270, 1995.
37. H. Sirringhaus, N. Tessler, and R. H. Friend, *Integrated optoelectronic devices based on conjugated polymers*, Science, **280**, 1741, 1998.
38. M. Muccini, *A bright future for organic field-effect transistors*, Nat. Mater., **5**, 605, 2006.
39. M. E. Gershenson, V. Podzorov, and A. F. Morpurgo, *Colloquium: Electronic transport in single-crystal organic transistors*, Rev. Mod. Phys., **78**, 973, 2006.
40. Ignacio Gutierrez Lezama, Masaki Nakano, Nikolas A. Minder, Zhihua Chen, Flavia V. Di Girolamo, Antonio Facchetti, and Alberto F. Morpurgo, *Single-crystal organic charge-transfer interfaces probed using Schottky-gated heterostructures*, Nature Mater., **11**, no. 9, 788–794, SEP 2012.
41. H. Klauk, U. Zschieschang, J. Pflaum, and M. Halik, *Ultralow-power organic complementary circuits*, Nature, **445**, 745, 2007.
42. H. Yan, Z. H. Chen, Y. Zheng, C. Newman, J. R. Quinn, F. Dotz, M. Kastler, and A. Facchetti, *A high-mobility electron-transporting polymer for printed transistors*, Nature, **457**, 679, 2009.
43. K. Myny, E. van Veenendaal, G. H. Gelinck, J. Genoe, W. Dehaene, and P. Heremans, *IEEE Journal of Solid-State Circuits*, **47**, 284, 2012.
44. M. Granström, K. Petritsch, A. C. Arias, M. R. Andersson, and R. H. Friend, *Laminated fabrication of polymeric photovoltaic diodes*, Nature, **395**, 257, 1998.
45. Anders Hagfeldt, Gerrit Boschloo, Licheng Sun, Lars Kloo, and Henrik Pettersson, *Dye-Sensitized Solar Cells*, CHEMICAL REVIEWS, **110**, no. 11, 6595–6663, 2010.
46. C. Barraud, P. Seneor, R. Mattana, S. Fusil, K. Bouzeshouane, C. Deranlot, P. Graziosi, L. E. Hueso, I. Bergenti, V. Dediu, F. Petroff, and A. Fert, *Nature Phys.*, **6**, 615, 2010.
47. M. Gobbi, F. Golmar, R. Llopis, F. Casanova, and L. E. Hueso, *Adv. Mater.*, **23**, 1609, 2011.
48. Alberto Riminucci, Mirko Prezioso, Chiara Pernechele, Patrizio Graziosi, Ilaria Bergenti, Raimondo Cecchini, Marco Calbucci, Massimo Solzi, and V. Alek Dediu, *Hanle effect missing in a prototypical organic spintronic device*, Appl. Phys. Lett., **102**, 092407, 2013.
49. J.-L. Brédas, J. P. Calbert, D. A. da Silva Filho, and J. Cornil, *Organic Semicon-*

- ductors: *A theoretical characterization of the basic parameters governing charge transport*, Proc. Natl. Acad. Sci. USA, **99**, 5804, 2002.
50. M. Plazanet, M. R. Johnson, J. D. Gale, T. Yildirim, G. J. Kearley, M. T. Fernández-Díaz, D. Sánchez-Portal, E. Artacho, J. M. Soler, P. Ordejón, A. Garcia, and H. P. Trommsdorff, *The structure and dynamics of crystalline durene by neutron scattering and numerical modelling using density functional methods*, Chem. Phys., **261**, 189, 2000.
 51. F. Ortmann, F. Bechstedt, and K. Hannewald, *Charge transport in organic crystals: Theory and modelling*, Phys. Stat. Sol., **248**, 511, 2011.
 52. T. Holstein, *Studies of Polaron Motion, Part I*, Ann. Phys., **8**, 325, 1959.
 53. T. Holstein, *Studies of Polaron Motion, Part II*, Ann. Phys., **8**, 343, 1959.
 54. K. Hannewald, V. M. Stojanović, J. M. T. Schellekens, P. A. Bobbert, G. Kresse, and J. Hafner, *Theory of polaron band width narrowing in organic molecular crystals*, Phys. Rev. B, **69**, 075211, 2004.
 55. F. Ortmann, F. Bechstedt, and K. Hannewald, *Charge transport in organic crystals: interplay of band transport, hopping and electronphonon scattering*, New J. Phys., **12**, 023011, 2010.
 56. W. Warta and N. Karl, *Hot holes in naphthalene: High, electric-field-dependent mobilities*, Phys. Rev. B, **32**, 1172, 1985.
 57. F. Ortmann, F. Bechstedt, and K. Hannewald, *Theory of charge transport in organic crystals: Beyond Holstein's small-polaron model*, Phys. Rev. B, **79**, 235206, 2009.
 58. K. Hannewald and P. A. Bobbert, *Ab initio theory of charge-carrier transport in ultrapure organic crystals*, Appl. Phys. Lett., **85**, 1535, 2004.
 59. F. Ortmann, K. Hannewald, and F. Bechstedt, *Charge Transport in Guanine-Based Materials*, J. Phys. Chem. B, **113**, 7367, 2009.
 60. F. Ortmann and S. Roche, *Charge transport in organic crystals: Temperature Tuning of Disorder Effects*, Phys. Rev. B, **12**, 023011, 2011.

Electronic Structure of Organic/Organic Interfaces: A Quantum-Chemical Insight

Jérôme Cornil

Laboratory for Chemistry of Novel Materials
University of Mons, Place du Parc 20, B-7000 Mons, Belgium
E-mail: Jerome.Cornil@umons.ac.be

We review here some of our recent theoretical works addressing the nature of the electronic processes occurring at interfaces between two different organic semiconductors. We illustrate that charge-transfer or polarization effects dominate the interface dipole depending on the nature of the compounds under study and that the choice of the DFT functional is critical to get a proper picture. Our discussion is also extended to the energy landscape in the vicinity of organic-organic interfaces, which has strong implications for charge separation processes in solar cells or charge recombination processes in OLEDs.

1 Introduction

The field of organic electronics has experienced a rapid progress during the last decade. The applications of organic semiconductors encompass light-emitting devices (LEDs), solar cells, field-effect transistors and sensors. Many of these devices incorporate several components; this is especially the case in solar cells in which π -donor (D) and π -acceptor (A) compounds are used under the form of a bilayer or a homogeneous blend to dissociate excitations into free charge carriers at their interface. This also applies to light-emitting devices that are generally made of several layers with specific functions (hole/electron transporting layers, exciton blocking layers, emitting layers). Since key mechanisms such as exciton dissociation in solar cells or charge recombination in LEDs occur at the interface between organic semiconductors, a deep understanding of the electronic processes at organic/organic interfaces will prove very useful to develop new strategies towards devices with enhanced efficiencies.

A central issue is to determine the way the frontier electronic levels of two adjacent organic layers align ones with respect to the others at the interface. The Schottky-Mott model is the simplest one that can be applied to organic conjugated materials. In this model, two adjacent organic layers share a common vacuum level. If this holds true, the energetic characteristics of the interface can be designed by tailoring separately the electronic properties of the two materials. This is typically done with organic solar cells by inferring the alignment of the HOMO and LUMO levels of the donor and acceptor units from cyclic voltammetry measurements performed separately for the two compounds. However, recent experimental studies have clearly shown that this picture is usually incorrect.^{1,2} An interface dipole is often induced at the donor/acceptor interface, which shifts the vacuum level of one layer with respect to the other. When approximating the interface dipole by two infinite charged plates, the magnitude of the vacuum level shift (VLS) originating from the charge distribution is given by:

$$\text{VLS} = \frac{eM_z}{\epsilon_0 S} \quad (1)$$

where M_z is the component of the dipole moment of a pair of interacting donor/acceptor molecules in the direction perpendicular to the interface and S is the surface area occupied by the donor-acceptor complex at the interface. Vacuum level shifts at metal/organic interfaces are well documented at both the experimental and theoretical levels.^{1,3,4} In contrast, although there is considerable experimental evidence for vacuum level shifts at organic/organic interfaces, theoretical papers addressing this issue at the quantum-chemical level are still scarce.^{5,6}

In these lectures notes, we will review some recent works aiming at the quantum-chemical description of interface dipoles in donor-acceptor complexes. We will first consider in Section 3 model systems made of a strong donor (tetrathiafulvalene, TTF) and a strong acceptor (tetracyanoquinodimethane, TCNQ, see chemical structures in Figure 2).⁷ This choice is primarily motivated by the fact that: (i) the largest interactions are expected to occur between the molecules facing each other at the interface; and (ii) the VLS between TTF and TNCQ layers has been recently characterized experimentally by Ultraviolet Photoelectron Spectroscopy (UPS) and estimated to be on the order of 0.6 eV.⁸ In Section 4, we will extend the study to large TTF/TCNQ stacks.⁹ We will then consider complexes made of pentacene and C₆₀ molecules¹⁰ in Section 5 before addressing in Section 6 the nature of energy landscapes around organic/organic interfaces and the implications for organic solar cells.¹¹

2 Interface Dipole: Charge Transfer and Polarization Components

The formation of an interface dipole between two organic layers originates from two dominant effects:

i) When the neutral state (DA) is more stable than any charge-transfer excited state (D^+A^- or D^-A^+), the formation of the interface dipole might stem from the admixture of a charge-transfer (CT) character in the ground-state wavefunction describing the donor/acceptor interface. Such a partial charge transfer in donor/acceptor complexes is a well-known phenomenon described previously at the theoretical level.^{12,13} At the second order of perturbation theory, the ground-state wavefunction of a donor-acceptor complex acquires some charge-transfer character due to the admixture of terms corresponding to excited CT states:

$$\Psi(D, A) = a\Psi_0(D, A) + \sum_i b_i\Psi_i(D^+A^-) + \sum_i c_i\Psi_i(D^-A^+) \quad (2)$$

where $\Psi_0(D, A)$ is an antisymmetrized product of the unperturbed wavefunctions of the donor and acceptor molecules in the complex; $\Psi(D^+A^-)$ is the wavefunction of a CT state corresponding to a charge transfer from one occupied level of the donor to one unoccupied level of the acceptor. The first-order correction coefficients (b_i and c_i) to the wavefunction Ψ_0 are equal to $V_i/\Delta E_i$, with V_i the electronic coupling between the ground-state (GS) and the charge-transfer excited state CT_{*i*}, and ΔE_i the corresponding energy separation. Accordingly, the charge transfer admixture in the ground state is given by:

$$q \div \sum_i \pm \left(\frac{V_i}{\Delta E_i} \right)^2 \quad (3)$$

where the sign is determined by the direction of the charge transfer. It is usually assumed that $c_i \ll b_i$, since the $\Psi_i(D^-A^+)$ states are lying at much higher energies, so that a back charge transfer (from the acceptor to the donor) is generally less efficient. From Equation (3), it is clear that the amount of charge transferred at the interface is controlled not only by the difference in electronegativities (which is closely related to ΔE) but also by the electronic coupling between the two molecules. The electronic coupling V is very sensitive to the mutual orientation of the molecules and exponentially decreases with the distance between the molecules (due to the exponential decay of the overlap between the wavefunctions).¹⁴

ii) The interfacial dipole layer may be formed by a polarization of the electronic cloud within the molecules. This contribution stems from the admixture of locally excited states in the ground-state wavefunction:

$$\Psi(D, A) = a\Psi_0(D, A) + \sum_i b_i\Psi(D_0A_i^*) + \sum_i c_i\Psi_i(D_i^*A_0) \quad (4)$$

where D_0 [A_0] and D^* [A^*] stand for the unperturbed ground and excited states of the isolated donor [acceptor]. This mixing results from the fact that the singly excited configurations D^*A and DA^* built from the molecular orbitals of the isolated units are not orthogonal to the ground-state wavefunction calculated for the whole dimer.

3 TTF/TCNQ Model Systems

The choice of the computational method has been guided by its performance in predicting the induced dipole moment in donor/acceptor complexes. For the sake of comparison, we will discuss below the results obtained at the semi-empirical Hartree-Fock Austin Model 1 (AM1)¹⁵ *ab initio* Hartree-Fock (HF) and density functional theory (DFT) levels. The hybrid B3LYP^{16,17} and BHandHLYP functionals introducing 20% and 50% of exact Hartree-Fock exchange, respectively, were used in the DFT calculations, as implemented in Gaussian03.¹⁸ We used a split-valence 6-31G(d) basis set for most HF and DFT calculations. A basis set incorporating polarization functions is expected to be sufficient for the description of the polarization component. The Mulliken charge partitioning scheme exploited here might prove a rather crude approximation, especially when the basis set is augmented with diffuse functions. However, we are confident that the Mulliken charges summed over the individual molecules are meaningful in our case since: (i) the intermolecular overlap is relatively small for the intermolecular distances considered in this study (≥ 3.5 Å); and (ii) we used in most cases the 6-31G(d) basis set which does not contain diffuse functions. The dipole moments and Mulliken charges reported hereafter were corrected for basis set superposition error, using the counterpoise correction of Boys and Bernardi,¹⁹ except for the AM1 results.

Figure 1 shows the evolution of the M_z component of the dipole moment in a TTF/TCNQ dimer as a function of the degree of translation of the TCNQ molecule along its main molecular axis (Y axis), as calculated at different levels of theory (using a 6-31G(d) basis set in the *ab initio* HF and DFT calculations). The initial geometry of the isolated molecules was first optimized at the B3LYP/6-31G(d) level and the dimer was then built in a cofacial geometry by fixing the separation between the molecular planes at 3.5 Å.

The results show that the amount of charge transferred critically depends on the chosen computational approach and on the relative position of the two interacting molecules.

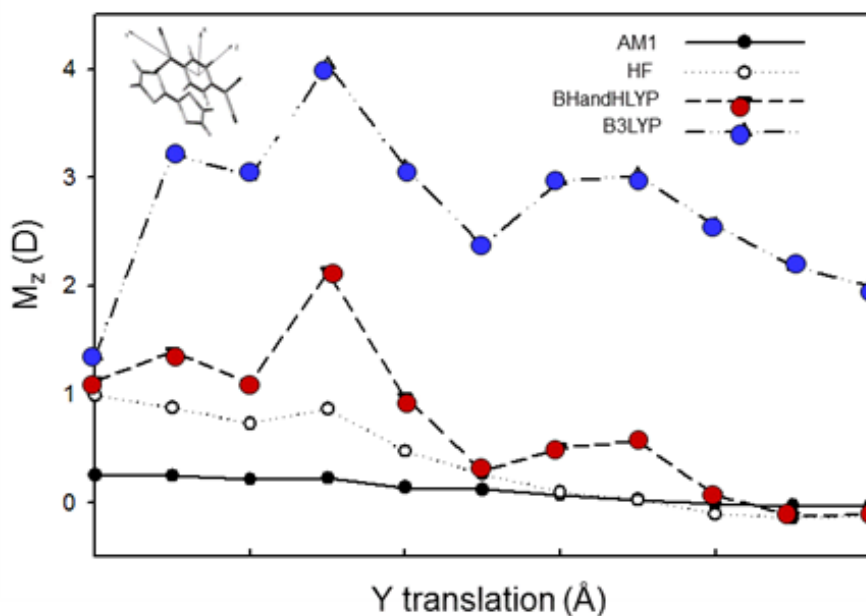


Figure 1. Evolution of the total dipole moment in the direction normal to the molecular planes (M_z) as a function of the degree of lateral translation of the TTF molecule along the Y axis, as obtained with different computational methods. The distance between the molecular planes is fixed at 3.5 Å. Adapted with permission from Ref. 7. Copyright 2009 John Wiley and Sons.

The magnitude of the charge transfer is governed by the calculated energy gap between the frontier orbitals of the donor and acceptor as well as by their electronic coupling. The Hartree-Fock (HF) method yields very large HOMO/LUMO gaps; this is partly due to the overestimation of the energies of the unoccupied levels. Moreover, the energy of a charge-transfer state calculated at the Hartree-Fock level for an isolated complex is larger than the value expected in a condensed medium (*i.e.*, at the interface) due to the neglect of the polarization of the surrounding medium.²⁰ These two effects should lead to an overestimation of the energy of the CT states and hence to a reduced charge transfer with Hartree-Fock. In contrast, DFT is known to provide electronic HOMO/LUMO gaps for *isolated* molecules much smaller than Hartree-Fock-based values and actually close to experimental optical gaps;²¹ by strongly underestimating electronic gaps, DFT thus tends to incorporate artificially medium polarization effects, which prove very useful in the present context.

These considerations are supported by the results of the electronic structure calculations based systematically on the same B3LYP/6-31G(d) input geometry (see Figure 1). AM1 yields the smallest induced dipole moment; the negligible charge transfer occurring

between the donor and acceptor is rationalized by the overestimation of the energy of the CT states and by a smaller polarization component due to the use of a minimal basis set. B3LYP calculations predict the largest amount of charge transfer, with a maximum observed for a shift of 3 Å ($q \sim 0.25|e|$). However, it is worth stressing that with B3LYP the LUMO of TCNQ is found to be lower in energy than the HOMO of TTF. This is certainly due to an insufficient admixture of the Hartree-Fock exchange (20%), which leads to the strong underestimation of HOMO-LUMO gaps in both TTF and TCNQ molecules and of the corresponding gap relevant for charge transfer.

At the BHandHLYP level (incorporating 50% of Hartree-Fock exchange), the maximum charge transfer is $\sim 0.12|e|$ for a shift of 3 Å; the M_z component of the dipole moment reaches 2 D, which yields an estimate for VLS of 0.75 eV on the basis of Equation (1) (with $S=100 \text{ Å}^2$ and by neglecting depolarization effects). This value has the same order of magnitude as the VLS of 0.6 eV measured experimentally,⁸ thus motivating the choice of the BHandHLYP functional in the following. This is also consistent with a number of theoretical studies showing that the BHandHLYP functional provides good estimates for the geometries and transition energies of charge-transfer complexes.^{22,23} The performance of different basis sets has been compared at the BHandHLYP level. The results show that the inclusion of diffuse functions (6-31+G(d)) and of a larger number of polarization functions (6-31G(2df,p)) slightly increases the dipole moment while keeping exactly the same evolution of the charge transfer as a function of the translation.

The consideration of a cofacial TTF/TCNQ complex allows us to separate the charge-transfer *versus* polarization contributions to the dipole moment in a straightforward way. The charge-transfer admixture in the ground state can be evaluated by summing up the Mulliken charges in each molecule of the dimer and by calculating the dipole moment from these. The remaining part of the total dipole moment obtained from the quantum-chemical calculations is then attributed to the polarization component. Figure 2 shows that an increase in the intermolecular distance leads to a fast decrease in the amplitude of the induced dipole moment. Interestingly, we observe that the value of the dipole moment computed from the Mulliken charges decreases almost to zero already for an intermolecular distance of 5 Å; in contrast, the total dipole moment obtained directly from the SCF procedure decreases much slowly when the intermolecular distance is increased from 3.5 to 5 Å and does not reach zero even for an intermolecular distance of 8 Å. This clearly demonstrates that the dipole is induced not only by the charge transfer but also, to a large extent, by polarization effects.

When shifting one molecule with respect to the other along the long molecular axis (Figure 1), the geometry for which the centers of mass of the two molecules are exactly superimposed (*i.e.*, a structure with a C_{2v} symmetry and no shift along the Y axis) does not yield the largest charge transfer, as could be intuitively expected, due to symmetry effects (see below). The latter is actually obtained when TCNQ is shifted by 3 Å along the Y axis. We have also found that there is a full parallelism between the amount of charge transfer and the amplitude of the total dipole moment. In most cases, the charge-transfer contribution dominates the induced dipole at such short intermolecular distances. The non-monotonic dependence of the amount of charge transferred in the ground state is related to variations in the electronic coupling between the highest occupied levels of the donor and the lowest unoccupied levels of the acceptor. Intuitively, the largest contribution to the charge transfer should stem from the HOMO (TTF) \rightarrow LUMO (TCNQ) transition

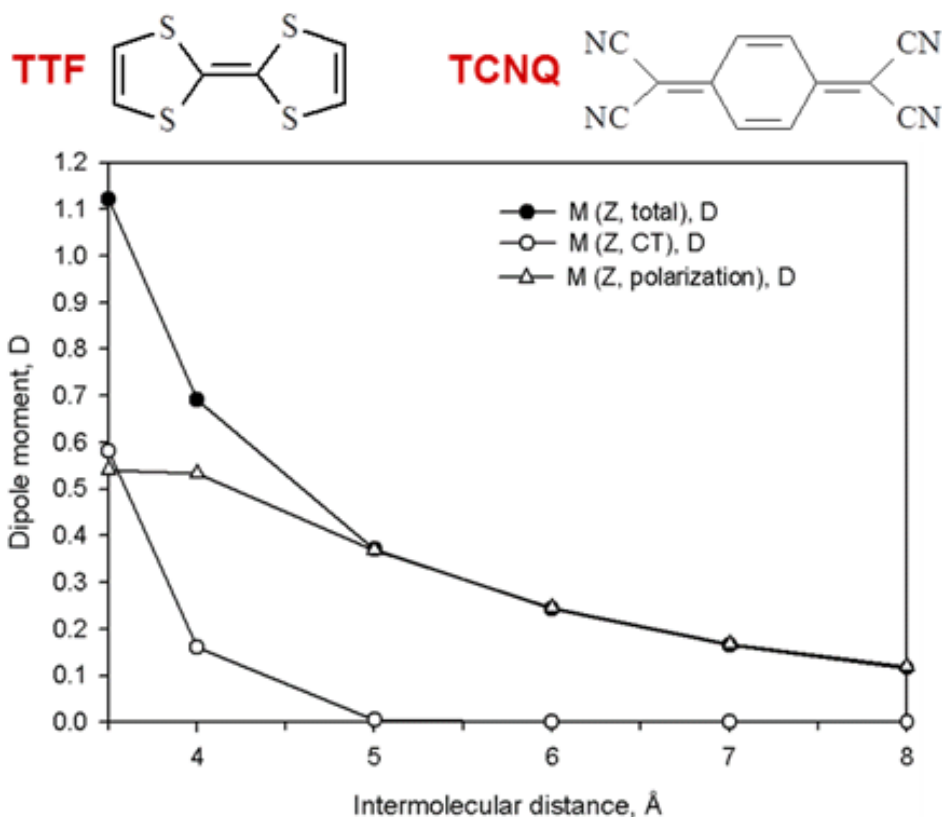


Figure 2. Evolution of the component of the dipole moment normal to the molecular planes obtained from the SCF calculations (filled circles) and from the Mulliken charges (open circles) in a cofacial TTF/TCNQ dimer as a function of the intermolecular distance. The curve with open triangles shows the polarization component of the dipole. We display on top the chemical structures of the two molecules. Adapted with permission from Ref. 7 Copyright 2009 John Wiley and Sons.

due to the fact that the energy separation between them is the smallest. However, the electronic overlap (and hence the electronic coupling) between the HOMO of TTF and the LUMO of TCNQ is equal to zero in the cofacial dimer due to symmetry effects. In this geometry, the largest CT contribution actually arises from HOMO-1 (TTF) \rightarrow LUMO (TCNQ) transition (the electronic coupling between the HOMO of TTF and the LUMO+1 of TCNQ is calculated to be two orders of magnitude smaller). When going away from the cofacial geometry, the amount of CT character in the ground state is generally mostly governed by the HOMO (D) \rightarrow LUMO (A) transition.

The alignment of the frontier electronic levels of the donor and acceptor units is also affected by the creation of the interface dipole when compared to the energy diagram established from the isolated compounds. In the case of the TTF/TCNQ cofacial dimer, both the occupied and unoccupied MOs of TTF experience a decrease in their energy with respect to the MOs of the isolated TTF molecule; this shift is as high as 0.51 eV for the

HOMO level and 0.41 eV for the LUMO. On the contrary, the energies of the frontier MOs of the TCNQ molecule are increased in the TTF/TCNQ dimer (shift of 0.25 eV for the HOMO and of 0.30 eV for the LUMO). The amplitude of the shift varies from orbital to orbital but the direction of the shift is the same for all orbitals of a given compound, including the σ -orbitals. A very nice correlation is actually observed between the amplitude of the induced dipole moment in cofacial dimers with various degrees of translation and the corresponding shift in the orbital energies (see Figure 3). In some cases, the energy shift of a particular orbital is reinforced by a resonant interaction with a deeper orbital of the other molecule; for example, the large shift of the HOMO level of TCNQ for $Y = 1 \text{ \AA}$ is partially promoted by a resonant interaction with the HOMO-1 level of TTF. These results have strong implications for organic solar cells since they demonstrate that the actual offset between the frontier electronic levels of the donor and acceptor components in the device might be significantly different from the value inferred from measurements performed on the isolated compounds.

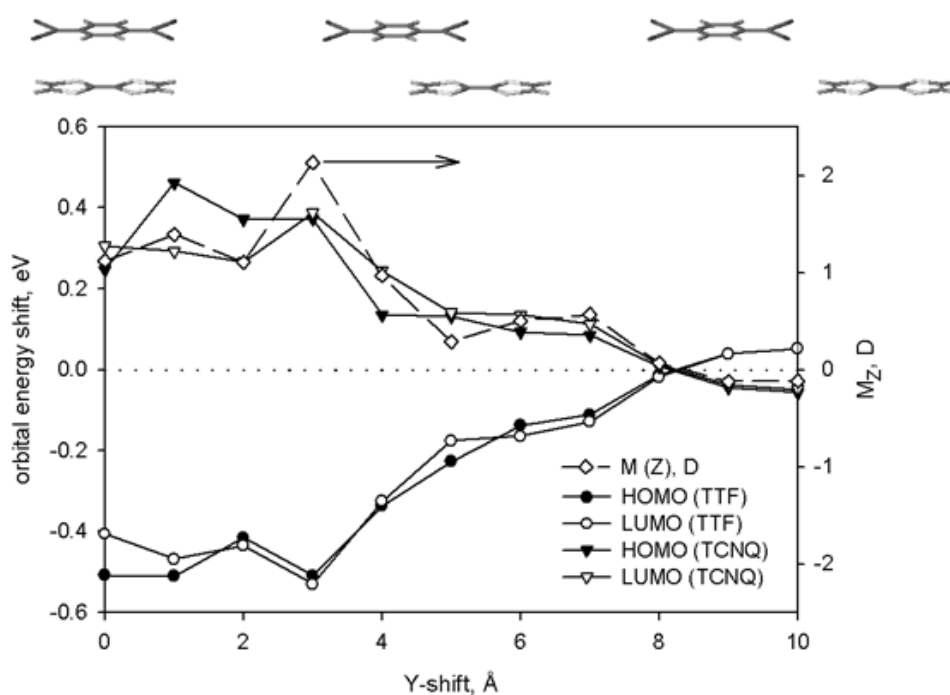


Figure 3. Evolution of the shift of the frontier MOs of TTF (circles) and TCNQ (triangles) versus the amplitude of the dipole moment normal to the molecular planes in a cofacial TTF/TCNQ dimer (diamonds), as a function of the lateral translation along the long molecular axis. The distance between the molecular planes is fixed here at 3.5 Å. Reproduced with permission from Ref. 7. Copyright 2009 John Wiley and Sons.

4 Extended TTF-TCNQ Stacks

We now turn to a description of the evolution of the charge transfer between cofacial TTF and TCNQ stacks of increasing size using different DFT functionals. We start here with the displaced geometry of the complex characterized by a 3-Å translation (that yields the largest charge transfer) and include additional TTF and TCNQ molecules in a perfect cofacial orientation. This results in a slip-stacked structure between a cofacial stack of TTF and a cofacial stack of TCNQ. The term layer used in the following corresponds to one molecule of TTF and one molecule of TCNQ on each side

The evolution with stack size of the dipole moment along the stacking axis, as calculated with BHLYP and a SVP basis set, is presented in Figure 4 which clearly shows that the dipole moment along the stacking axis reaches unrealistic values of about 90 Debyes in the largest stacks; in addition, no convergence is reached with system size. This behavior appears to be in contradiction with UPS measurements that point to a vacuum level shift around 0.6 eV, associated with a much smaller interface dipole.⁸ In order to understand the origin of these large dipole moments, we have performed a Mulliken charge analysis on a stack comprising 6 layers, see Figure 4.

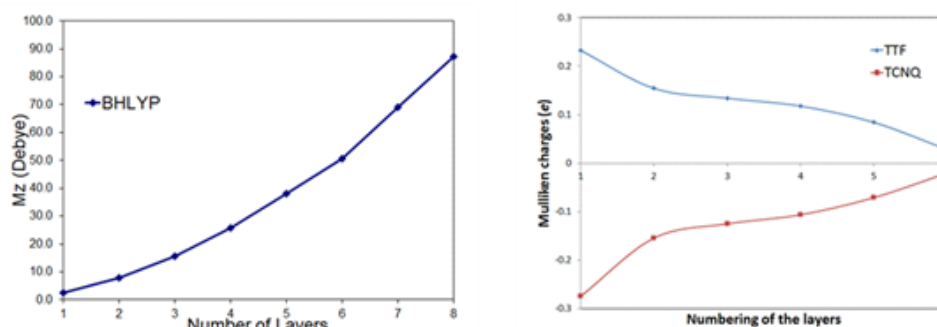


Figure 4. Left: Evolution of the dipole moment along the z-axis with an increasing number of layers at the BHLYP/SVP level; right: Evolution of the charge per molecule within a stack of 6 layers, as calculated with BHLYP/SVP. Adapted with permission from Ref. 9. Copyright 2012 American Institute of Physics.

Figure 4 highlights the large delocalization of the charges within the entire stack. The molecules at the interface bear a significant charge that decreases along the stack though without vanishing at the end of the stack. This unphysical evolution of the dipole moment linked to a rapid crossing of $\text{LUMO}_{\text{TCNQ}}$ and HOMO_{TTF} when increasing the number of layers rules out the use of BHLYP to study extended donor-acceptor complexes. The pronounced charge delocalization is most likely related to the poor description of the long-range interactions in the BHLYP functional. Accordingly, we next turn to long-range corrected (LRC) functionals and present in Figure 5 the evolution of the dipole for stacks containing from 1 to 8 layers, using the $\text{LC-}\omega\text{PBE}^{24}$ and ωB97X^{25} functionals as well as the Hartree-Fock and MP2 methods with the SVP basis set.

The dipole moment calculated with $\text{LC-}\omega\text{PBE}$ and ωB97x for an eight-layer stack amounts to 4.61 D and 6.61 D, respectively, and appears to have nearly converged. Furthermore, ωB97x fits best the values obtained with MP2 considered as benchmark. Note that

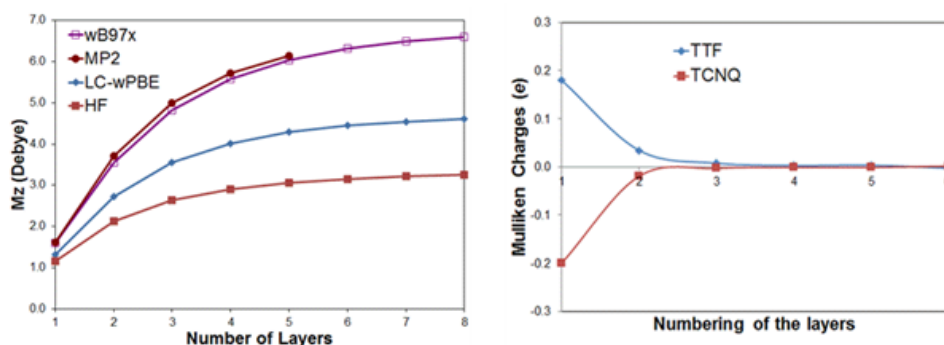


Figure 5. Left: Evolution of the dipole moment along the z-axis with the number of layers for two long-range corrected DFT functionals (LC- ω PBE and ω B97x), MP2, and HF methods combined with the SVP basis set; right: Evolution of the charge per molecule within a stack of 6 layers at the ω B97x/SVP level. Adapted with permission from Ref. 9. Copyright 2012 American Institute of Physics.

the dipole moment calculated with the HF method converges with the number of layers, but tends to an upper limit around 3.25 D for a stack of 8 layers due to the HOMO-LUMO gap overestimation which reduces the amount of charge transfer. In order to understand the difference in behavior between BHLYP and ω B97x, the Mulliken charge distribution within a stack of 6 TTF/TCNQ layers obtained at the ω B97x level (Figure 5) has been compared with the corresponding distribution at the BHLYP level (Figure 4). Figure 5 illustrates that the charges are delocalized along the entire stack with the BHLYP functional. On the other hand, the charge distribution obtained with ω B97x is strongly localized on the interfacial molecules. It gets vanishingly small already on the 3rd layer of the stack and decreases even further away from the interfacial region. This evolution explains the saturation of the dipole moment. The reason for which it is preferable to introduce HF exchange in the long-range region only rather than everywhere in space can be related to a subtle balance of errors between exchange and correlation components of optimized exchange-correlation functionals in the electron-rich short-range region.

5 C₆₀ / Pentacene Complexes

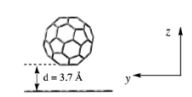
We now turn to the case of pentacene/C₆₀ complexes since these compounds have been widely used in organic solar cells.²⁶ We have first considered a cofacial pentacene/C₆₀ dimer in which a six-membered cycle of the fullerene lies above the pentacene at a distance of 3.7 Å within a C_s symmetry (Table 1). Although not representative of the real morphology of the interface, this simple system provides useful qualitative information on the magnitude and orientation of the induced molecular dipole moment, as well as on the strength of the electrostatic and polarization interactions between the two monomers.

Table 1 reports the component of the total dipole moment parallel to the stacking direction, as calculated using conventional semi-empirical, *ab initio* HF, *post*-HF and DFT methods, as well as the total net charge carried by the fullerene molecule. By convention, the dipole vector is oriented from the negative to the positive pole. Two computational schemes providing atomic charge populations are used, namely the Mulliken and NPA

(Natural Population Analysis) schemes. The former is based on the assumption that off-diagonal elements of the density matrix can be distributed equally among the contributing atomic centers independently of their relative electronegativities. This approximation can lead to an overestimation of charge separations, especially when calculated using diffuse basis sets. In the NPA analysis, the density matrix is divided into blocks of basis functions belonging to one particular atom. Each block is then diagonalized to produce a set of natural atomic orbitals (NAOs) for each atom. The NAOs are eventually orthogonalized such that the diagonal elements of the density matrix in this basis correspond to the orbital populations.

Table 1. The z -component of the total dipole moment of the model dimer (m_z , in Debye), as well as total net charge of the C_{60} molecule (Q_{C60} , in $|e|$) calculated using the Mulliken and NPA schemes.

Level of theory	m_z	Q_{C60} (Mulliken)	Q_{C60} (NPA)
AM1	-0.525	-0.0003	/
RHF/sto-3g	-0.237	-0.0005	-0.0005
RHF/6-31G(d)	-1.048	-0.0067	-0.0046
MP2/6-31G(d)	-1.048	-0.0068	-0.0045
B3LYP/6-31G(d)	-0.997	-0.0106	-0.0076
BH&HLYP/6-31G(d)	-1.008	-0.0086	-0.0059



As shown in Table 1, all theoretical levels provide the same qualitative results: a significant dipole moment is found pointing from the C_{60} towards the pentacene, together with a weak charge transfer between the two molecules (the net charge on C_{60} being slightly negative). This indicates that the major part of the interface dipole originates from polarization effects rather than from a partial charge transfer between the two fragments, in contrast to the situation in TTF/TCNQ complexes. Besides, the dipole magnitude strongly depends of the size of the basis set, as shown by the significant difference between values obtained at the RHF/sto-3g and RHF/6-31G(d) levels. On the contrary, including electron correlation at the MP2 or DFT level (using either the B3LYP or BH&HLYP functional) does not introduce significant changes in the dipole value. Moreover, although underestimated, AM1 provides dipole moments in good qualitative agreement with *ab initio* and DFT results. The fact that AM1 gives smaller absolute values for the dipole is first related to the residual charge transfer which, although small, still exists when using *ab initio* and DFT schemes. Moreover, AM1 is known to underestimate the normal polarizability component with respect to the in-plane components in π -conjugated compounds, due to the lack of flexibility of the minimal valence basis set.

We consider now a single C_{60} molecule interacting with a surface containing 49 pentacene units. Figure 6 illustrates the variation of the induced dipole moment on the C_{60} molecule as a function of its location on the pentacene plane, as calculated at the VB/HF-AM1 level. The induced dipole moment changes sign when the fullerene is translated parallel to the long axis of the pentacene molecules. We note also that the dipole amplitude is weaker than in the dimer, with absolute values smaller than 0.15 D, due to the antagonistic quadrupolar electric fields originating from each pentacene molecule. The orientation of the interfacial dipoles depends on whether the C_{60} center-of-mass is located

on top of the pentacene molecular backbone, or at the edge of the pentacene molecule. This effect can be traced back to the uncompensated quadrupolar field at the interface. The pentacene quadrupole can be viewed as the result of a collection of 14 CH units that are polarized with negative partial charges on the inner carbon atoms and positive partial charges on the outer hydrogen atoms. When the C_{60} molecule mainly interacts with the π -electronic density of the carbon atoms of pentacene, the reorganization of the electronic cloud over the fullerene molecule promotes a sizeable intramolecular charge transfer away from pentacene. Interactions with the hydrogens atoms of pentacene generate the opposite polarization of C_{60} .

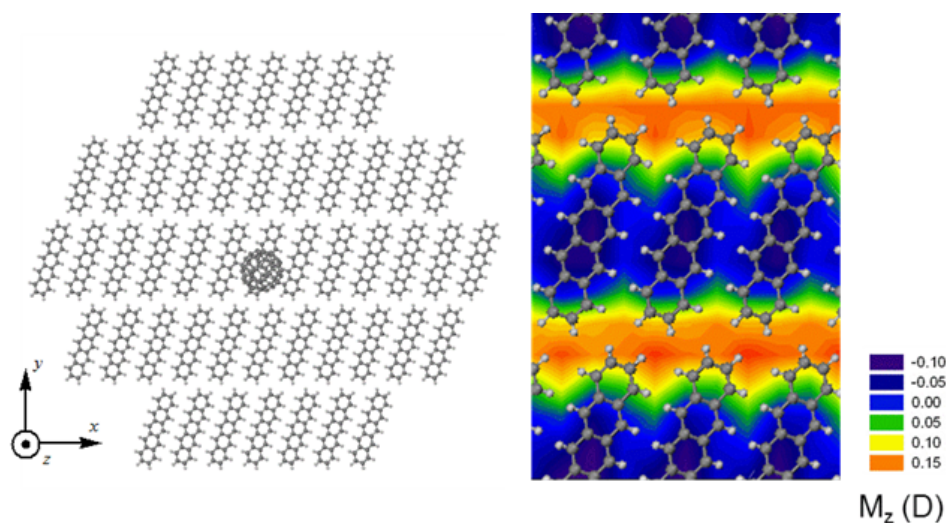


Figure 6. C_{60} molecule above a plane of pentacenes and amplitude of the z -component of the induced dipole on the C_{60} molecule as a function of its position on the (x, y) plane, as calculated using the VB/HF-AM1 model. Reprinted with permission from J. Phys. Chem. C 114, 3215, 2010. Copyright 2010 American Chemical Society.

Interactions between two molecular surfaces have been further investigated by considering aggregates in which C_{60} units are progressively added above a pentacene plane containing 55 molecular units (Figure 7). The evolution of the z -component of the total interface dipole m_z , as well as of the average induced dipole per fullerene unit m_z/N with the number of fullerenes (N) are reported in Figure 7.

The chaotic evolution of the total induced dipole moment m_z is related to the way the C_{60} molecules are progressively added on the pentacene surface. As previously discussed, when a C_{60} molecule is added above the carbon body of a pentacene unit, its molecular induced dipole moment points towards the pentacene plane leading to the decrease of the total interface dipole. On the contrary, m_z increases as the additional C_{60} molecules are located above interstices between pentacene units. These local induced dipoles compensate each other, which has for consequence that the averaged induced dipole per fullerene unit, m_z/N , tends to saturate with N , with a weak asymptotic value. These calculations evidence that the measure of the interfacial dipole averaged over the interface is not rep-

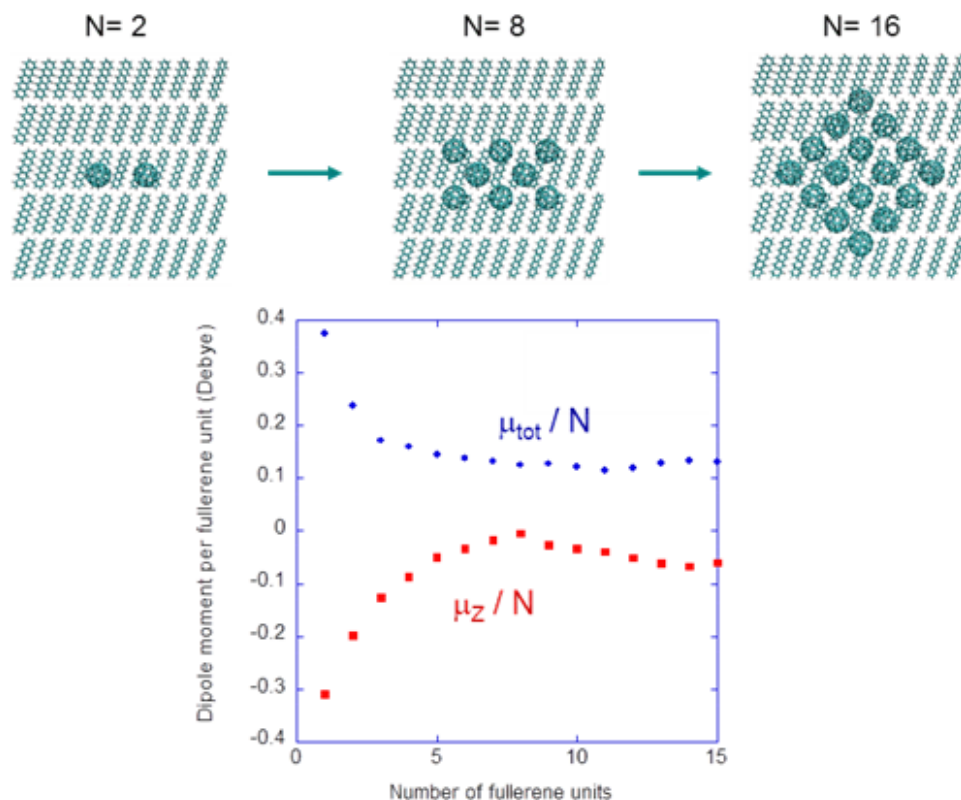


Figure 7. Evolution of the z -component of the total induced dipole moment (black squares) and of the averaged induced dipole (white squares) divided by the number N of C_{60} molecules in the interface (Debye) as a function of N , as calculated at the AM1 level. Reprinted with permission from J. Phys. Chem. C 114, 3215, 2010. Copyright 2010 American Chemical Society.

representative of the local quadrupole-induced dipoles (QID) on the molecular units at the interface.

6 Energy Landscape around Organic/Organic Interfaces

A comprehensive description of the exciton dissociation in photovoltaic cells entails a detailed knowledge of the electronic structure at the heterojunction between the donor (D) and acceptor (A) materials. In organic solar cells, the occurrence of photo-induced charge transfer to produce charge transfer (CT) states requires a proper tuning of the frontier electronic levels of the donor and acceptor molecules.²⁷ In most cases, the choice for the donor and acceptor materials used as active components in organic solar cells is driven by their bulk electronic and optical properties, thus neglecting the impact of interfacial electronic interactions. Here, we demonstrate that such interactions affect: (i) the alignment of the frontier electronic levels of the donor and acceptor molecules; (ii) the energy landscape

explored by charge carriers during the photo-conversion process.

As a proof of principle, we have applied complementary quantum-chemical methods to unravel the electronic structure at oligothiophene/ C_{60} and dicyanovinyl-substituted oligothiophene/ C_{60} interfaces. The ground-state geometry of isolated oligothiophenes (nT), dicyanovinyl-substituted oligothiophenes (DCVnT, with $n=2, 4$ and 6), all imposed planar, and C_{60} molecules has been optimized at the density functional theory (DFT) level using the B3LYP hybrid functional and the 6-31g(d) basis set. The corresponding one-electron energy diagram is shown in Figure 8.

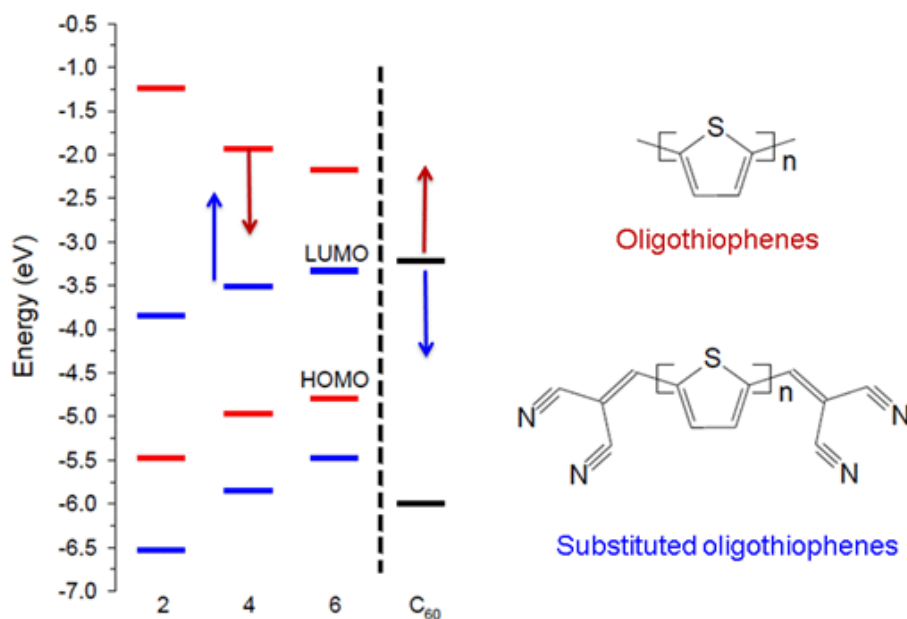


Figure 8. DFT/B3LYP one-electron energy diagram for isolated nT (red lines) and DCVnT (blue lines) with $n=2, 4$ and 6 , and C_{60} . Adapted with permission from J. Phys. Chem. Lett. 3, 2374, 2012. Copyright 2012 American Chemical Society.

As expected from their electron-withdrawing character, substitution of the oligothiophenes by dicyanovinyl end groups causes a down shift in the energy of the frontier molecular orbitals, the effect being larger for the LUMO. As a result, while the large energy offset between the LUMO of unsubstituted oligothiophenes and the LUMO of C_{60} is expected to promote efficient exciton dissociation into free charge carriers, the situation is drastically different in DCVnT/ C_{60} pairs where the driving force for free charge generation (related to first approximation to the LUMO energy offset between the donor and acceptor molecules²⁸) is close to zero or even negative. The changes in electronic structure and the reduced LUMO energy offset with respect to C_{60} in DCVnT compared to nT are consistent with experimental findings.²⁸ Despite the small driving force for charge separation, DCVnT molecules have been successfully exploited in solar cells using C_{60} as acceptor with power conversion efficiencies of 1-3%.²⁸ As described below, this apparent

inconsistency is lifted when accounting for the readjustment of the electronic levels due to interfacial effects.

By comparison to their values in the isolated molecules, both DFT/B3LYP and MP2 calculations show that the HOMO and LUMO levels of unsubstituted oligothiophenes are stabilized in the donor/acceptor dimer while those of the C_{60} molecule are destabilized. The shift in the frontier molecular orbitals at the heterojunction results from the appearance of an interfacial dipole. As the amount of ground-state charge transfer from (DCV)nT to C_{60} is negligible in all cases, the interfacial dipole is mainly associated with the polarization of the electronic cloud of the C_{60} molecule, with the positive pole next to the oligothiophene backbone in the case of nT/ C_{60} dimers. Similarly to pentacene, this effect is primarily attributed to the (uncompensated) quadrupolar electric field generated by the oligothiophenes.

In marked contrast, the HOMO and LUMO levels of dicyanovinyl-substituted oligothiophenes are slightly destabilized in presence of C_{60} while the corresponding C_{60} frontier electronic levels are shifted down. Thus, the strong dicyanovinyl electron-withdrawing moieties perturb the electronic cloud on the oligothiophene backbone and yield an opposite quadrupolar field that in turn swaps the interfacial dipole orientation (now pointing its negative pole towards the oligothiophene backbone within the C_{60} molecule). Very interestingly, unlike the nT/ C_{60} case, the electrostatic effects computed at the DCVnT/ C_{60} interface are found: (i) to increase (by 0.1-0.2 eV) the energy offset between the LUMO levels of DCVnT and C_{60} (which should affect the driving force for charge separation, Δ_{LUMO}); and (ii) to reduce the energy difference between the HOMO level of the donor and the LUMO level of the acceptor (which might affect the open-circuit voltage, V_{oc}). Though the simplicity of the dimer model used here does not allow pulling out a quantitative estimate for Δ_{LUMO} and V_{oc} , it nicely shows how the electronic structure at donor/acceptor interfaces can be controlled by tuning the chemical structure of the interacting molecules (here through grafting electroactive moieties on the donor molecules).

In a next step, model 1D stacks comprising 15 donor and 15 acceptor molecules in a cofacial arrangement (with a 3.5 Å separation) have been built. Changes in the ionization potential of dicyanovinyl-substituted and unsubstituted oligothiophenes as well as in the electronic affinity of C_{60} have been computed as a function of distance to the interface using the Valence Bond/Hartree-Fock (VB/HF) scheme²⁹ at the AM1 level. In the case of the oligothiophene/ C_{60} interfaces, our results are consistent with previous findings on cofacial pentacene/ C_{60} heterojunctions, in the sense that the interfacial electrostatic effects provide an improved driving force for electron-hole pair separation. Indeed, both the positive charged state of oligothiophenes and the negative charged state of C_{60} get destabilized at the vicinity of the interface, Figure 9 top left. Interfacial electrostatic effects thus push the charges into opposite directions, from the interface into the bulk, which might at least partly compensate the loss in Coulomb binding energy. A band bending effect in the opposite direction is predicted in the case of the dicyanovinyl-substituted oligothiophene, as a result of the opposite quadrupole induced interfacial dipole, Figure 9 bottom left. Therefore, in this case, we expect that the interfacial electronic interactions will add to the Coulomb attraction to further stabilize the charge transfer state across the heterojunctions. In the case of the hexathienyl/perylene-tetracarboxylic-dianhydride (PTCDA) interface, a larger reshuffling in the electronic structure at the interface is predicted compared to the 6T/ C_{60} heterojunction, Figure 9 top right. Unlike C_{60} , the PTCDA acceptor molecule in-

deed generates a quadrupolar electric field that adds constructively to the corresponding field generated by the 6T donor molecule, hence the larger band bending. Comparing the DCV6T/PTCDA and 6T/PTCDA interfaces (Figure 9 bottom right), much smaller changes in the energetic positions of the positively and negatively charged states is predicted close to the heterojunction with respect to the bulk in the former case, owing to a partial cancellation of the quadrupolar fields sourced by the two partners. As a rule of thumb, imparting a quadrupolar moment polarized in opposite directions for the donor (with local dipoles having their positive poles lying outwards with respect to the center of the molecule, e.g. C-H in pentacene) and the acceptor (with local dipoles having their positive poles lying inwards with respect to the center of the molecule, e.g. C=O in PTCDA) appears as an attractive strategy to bias the energy landscape in favor of full separation of the charge transfer pairs. Yet, we would like to stress that a complete picture of the exciton dissociation process requires addressing the influence of the relative positions of the molecules on the calculated interfacial dipole and the role of solid-state polarization effects as the charges separate.³⁰

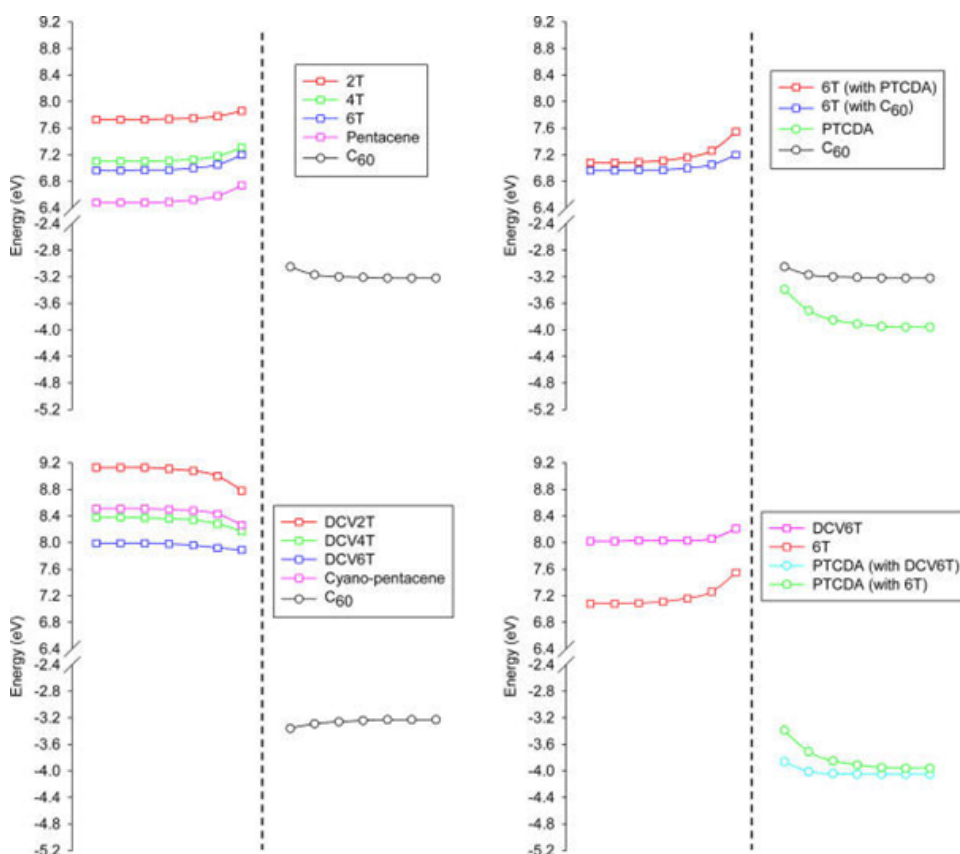


Figure 9. Left: ionization potentials of dicyanovinyl-substituted oligothiophenes and cyano-pentacene (bottom), oligothiophenes and pentacene (top), and electronic affinities of C₆₀ molecules in model 1D stacks, as a function of distance to the interface. Right: ionization potentials of dicyanovinyl-substituted hexathieryl (bottom) and hexathieryl (top), and electronic affinities of perylene-tetracarboxylic-dianhydride (PTCDA) and C₆₀ molecules in model 1D stacks, as a function of distance to the interface. The dashed line indicates the interface. Adapted with permission from J. Phys. Chem. Lett. 3, 2374, 2012. Copyright 2012 American Chemical Society.

7 Conclusions

We have illustrated here through quantum-chemical calculations that interfacial electronic effects cannot be ignored at organic/organic interfaces to provide a proper description of their electronic structure. These interactions promote the formation of an interface dipole that affects in turn the alignment of the electronic levels of the two components; they also generate a gradient of the electronic levels going from the interface to the bulk due to uncompensated electrostatic interactions at the interface. It is now of prime interest to generate realistic morphologies of organic/organic interfaces using force-field calculations to exploit them as input for electronic structure calculations, using in particular for such large systems micro-electrostatic models properly parameterized on the basis of quantum-chemical calculations.³⁰

Acknowledgements

This work has been supported by the European project MINOTOR (FP7-NMP-228424), the Interuniversity Attraction Pole program of the Belgian Federal Science Policy Office (PAI 6/27) and the Belgian National Fund for Scientific Research. I would like to acknowledge all the people that have contributed to the results presented in these lecture notes, namely: Igor Avilov, Victor Geskin, Tanguy Van Regemorter, Maxime Guillaume, Sébastien Mothy, Mathieu Linares, David Beljonne (University of Mons), Kelly Lancaster, Jean-Luc Brédas (Georgia Institute of Technology), Stijn Verlaak, Alexander Mityashin, Paul Heremans (imec), Andreas Fuchs, Christian Lennartz (BASF), Julien Idé, Raphaël Méreau, Philippe Aurel, Laurent Ducasse, Frédéric Castet (University of Bordeaux).

References

1. H. Ishii, K. Sugiyama, E. Ito, K. Seki, *Adv. Mater.* **1999**, *11*, 605.
2. C. Shen, A. Kahn, I. Hill, in *Conjugated Polymer and Molecular Interfaces: Science and Technology for Photonic and Optoelectronic Applications*; (Eds: W.R. Salaneck, K. Seki, A. Kahn, J.-J. Pireaux); Marcel Dekker, New York **2001**, pp. 351-400.
3. D. Cahen, A. Kahn, *Adv. Mater.* **2003**, *15*, 271.
4. X. Crispin, V. Geskin, A. Crispin, J. Cornil, R. Lazzaroni, W. R. Salaneck, J. L. Brédas, *J. Am. Chem. Soc.* **2002**, *124*, 8131.
5. H. Vázquez, W. Gao, F. Flores, A. Kahn, *Phys. Rev. B*, **2005**, *71*, 041306.
6. S.R. Yost, T. Van Voorhis, *J. Phys. Chem. C* **2013**, *117*, 5617.
7. I. Avilov, V. Geskin, J. Cornil, *Adv. Funct. Mat.* **2009**, *19*, 624.
8. R. J. Murdey, W. R. Salaneck, *Jap. J. Appl. Phys.* **2005**, *44*, 3751.
9. T. Van Regemorter, M. Guillaume, A. Fuchs, C. Lennartz, V. Geskin, D. Beljonne, J. Cornil, *J. Chem. Phys.* **2012**, *137*, 174708.
10. M. Linares, D. Beljonne, J. Cornil, K. Lancaster, J.L. Brédas, S. Verlaak, A. Mityashin, P. Heremans, A. Fuchs, C. Lennartz, J. Idé, R. Méreau, P. Aurel, L. Ducasse, F. Castet, *J. Phys. Chem. C* **2010**, *114*, 3125.
11. S. Mothy, M. Guillaume, J. Idé, F. Castet, L. Ducasse, J. Cornil, D. Beljonne, *J. Phys. Chem. Lett.* **2012**, *3*, 2374.

12. R. S. Mulliken, W. B. Person *J. Am. Chem. Soc.* **1969**, *91*, 3409.
13. C. J. Bender, *Chem. Soc. Rev.* **1986**, *15*, 475.
14. J. L. Brédas, J. P. Calbert, D. A. da Silva Filho, J. Cornil, *Proc. Natl. Acad. Sci. USA*, **2002**, *99*, 5804.
15. M. J. S. Dewar, E. G. Zoebisch, E. F. Healy, *J. Am. Chem. Soc.* **1985**, *107*, 3902.
16. A. D. Becke, *J. Chem. Phys.* **1993**, *98*, 5648.
17. C. Lee, W. Yang, R. G. Parr, *Phys. Rev. B*, **1988**, *37*, 785.
18. M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, Jr., T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez, and J. A. Pople, *Gaussian 03*, Revision C.02; Gaussian, Inc.: Wallingford CT, 2004.
19. S. F. Boys, F. Bernardi, *Mol. Phys.* **1970**, *19*, 553.
20. E.V. Tsiper, Z. G. Soos, W. Gaob, A. Kahn, *Chem. Phys. Lett.* **2002**, *360*, 47.
21. U. Salzner, P. G. Pickup, R. A. Poirier, J. B. Lagowski, *J. Phys. Chem. A*, **1998**, *102*, 2572.
22. J. C. Sancho-García, *Chem. Phys.* **2007**, *331*, 321.
23. Y. Zhao, D. G. Truhlar, *J. Chem. Theory Comput.* **2005**, *1*, 415.
24. M.A. Rohrdanz, K.M. Martins, J.M. Herbert, *J. Chem. Phys.* **2009**, *130*, 054112.
25. J.D. Chai, M. Head-Gordon, *J. Chem. Phys.* **2008**, *128*, 084106.
26. S. Yoo, B. Domercq, B. Kippelen, *Appl. Phys. Lett.* **2004**, *85*, 5427.
27. J. L. Brédas, J. E. Norton, J. Cornil, V. Coropceanu, *Acc. Chem. Res.* **2009**, *42*, 1691.
28. R. Fitzner, E. Reinold, A. Mishra, E. Mena-Osteritz, H. Ziehlke, C. Körner, K. Leo, M. Riede, M. Weil, O. Tsaryova, A. Weiß, C. Urich, M. Pfeiffer, P. Bäuerle, *Adv. Funct. Mater.* **2011**, *21*, 897.
29. F. Castet, P. Aurel, A. Fritsch, L. Ducasse, D. Liotard, M. Linares, J. Cornil, D. Beljonne, *Phys. Rev. B* **2008**, *77*, 115210-1.
30. S. Verlaak, D. Beljonne, D. Cheyns, C. Rolin, M. Linares, F. Castet, J. Cornil, P. Heremans, *Adv. Funct. Mat.* **2009**, *19*, 3809.

UNICORE Rich Client User Manual

**Bastian Demuth, Lara Flörke, Björn Hagemeyer,
Daniel Mallmann, Michael Rambadt, Mathilde Romberg, Rajveer Saini,
Bernd Schuller on behalf of the UNICORE Team**

Jülich Supercomputing Centre, Research Centre Jülich, 52425 Jülich, Germany
E-mail: unicore-info@fz-juelich.de

Nowadays more and more scientists require a lot of high performance computing power to run their complex parallel applications. Even if the available systems get increasingly powerful this often is not enough. As a consequence many applications are designed to run not only on one supercomputer but on several in parallel. This implies special software as an essential offer to the scientists to hide the complexity and the heterogeneousness of the underlying systems and architectures.

The UNICORE software provides these features. It comes with a seamless interface for preparing and submitting jobs to a wide variety of heterogeneous distributed computing resources and data storages. It supports users to generate scientific and engineering applications, to submit them and to monitor the results. UNICORE has an integrated extended workflow engine that allows the scientist to create complex multi-step and multi-site jobs.

1 Introduction

This document describes how to install and use the Eclipse based Rich Client for the UNICORE workflow system. UNICORE is a European project that facilitates the access to modern heterogeneous computer networks, so called ‘Grids’. It offers a client-server framework for accessing Grid resources. It has a service oriented architecture (SOA) which means that the functions of the software are grouped into small coherent chunks (named ‘services’) which can be installed on different computer systems.

The client software enables users to create descriptions of work to be performed on the Grid, so called ‘jobs’. A single job usually corresponds to the execution of a computer program on one of the available computer systems in the Grid. Once a job has been created, the UNICORE Rich Client can submit it to a selected computer system. The remote execution of the job can be monitored and output files of the executed program can be downloaded to the user’s computer. In order to accomplish more complex tasks on the Grid, jobs can be embedded into workflows. In our terminology, a workflow is a set of activities (the execution of a single job would be considered an activity), interconnected by transitions that define the order in which the activities must be performed. Workflows can be created and edited graphically. Similar to jobs, they can be submitted to a designated service on the Grid which executes them. Workflow execution can be monitored in multiple ways and resulting output files can be downloaded to the local harddisk. Apart from these basic features, the UNICORE Rich Client offers a bunch of additional functions like browsing and monitoring services on the Grid, managing user certificates, and transferring files to and from Grid storages.

This document is structured into the following parts: Section 2 provides information about the UNICORE history. Section 3 describes the installation procedure and how to

startup the client application. Section 4 gives a brief overview of the basic features and most frequent use cases of this application.

2 A Brief History of UNICORE

The UNICORE (Uniform Interface to Computing Resources) system was originally conceived in 1997 to enable German supercomputer centres to provide their users with a seamless, secure, and intuitive access to the heterogeneous computing resources at the centres. As a result, the projects UNICORE and UNICORE Plus were funded by BMBF, the German Ministry for Education and Research, with the following objectives:

UNICORE was designed to hide the seams resulting from different hardware architectures, vendor specific operating systems, incompatible resource management systems, and different application environments. Retaining organisational and administrative autonomy of the participating centres was a key objective of UNICORE. None of the service providers should be forced to change historically grown computer centre practices, naming conventions, and security policies to be able to use the full benefits of UNICORE. Security was built into the design of UNICORE from the start relying on the X.509 standard. Certificates are used to authenticate servers, software, and users as well as to encrypt the communication over the open internet. Finally, UNICORE had to be usable by scientists and engineers without having to study vendor or site-specific documentation.

Version 6 is a major milestone in the continuous development of the proven Grid software. It retains the rich functionality of previous versions, like seamless access to heterogeneous resources, complex workflows, and secure computing in a Grid environment. Application level brokering has been added to meet user requirements. The graphical user interface has been improved for greater efficiency and ease of use. Some user actions that turned out to be redundant were consequently removed. In addition, the performance of UNICORE 6 has been improved substantially. Both the specific feedback from users and the advent of Grid standards and new implementation tools have contributed greatly to this version. The software has been cleanly implemented from scratch using web service technology and modern programming environments, like Eclipse. This allows to remain interoperable with other standards based Grid solutions, become easily extensible to meet new demands, and - most importantly - stay a safe investment in the future. UNICORE development continues as an open source project that is driven and supported by a dedicated team at the Jülich Supercomputing Centre.

3 Installation and Startup

3.1 Prerequisites

- Operating Systems: currently Linux and Microsoft Windows are supported. ²⁸¹
- Java Runtime Environment: Sun Java 6 or higher is required. ²⁸²

3.2 Procedure

- Download the installation archive that matches your operating system.

- Unzip the archive to the desired location.
- Run the executable called ‘UNICORE_Rich_Client.exe’ (or ‘UNICORE_Rich_Client’, on a Unix/Linux machine). A splash screen will indicate the startup of the client.
- Specify location and passphrase of the keystore file that holds your certificates (see Section 4.3 for details about why this is necessary).

4 Basic Usage Guide

4.1 Welcome screen

When the client is started for the first time, it will display a welcome screen that provides valuable information and helps in making the first steps with the UNICORE Rich Client (see Figure 1).

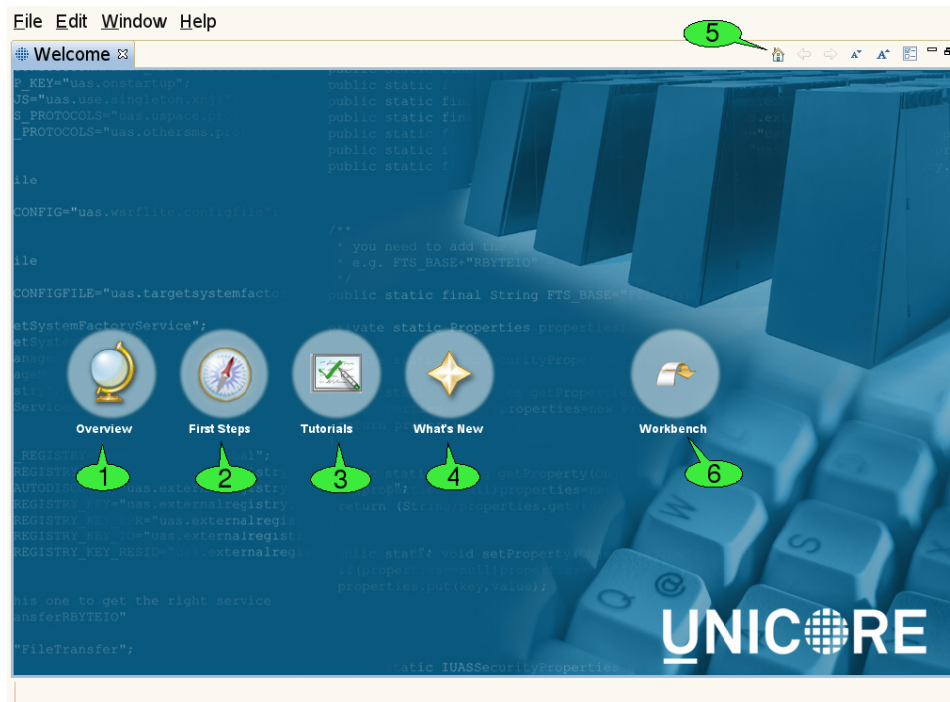


Figure 1. The Welcome screen

The welcome screen is composed of several web pages that are displayed in the internal web browser of the client:

- The *Overview* page 1 contains links to parts of this document and the Eclipse framework’s user manual.

- The *First Steps* page 2 helps in configuring the client for accessing different Grids.
- The *Tutorials* page 3 offers links to Flash-based online tutorials that will be displayed in a web browser.
- The *What's New* page 4 summarizes the most important new features of the current client version and lists general UNICORE related news.

A navigation bar on top of each page contains cites to the other pages. The toolbar of the welcome screen can also be used to navigate back and forth between the pages 5. In order to leave the welcome screen and start working with the client, click the *Workbench* hyperlink 6. The welcome screen can later be re-opened via the *Help* → *Welcome* pull down menu item.

4.2 The Eclipse workbench

The client's main window is called the *workbench* (see Figure 2). It has different components which can be opened, closed, resized, re-ordered and even detached from the main window.

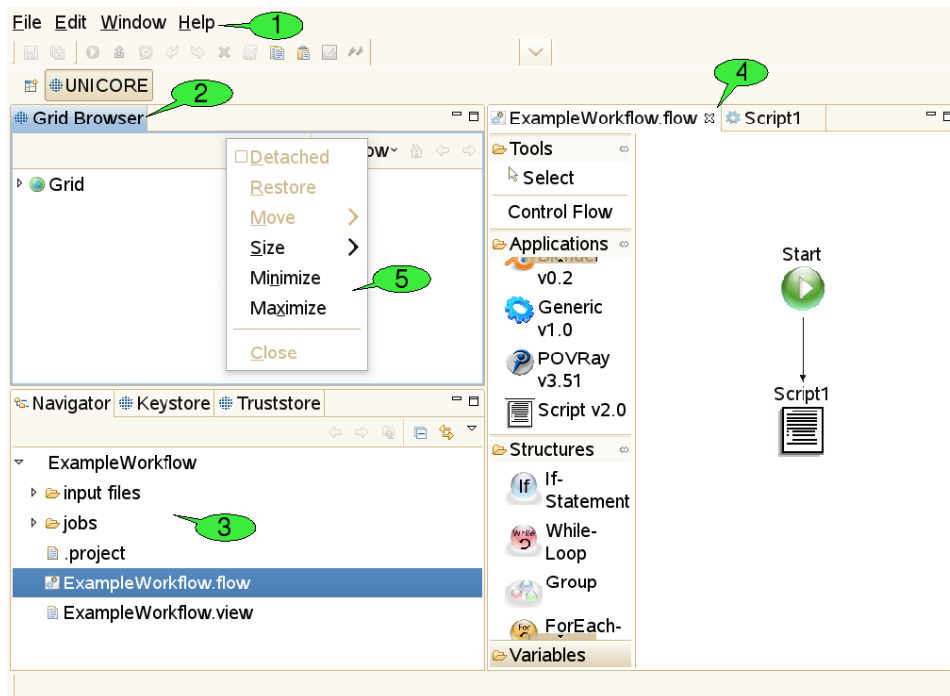


Figure 2. The Eclipse workbench

4.2.1 Menu bar and tool bar

At the top of the workbench, there is a menu bar from which different pull down menus containing ‘global’ actions can be opened ¹. For convenience, some actions are available via shortcuts from the tool bar just below the menu bar. The items in the tool bar can change depending on the selection of objects in the client, mirroring the fact that different actions can be performed on different objects.

4.2.2 Views

Resizable and draggable tab panels containing buttons and other controls are an integral part of all Eclipse based clients. These panels are called *views* ². Apart from being resized and moved, they can also be closed and re-opened. Detaching a view from the workbench will embed the view in its own window. Double-clicking its title will maximise it and double-clicking the title once more will restore its original size. Some views are ‘single-tons’, so only one instance of the view can be opened, whereas other views can be opened multiple times, showing a different content in each instance.

4.2.3 The workspace

The workspace is a directory, usually located on the local hard drive ³. It is supposed to hold all relevant user data needed for the daily work with an Eclipse-based client. Inside the workspace, the user data is organised in subfolders, so-called *projects*. All files within a project should be thematically related. In the UNICORE Rich Client, each job description file (with the extension ‘.job’) and each workflow description file (‘.flow’ file) is stored in its own project, together with its input files. Having a separate project for each job or workflow has the following advantages:

1. Jobs and workflows can get complex. They may need a large number of input files that might be organised in their own directory structure. Mixing up multiple jobs or workflows in a single project can therefore lead to mixing up input and/or output files.
2. Eclipse has its own notion of importing and exporting projects. This provides a nice mechanism for exporting jobs and workflows (e.g. to a single zipped file that contains all necessary input data) and sharing it with co-workers. In the UNICORE Rich Client, job input files should be put into a directory called ‘input files’ inside the project. Relative paths can then be interpreted relative to this directory, which makes sharing of projects very easy.

Apart from the data that are relevant to the user, the workspace also contains metadata that are used in order to manage user preferences and store the state of the Eclipse workbench. In the Eclipse framework, there are different views for displaying the content of the workspace. The most widely used view is called the *Navigator* view. It represents the workspace as a file tree and is very similar to most graphical file browsers. It can be used for creating, renaming, copying, and deleting projects, files and directories. Projects can also be ‘closed’ if unneeded. This will hide their content from the Navigator view.

4.2.4 Editors

When a file is supposed to be opened (e.g. after double clicking it in the *Navigator* view, Eclipse tries to identify a suitable editor by looking at the file's extension. If an associated editor can be found, it is invoked and will display the file content. For example, '.txt' files invoke a text editor, the '.flow' extension invokes the workflow editor 4. File types can also cause associated external applications to be started; for example, a web browser for '.html' files. If the filetype is not supported, an error message is displayed. Associations between file types and editors are defined in the preference page that can be reached via *Window* → *Preferences* → *General* → *Editors* → *File Associations*.

4.2.5 Context menus

Many functions in the client are available via context menus 5. In order to open a context menu, right click an object or a view. The items available in the context menu are different, depending on the object on which the context menu was opened.

4.2.6 Perspectives

The outer appearance of the workbench is very flexible and can change a lot over time. The user benefits from being able to hide information he does not want to see at the moment and arrange the remaining components in a way that fits his needs best. However, less experienced users may have to search for information they accidentally hid in the first place. In order to deal with this problem, the Eclipse framework has introduced the notion of *perspectives*. A perspective is a well defined arrangement of views and editors in the workbench. In addition to determining which components are visible in which spots, it can also influence the actions that can be performed from the tool bar of the workbench. A given arrangement can be saved as a perspective for later re-use and a user can always restore the original appearance of a perspective by resetting the perspective.

4.3 Basic security configuration

4.3.1 How does encryption with X.509 certificates work?

Most security mechanisms on a UNICORE Grid are based on X.509 certificates. For each X.509 certificate, there is a pair of cryptographic keys, that fit each other. These keys can be used to encrypt and decrypt messages: whatever has been encrypted with one of the keys can only be decrypted with the other key - but the keys are not equal. This is why this type of encryption is called 'asymmetric'. Such an asymmetric pair of keys can be used in a public key infrastructure (PKI): The trick is that one of the two keys, called the 'public' key is published and therefore open to everyone, whereas the other key - called the 'private' key - is kept secret by the owner of the key pair. In order to be able to keep the private key secret, it must be very difficult to reconstruct or guess the private key by looking at the public key.

Everyone can use the public key to encrypt messages that only the owner of the private key can read. And, equally important, the owner of the private key can prove that he owns the private key by encrypting a meaningful message with it: everyone can use the public

key to decrypt the message and make sure that it is meaningful, but only the owner of the private key can produce the encrypted message. Asymmetric encryption can also be used for digitally signing documents. With a digital signature, a person can prove that he really is the author of a document, or that he approves the content of a document. The most common way of creating digital signatures comprises two steps: first, a checksum for the document to be signed is computed. The checksum is a relatively short sequence of characters (compared to the document). It is computed by applying a well-known checksum function that always generates the same checksum as long as the content of the document is unchanged. Second, the checksum is encrypted with a private key. The encrypted checksum is published together with the document and forms the digital signature. A reader of the document can use it for checking whether the document was changed. To this end, he applies the same checksum function to the document and compares the result to the checksum that he obtains by decrypting the digital signature (using the public key).

In order to obtain an X.509 certificate from a key pair, the public key is stored in a document, together with some information about the certificate's owner-to-be (e.g. name, email address, organisation). This document is then digitally signed with the private key of a certificate authority (CA), which means that the CA approves the creation of the certificate. This process is called 'issuing a certificate'. Everyone can use the CA's public key to check, whether the certificate has been signed by the CA.

4.3.2 How does UNICORE use X.509 certificates?

With X.509 certificates, UNICORE ensures two things: First, each client or server on the Grid can attest that he is who he claims to be. He does so by presenting his certificate - which contains the public key - and providing evidence that he knows the private key belonging to this public key (by encrypting a previously defined message). Since private keys are kept secret, he must be the owner of the certificate. Second, the public key is used to encrypt messages that only the person knowing the private key (the owner of the certificate) can read. This way an encrypted communication channel between different actors on the Grid is established (by secretly sending a newly created key that can be used for both encryption and decryption of additional messages). The protocol defining the details of establishing the encrypted channel is called Transport Layer Security (TLS), a successor of the Secure Sockets Layer (SSL).

4.3.3 What does this mean to the user?

Before accessing a UNICORE based Grid, each user needs to obtain a valid X.509 certificate which is issued by one of the certificate authorities (CAs) that the UNICORE servers trust. The client presents this certificate to the server whenever he is asked for authentication. The server then checks whether it trusts the CA that issued the certificate. It does so by searching for the CA's certificate in a so-called 'truststore' i.e. a file that contains a list of trusted CAs' certificates. If the CA's certificate is found, it knows it can trust the client. Analogously, the client checks whether it trusts the server. If both checks are successful, a communication channel is created.

All private keys for certificates that the user may want to use on the Grid are stored in a special file called 'keystore'. The keystore is encrypted and secured by a passphrase that

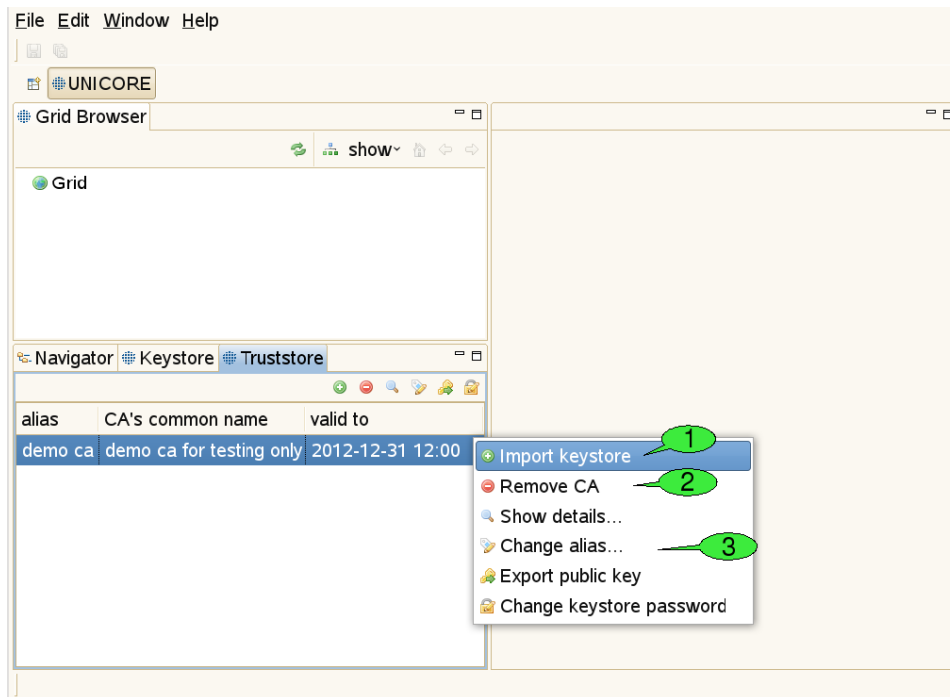


Figure 3. The Truststore view

the user has to remember. During first startup, the Rich Client can create a new keystore file. It is also possible to reuse an existing keystore file. For simplicity, there is only one file that contains both truststore and keystore, so the list of trusted CAs is written to the same encrypted file that holds the private keys.

4.3.4 The Truststore view

Use this view to add certificates of trusted certificate authorities (CAs) to the truststore (see Figure 3). This is necessary in order to communicate with secure Grid services via an SSL encrypted channel. Failing to add the required certificates for the Grid infrastructure that you would like to use will result in errors when trying to contact any of the Grid services.

For each CA certificate contained in your keystore/truststore file, the truststore view displays the alias identifying the certificate (must be unique), the name of the CA, and the end of the certificate's validity period.

In order to add trusted CA certificates, import a file containing these certificates (the file extension should be one of '.jks', '.p12', or '.pem') 1. Certificates can also be removed from the truststore 2. Additional actions allow for opening a detailed certificate description, changing a certificate's alias (used aliases must be unique) exporting public keys to '.pem' files and setting the keystore password 3.

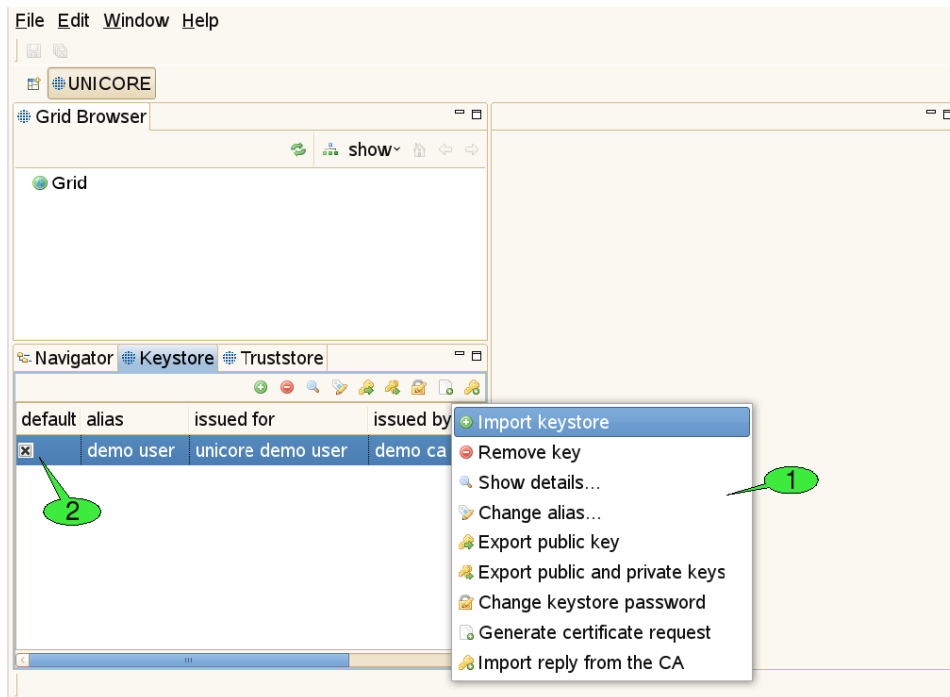


Figure 4. The Keystore view

4.3.5 The Keystore view

This view is used to manage private keys and the associated X.509 user certificates (Figure 4). Different actions may be performed via the view's context menu *1*. The first item is used to import all private and public keys from an existing keystore file into the client's keystore. The second item can permanently delete private keys from the client's keystore. Additional items allow for displaying more details about a selected key, changing the alias that identifies the selected private key, exporting the certificate that belongs to the selected private key, exporting a number of private and public keys to an external keystore file and modifying the client keystore's passphrase. In order to obtain a valid certificate from an existing CA, a certificate request can be created. For each request, a pair of private and public keys is generated. The private key is saved in the keystore. The certificate request must be sent to the administrator(s) of a CA. The response to such a request is usually a '.pem' file, containing the certificate, now signed by the CA. By importing this file into the keystore (using the last item in the context menu), the private key associated to the certificate becomes functional. If the keystore contains multiple user certificates, a default certificate for accessing Grid services should be set *2*.

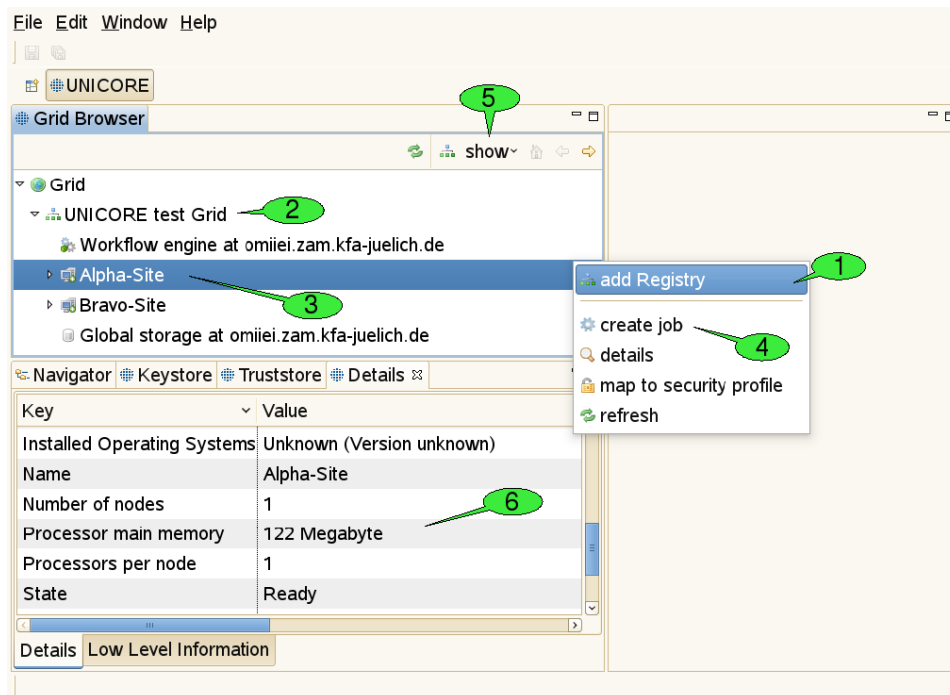


Figure 5. The Grid Browser and Details views

4.4 Browsing and monitoring the Grid

4.4.1 The Grid Browser view

This view represents the Grid as a tree structure (see Figure 5, top left). The items that form the tree are called ‘nodes’ and represent Grid services and files.

There are numerous actions that can be performed on this view or its nodes:

1. Adding registries: For getting started, open the context menu (by right-clicking inside the Grid Browser) and select *add Registry* 1. In the appearing dialogue, enter the URL of a registry that serves as an entry point to the Grid: A registry is used for looking up all available services. For each added registry, a new node should appear just below the root node called *Grid* 2.
2. Refreshing nodes: By double-clicking a node, the represented Grid service is contacted and information about its state is gathered. This is called a *refresh*. After refreshing the registry, a new sub tree should open, displaying the target system and workflow services known by the registry 3. Target system services are used for job execution, workflow services are used for workflow execution.
3. Opening the context menu on a selected node: By right-clicking a node, a context menu that contains all available actions for the associated service will appear. For

instance, users can create job descriptions for job submission to a target system by selecting the *create job* action from the target system's context menu 4.

4. Filtering of Grid services: In large Grids, keeping an overview of the available services and finding relevant information might become difficult. In order to support the user with these tasks, configurable filters can be applied to the Grid Browser. Nodes that do not pass the set of active filters, will not be displayed to the user. The default filter shows job or workflow execution services and storages only. Services that are less frequently used can be revealed by using the *show* menu to the top of the Grid Browser view 5 and selecting *All services*. Additional filters allow to search for services of a specific type, display jobs and workflows that yield a particular state, or have been submitted within a given period of time. A file search filter can be used to retrieve all files that match a certain file name pattern.

Although the Grid Browser displays the Grid as a tree, the actual topology of the Grid can only be modelled with a graph. The Grid Browser deals with this situation by depicting a single Grid service with multiple nodes. For instance, a job that is part of a workflow will be represented by two different nodes in the Grid Browser: one beneath the target system service that executed the job and the other one beneath the workflow management service that corresponds to the job's parent workflow. These two nodes, however, share the same data model: whenever you refresh one of the nodes, the other one is being refreshed at the same time.

4.4.2 Grid Files

Remote files in UNICORE based Grids are accessible through UNICORE storages that can be searched directly in the Grid Browser. Directories and files are displayed as child nodes of the storage node. Double-clicking a directory will open it and list contained files and folders, while double-clicking a file will download that file to the local hard disk and open its content in an associated editor. Saving the file with the associated editor will also update the remote file's content (except when the file is opened with an external editor). Data can be moved between different remote file systems. For instance, you can move a directory from one UNICORE storage to another with a single mouse drag. Files can also be uploaded to remote storages by dragging them from the workspace, a local file browser or the desktop. Due to a limitation of the Eclipse framework, files can only be downloaded to the workspace (via the Navigator view).

4.4.3 The Details view

When a node in the Grid Browser has been refreshed for the first time, information about the associated service is shown in the *Details* view 6. For target system services, this includes available resources like number of CPUs, amount of main memory, and a list of installed applications. For jobs and workflows, states and submission times are displayed, for Grid files, sizes and modification dates. Note, that this view is connected to the Grid Browser: Whenever a different node is selected, the *Details* view is being updated to display its details.

4.5 Job submission and visualisation of job outcomes

4.5.1 The Job editor

The UNICORE Rich Client offers graphical editors for setting up job descriptions. Instead of having to edit text-based job descriptions, the user is provided high level interfaces which are tailored to the applications he wants to execute on remote systems. The client is easily extensible with new application specific user interfaces as new applications are introduced to the Grid environment. Setting up a job description only requires a few simple steps and can be performed within a couple of seconds. The first step is the creation of a job project.

4.5.2 Creating a job project

There are different ways to create a new job project:

1. Select *File* → *New* → *Job Project* from the menu bar (see 1 in Figure 6).
2. Open the context menu of the *Navigator* view and select *New* → *Job Project*.
3. Use the *create job* item from the context menu of a target system node.
4. Choose the *restore Job description* item from a job's context menu in the Grid Browser.

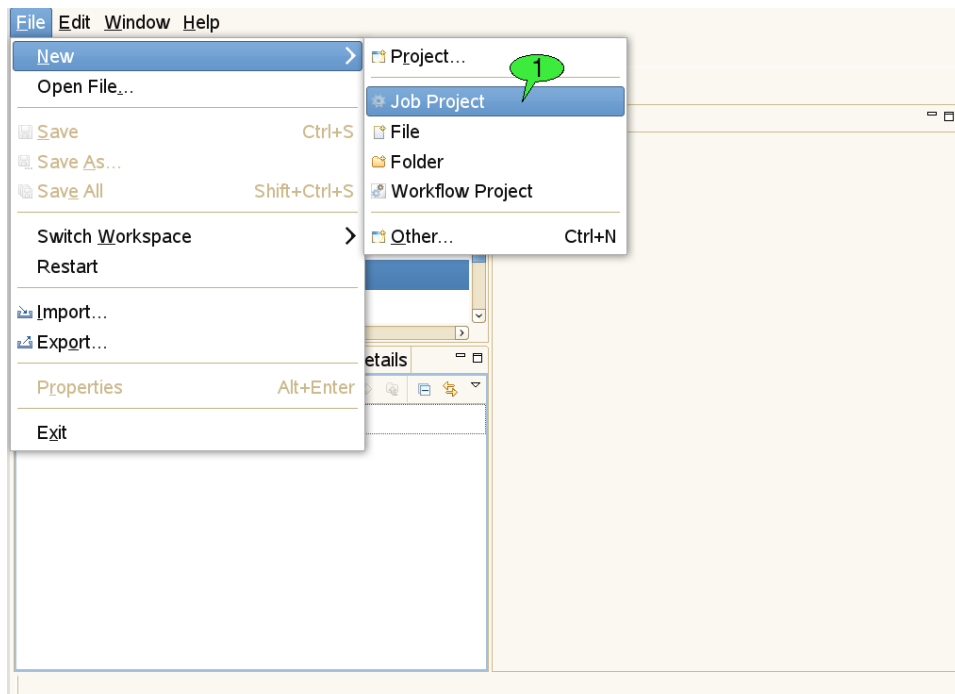


Figure 6. Creating a job or workflow project

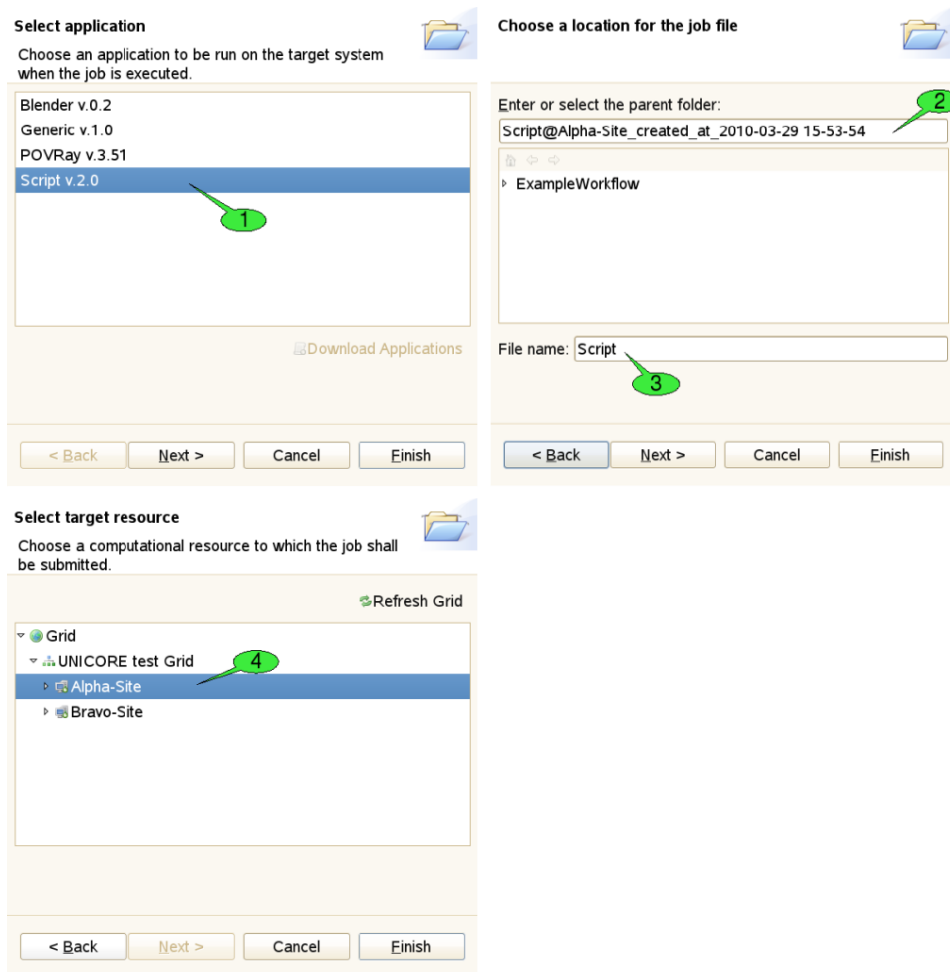


Figure 7. Wizard for creating a job project

The first three of these options will pop up a series of wizard dialogs which will guide the user through the creation of the job project (see Figure 7).

The first step of the wizard is used to choose an application to be run on the target system. In our example, we would like to execute a simple shell script. Therefore, we have selected the *Script* application 1. By pressing the *Finish* button the new job project is created. Click *Next* which will take you to the next wizard step. Here, a different name for the project 2 and the job file 3 can be set. The third wizard page allows for selecting a different target system for job submission 4. The selected target system can also be modified after the project has been created. When the job is created with the last option, both the target system selection and application to be run are restored from the server. Therefore, the job creation wizard shows the second wizard page, only (where you can set names for the project and job file).

4.5.3 Editing mode

The most convenient way to create a job project is using the context menu of a target system node (see 1 in Figure 8), as the corresponding target system will be pre-selected and the job creation wizard can be completed on the first page.

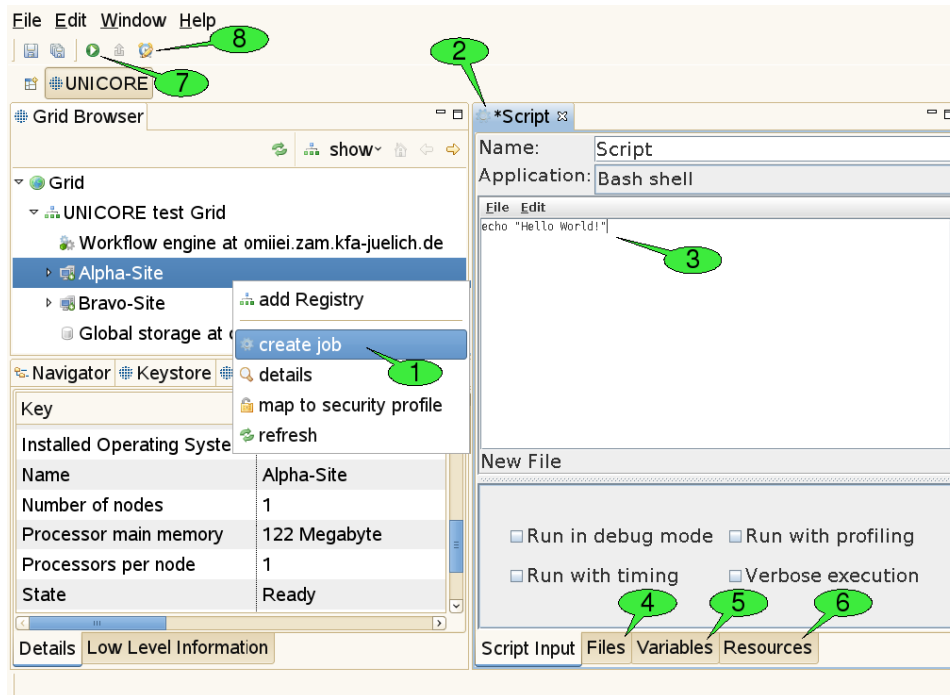


Figure 8. Job editing and submission

Once the job project and job file have been created, a new *job editor* will be opened in editing mode, displaying a graphical user interface (*GUI*) for the application 2. It allows for defining the input parameters of the job to be run. The GUI for the Script application provides an embedded text editor for typing in the shell script 3. New application GUIs can be installed by selecting *Help* → *Software Updates* → *Download Grid Applications* from the workbench's menu bar. This option requires an application GUI server to be available on the Grid (if no server has been found, the option is not available). The job editor holds several tabs. First the application specific tabs are shown for setting parameters in a user friendly way.

In addition, the editor holds three generic panels:

- The *Files* panel 4: This panel can be used to define file imports from remote locations or preceding activities in a workflow. The application specific panels usually only allow for defining imports from the local file system. File exports to remote locations can also be set up here.

- The *Variables* panel 5: This panel can be used to set the application's input parameters directly (circumventing the application specific panels that usually operate on these parameters, too). All parameters are passed to the application via environment variables. Furthermore, the panel allows for setting up additional environment variables for your application run.
- The *Resources* panel 6: This panel can be used for specifying resource requirements of the job, like the number of CPUs needed for a calculation or the amount of memory. The tree-like view on the Grid to the right serves for changing the selected target system for job execution. Note that the list of suitable target systems is updated when changing resource requirements. Also note that the boundaries for resource requirements change when a different target system is selected. The selection can be undone by choosing a node that is not a computational resource (e.g. the *Grid* node or a registry node).

When all parameters are set, click the green *submit* button (see 7 in Figure 8) to submit the job to the selected target system.

An additional action in the tool bar of the job editor is used to set the job's lifetime. When the job has reached the end of its lifetime, the job resource representing the submitted job is destroyed and its working directory is cleaned up automatically. This implies that the job's outcomes cannot be accessed hereafter. The default lifetime for jobs is set to 720 hours (30 days).

4.5.4 Monitoring Mode

As soon as a job is being submitted, the job file is copied into a newly created subfolder of the 'submitted' folder in the job project. The subfolder's name consists of the String 'submitted at', followed by a timestamp, e.g. '2010-03-29 16-00-34' that indicates when the job was submitted (1 in Figure 9). This way, a history of all submitted versions of the job is kept and the user can later look up old job descriptions and compare the results of the associated job executions. The copied version of the job file is then opened in a new job editor.

In order to inform the user about the execution state of the job, the editor is put into *monitoring mode*. This means that the job description cannot be edited anymore and the title of the job editor indicates the current execution status 2. The status may be one of the four values *submitting*, *running*, *finished*, and *failed*. If the job editor is closed in state *submitting* the job submission cannot be performed successfully and the subfolder with the copy of the job file is deleted automatically. If the editor is closed in state 'running', execution of the job will continue normally on the server side. By double-clicking the job file copy in the *Navigator* view, the job editor will be re-opened in monitoring mode and continue to watch the job execution. Jobs can be aborted by selecting the 'abort' item in their context menu. Aborting a job will interrupt the execution of the associated application as soon as possible (this depends on the target system's ability to abort application runs), but leave the job node (and its working directory node) accessible in the Grid Browser. In contrast, destroying a job will first abort the job and then clean up all used resources including the job's working directory.

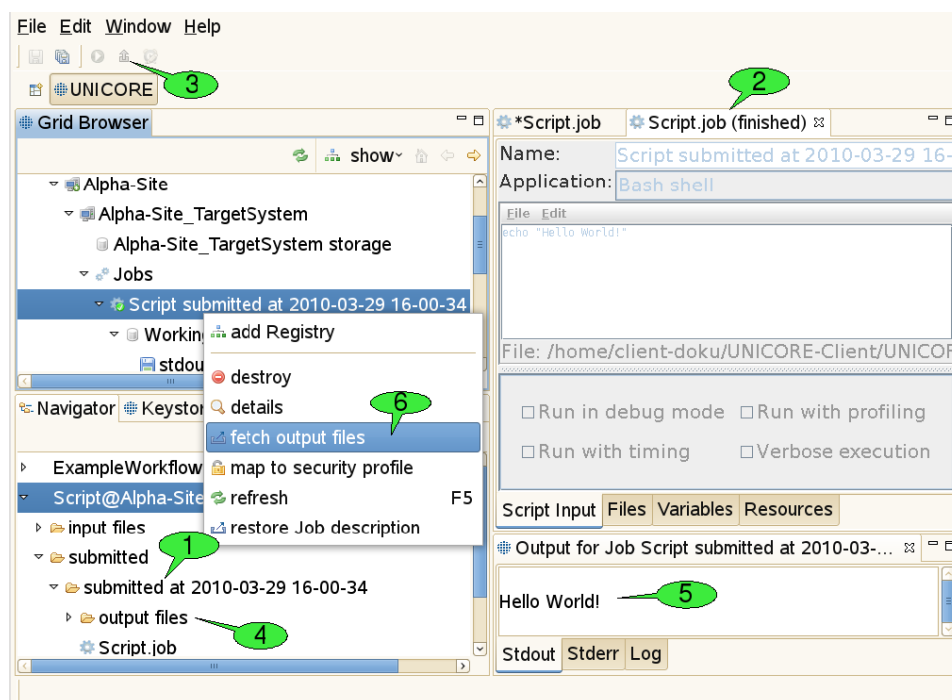


Figure 9. Job monitoring and fetching job outcomes

4.5.5 Fetching Job Outcomes

Once the job has finished successfully, the *fetch output files* action becomes available in the tool bar of the job editor in monitoring mode 3. After clicking it, a dialog will appear that shows all produced output files and allows you to deselect files you do not want to download. After clicking *OK* the selected files are downloaded to the ‘output files’ directory in the subfolder that contains the copy of the submitted job 4. Finally, a new application specific *Job Outcome* view will appear showing the contents of the job’s output files 5. In our example a simple text editor shows the output of the script, but more advanced visualisation software is used for displaying the results of scientific applications (e.g. 3D molecule visualisations for chemical applications). Alternatively, job outcomes can be fetched by selecting *fetch output files* from the context menu of job nodes in the *Grid Browser* view 6.

4.6 The Workflow editor

This software component provides a graphical editing tool for workflows, offering features like copy & paste, undoing changes, performing automatic graph layouts, zooming, and printing of diagrams. Each workflow is created in its own project and can be submitted and monitored like a single job.

4.6.1 Creating a workflow project

In order to create a new workflow project, either select *File* → *New* → *Workflow Project* from the workbench's menu bar or select *New* → *Workflow Project* from the context menu of the Navigator view (see Figure 6). After providing a valid name for both the parent folder and the workflow file, the project is created.

4.6.2 Editing mode

When creating a new workflow project or opening an existing workflow file, a new workflow editor instance is opened for setting up the workflow description (see Figure 10).

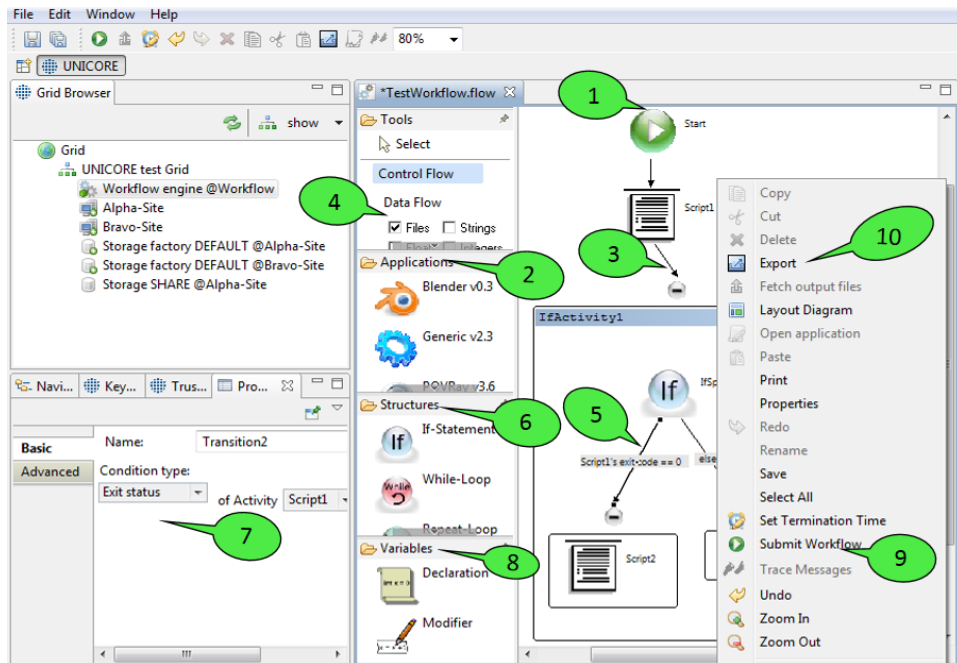


Figure 10. The workflow editor: editing mode

Workflow descriptions are graphs consisting of nodes (commonly called activities in workflow terminology) and edges (called transitions). When a workflow diagram is created, it only displays a single activity: the starting activity of the workflow *I*. Execution of the workflow begins at this activity. In order to add new elements to the workflow, select them from the palette on the left hand side and click in the diagram where you want to place them. Currently, the palette offers the following elements that can be added to the workflow:

1. Application activities These activities represent jobs that are submitted to target systems during workflow execution in order to run specific applications there. For each

application GUI that is installed in the client platform, the palette shows a small icon and the name of the application 2. By selecting an icon and left-clicking a free spot within the workflow editor, a new activity for the associated application will be created. This leads to the creation of a job file in the 'jobs' directory of the workflow project as soon as the workflow is saved. When being double-clicked, application activities will open the job editor for the associated job file. The editor can be used in order to change the job description. When a job is embedded in a workflow, there are several additional possibilities for specifying the job's inputs and outputs that are not available for single jobs:

- Additions to the *Files* panel: A File can now be exported as a *Workflow file* meaning that the file will be stored on some global storage and will be available to subsequent workflow activities.
 - Additions to the *Variables* panel: This panel can be used to set the application's input parameters to the values of workflow variables. Workflow variables can be declared by special activities and modified while the workflow is executed. Their current value during workflow execution is maintained by the workflow engine and may be fed into a job's description before the job is submitted. This mechanism allows for running the same job multiple times, with different parameter values e.g. for performing parameter sweeps.
 - Additions to the *Resources* panel: Workflow jobs do not require users to select a single target system for job execution. This is due to the fact that the workflow engine has a resource broker which is capable of distributing jobs to suitable target systems. In this process, specified resource requirements of the job (e.g. amount of memory) are compared to the target systems' offerings for finding a computing resource that fulfils the requirements. This is generally referred to as 'match-making'. In order to narrow down the choice of target systems used for match-making, the user may select one or more target systems as 'candidate resources' for the job. Again, the selection can be undone by choosing a node that is not a computational resource (e.g. the *Grid* node or a registry node).
2. Transitions Transitions represent the flow of control that passes from one activity to the next. Currently, there are two types of transitions: unconditional 3 and conditional 5 ones. Only unconditional transitions can be added to the workflow manually. Conditional transitions are used in If-statements and While-loops and are added automatically. The reason for this is that conditional transitions may require a different joining behaviour: the default joining behaviour when an activity has multiple incoming transitions is called 'synchronisation'. This means that the activity is only processed when all incoming transitions have been processed. As you might imagine, this behaviour is no longer appropriate when conditional transitions are used: the activity that joins the if and else branches of an If-statement would never be processed if it waited for both branches to finish. In order to hide this complexity from users that are unfamiliar with workflow processing and programming languages, If-statements and similar constructs will be modelled as sub-workflows that automatically define the appropriate joining behaviour.
- In addition to the *Control Flow* view *Data Flow* view can also be selected to

visualize the input and output of created workflow jobs. You can check option *Files 4* to have a graphical view of the input and output files of the activities. Moreover the output of one job can be used by other jobs by simply connecting (drag and drop) it to the respective input files of other jobs. You can check various workflow variable types (*Strings*, *Float* and *Integer*) to visualize the input parameters of the workflow jobs.

3. Workflow structures Workflow structures are subgraphs that bring their own semantics on how to process their child nodes. Currently, four workflow structures are provided: groups, If-statements, While-loops, and ForEach-Loops 6.
 - (a) Groups are the simplest of all subgraphs. They are just containers for other activities. Their content may be hidden by clicking the small minus symbol at their top.
 - (b) If-statements influence the flow of control and contain two additional subgraphs (which are modelled as groups): the if-branch and the else-branch. The if-branch is processed when a certain user-defined condition holds. If the condition evaluates to false the else-branch is processed instead. Both branches can contain arbitrary activities and transitions, thus permitting nesting of workflow structures. Conditions can be altered by double clicking the conditional transition. This will open up the *Properties* view which displays relevant properties of workflow elements 7. Most properties can be modified through this view. There are currently four types of conditions: the first type compares the exit status of an application to a value provided by the user, the second one tests whether an output file with a given name has been created by an application activity, the third one compares the value of a workflow variable to a given value, and the last one checks whether the current time lies before or after a given point in time.
 - (c) The While-loop provides a single subgraph called the loop-body that can be processed multiple times (as long as the loop's condition holds true). The While-loop declares a workflow variable that reflects the current number of loop iterations, the so-called loop 'iterator'. It also declares a variable modifier that increments the loop iterator. The variable declaration can be changed in the *Properties* view of the red activity at the top of the loop and the variable modifier can be set up in the *Properties* view of the associated modifier activity (at the bottom of the while-loop).
 - (d) The Repeat-Until-loop works just like the while loop, but its loop-body is always processed once before the condition is evaluated for the first time. Also, compared to the while-loop, the condition is negated, i.e. the loop ends when the condition becomes true.
 - (e) ForEach-loops can be used in order to create many similar jobs without having to set up each job individually. They have two different modes of operation. The first mode will iterate over a set of workflow variable values and run the job(s) contained in the loop body once for each value in the set. The workflow variable values can be used as input parameters for these jobs. Complex parameter sweeps are possible, as multiple workflow variables can be swept at the same time. The second mode is used to iterate over a set of files. The file set

may consist of any combination of local, remote or workflow files. This mode provides a convenient way to process many different files simultaneously. The operational mode and the parameters to the selected mode can be modified in the Properties view of the orange activity at the top of the ForEach-loop. The iterations of the ForEach-loop are usually executed in parallel. However, there is an upper bound of parallel iterations which results from the workflow engine's capabilities. There is also a way to lower this boundary by providing an Integer value for the *Number of parallel tasks* in the Properties view of the ForEach activity. Setting this value to '1' will lead to sequential execution of the loop iterations.

4. Variable declarations and modifiers 8 Additional workflow variables can be declared using the appropriate *Declaration* activity. The Properties view of this activity allows for (re-)naming the variable and assigning it a type (e.g. *String* or *Integer*) and initial value. A *Modifier* activity can be used to change the value of a workflow variable later.

When the user is pleased with the workflow description, the workflow can be submitted via the editor's context menu 9 or the workbench's tool bar. It can also be exported to an XML based workflow language that the workflow engine understands 10. The exported workflow can later be submitted to the workflow engine by the UNICORE commandline client. This feature is useful e.g. in order to make predefined workflows available via a web interface (the Chemomomentum web portal solution uses the commandline client for workflow submission).

4.6.3 Monitoring mode

The workflow editor is also used for monitoring the execution of workflows, so the basic graphical representation of a workflow stays the same before and after submission to the workflow engine (see Figure 11). This helps in identifying which part of the workflow is being executed at a given point in time.

When a workflow has been submitted, a new folder is created in the 'submitted' sub-folder of the workflow project. This folder contains a copy of the workflow file that is automatically opened in a new workflow editor panel - in monitoring mode. In this mode, the editor disallows any changes to the workflow. It displays the progress of workflow execution by adding small icons to the nodes of the workflow graph that symbolise the execution state of these parts 1.

Outcomes of jobs can be fetched as soon as the jobs have finished. This function is available via the context menu of application activities and the *fetch output files* action in the global tool bar (after selecting the activity for which to fetch outcomes). Job outcomes are downloaded to the 'output files' folder again, so they can easily be found later and associated with the workflow by which they were produced. Monitoring a workflow can be interrupted by simply closing the editor panel 2. By double-clicking the file that represents the submitted workflow (in the *Navigator* view), the editor panel will be re-opened and continue to monitor the execution of the workflow.

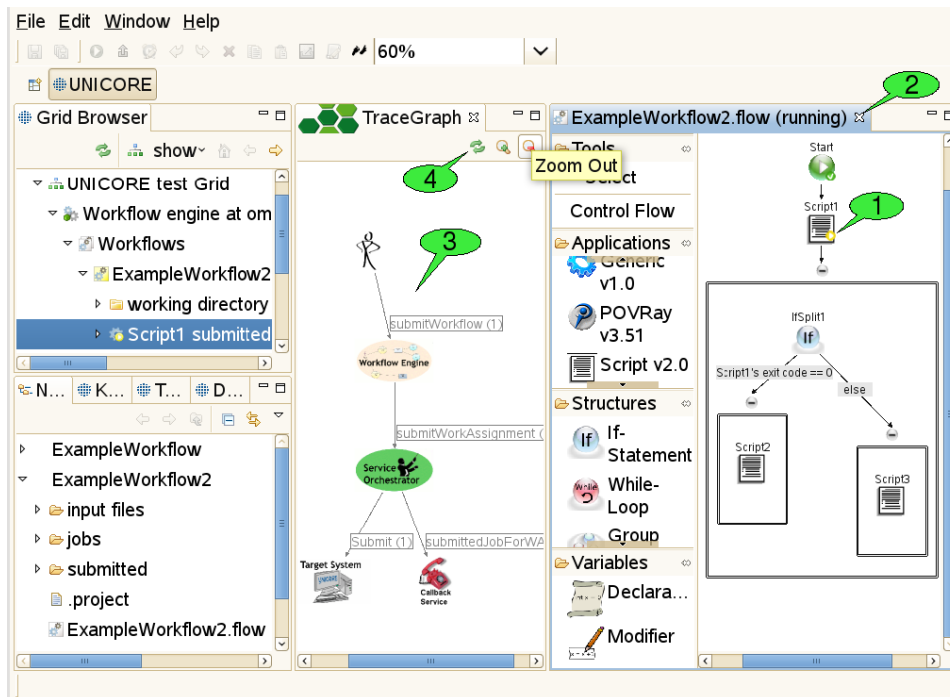


Figure 11. The workflow editor: monitoring mode and the Trace Graph view

4.6.4 The Trace Graph view

In addition to monitoring the execution states of activities in the workflow, the user may trace the workflow for finding out where his jobs were submitted. This action is available via the context menu of the workflow editor. A trace graph will open, showing all messages that were sent by the workflow system during the execution of the workflow 3. By hovering the mouse over a node or edge in the trace graph, additional information about the element is displayed in a tooltip. The set of traced messages can be updated by clicking the *Refresh* button in the tool bar of the *Trace Graph* view 4. Additional buttons allow to zoom in and out (zooming can also be achieved by rotating the mouse wheel while pressing the 'control' key).

4.7 Interactive site access

The UNICORE Rich Client features a Terminal view which can be used to log on to remote hosts via SSH and GSISSH. It complies to the VT100 standard for terminal emulation and can hold multiple terminal sessions (in multiple tabs). Sessions can be created via the *open terminal* action from the context menu of a target system node. Please note, that this action is only available, if the administrator of the UNICORE site has enabled interactive access and provided necessary information about the target system, i.e. the host name and port that should be used for establishing the interactive connection and the available connection

methods. Currently, the UNICORE Rich Client provides two different secure connection methods, Plain SSH and GSISSH. Apart from that, additional protocols can be used in the future — both UNICORE client and server are extensible in this regard.

4.7.1 Plain SSH

When connecting to an SSH server via plain (i.e. conventional) SSH, the user can choose between three different authentication methods:

- Password
- Keyboard-Interactive
- Public-key: If the user's private key path wasn't specified before, the UNICORE Rich Client tries to find the key in the appropriate default directory (e.g. `~/ssh/id_dsa`). If this fails the user is prompted to specify the path.

4.7.2 GSISSH

The GSISSH connection method provides access to GSISSH servers via RFC-3820 compliant proxy certificates. The proxy is created from the keystore of the UNICORE Client when the user starts to connect to the server. It can be stored on the local machine if required. It is possible to choose between different aliases representing different keys in the keystore, different delegation types, and different proxy types. Furthermore, the lifetime of the proxy certificate can be set (the default is 12 hours). When connecting to GSISSH servers the UNICORE Rich Client converts the PEM formatted CA certificates in the UNICORE client's truststore to GSISSH-conform certificates, and stores them on the local machine. By default these files are created in the `~/globus/certificates` folder. However, this can be changed in the client's preferences at *UNICORE* → *Terminal* → *GSISSH*.

4.7.3 How to open a terminal

There are different ways to open a terminal shell for a target site. The most convenient method is to right click a UNICORE target site in the Grid Browser view and select the *Open Terminal* menu item (see 1 in Figure 12).

If the user connects to a site for the first time, he will be prompted to choose one of the available connection methods, in case the administrator of the UNICORE site has provided the necessary information in the IDB. If no connection information is provided, the user can enter the hostname, port, and login name manually. When all required connection data has been gathered, the secure connection process is triggered and the terminal view should show up automatically (see 2 in Figure 12). Alternatively, the user can open terminal shells by using the Terminal Config view.

4.7.4 How to maintain and configure connection information

A second view, the so-called *Terminal Connection Config* view (see 3 in Figure 12), can be used for modifying the user's settings for interactive access to different target systems.

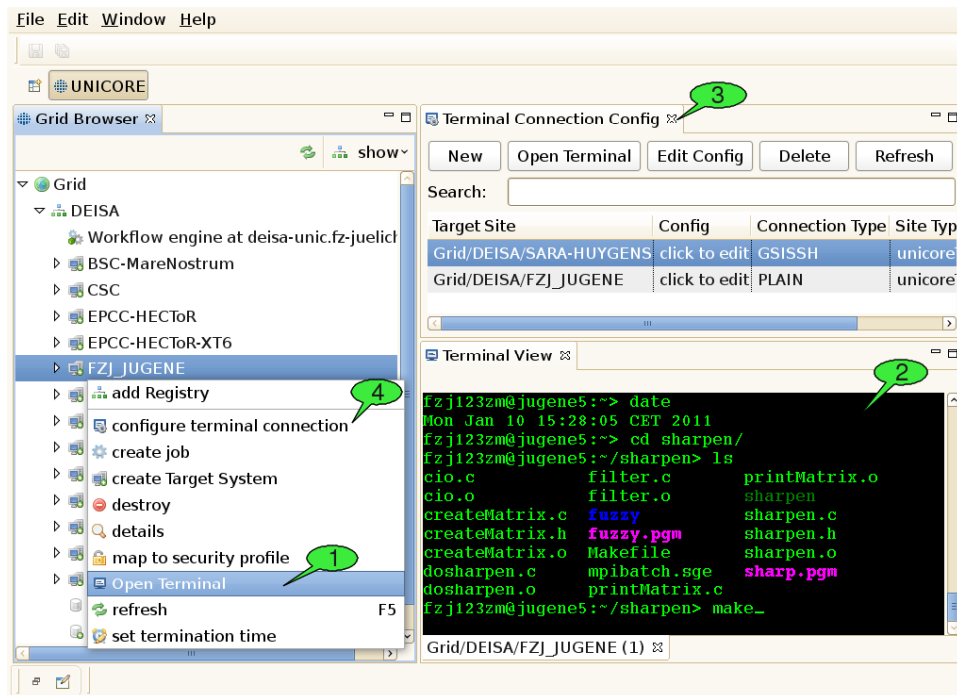


Figure 12. Opening and using terminal connections.

The Terminal Config View can be invoked by selecting the *configure terminal connection* menu item (see 4 in Figure 12) from the context menu of a target site in the Grid Browser.

The view provides a table with all SSH target sites that have previously been invoked by the user. The user can rename the target site, or set a default connection type in the table. To edit the site's terminal configuration the *Config* column or the Edit-button in the top menu can be clicked. This action will open a dialog for editing the parameters of different connection methods. The values are stored permanently in the UNICORE client after clicking *Ok*.

Custom target sites can be created by clicking the *New* button in the top menu. Terminals to such sites can only be opened from the *Terminal Config* view. They can be recognised by the *CustomTargetSite* tag in the Site Type column.

5 Concluding Remarks and Further Information

5.1 Future prospects

Users might have special requirements to UNICORE concerning a special application that has to be run on the target system. Therefore the UNICORE team introduced the GridBean concept <http://www.unicore.eu/documentation/manuals/unicore6/files/GridbeanDevelopersGuide.pdf>. The GridBean developer

guide shows how to implement individual GridBeans and so how to get maximum flexibility.

UNICORE is an open source software that will be enhanced with new features permanently. For instance the UNICORE team is developing the UNICORE portal. This web-based portal will enhance the already existing clients and will be an additional UNICORE client running in any webbrowser. Users will be able to create their jobs, submit them to the target system, get the output, transfer files, etc via a web browser without having to install any software on the local computer. A first stable version of the portal is planned for the end of 2013.

5.2 Documentation

Enhanced UNICORE documentation (manuals, video tutorials, papers, talks, etc.) and all UNICORE downloads are available at www.unicore.eu.

5.3 Contact

In case of any questions please refer to unicore-info@fz-juelich.de.

6 Glossary

CA Certification Authority: An entity which issues digital certificates for use by other parties. CAs are characteristic of many public key infrastructure (PKI) schemes.

GUI Graphical User interface: A set of visual controls that steer a computer program. In contrast to a command line interface, it usually requires less typing because most actions can be performed via mouse clicks.

HTTP Hypertext Transfer Protocol: A communications protocol. Its use for retrieving interlinked text documents (hypertext) led to the establishment of the World Wide Web.

JRE Java Runtime Environment: A set of computer programs and data structures which use a virtual machine model for the execution of JAVA programs.

OGSA Open Grid Services Architecture: An architecture of interacting services. It was described in the paper “The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration” and combines ideas and technologies from both Web- and Grid Services to provide a basis for service oriented Grid architectures (see <http://www.globus.org/alliance/publications/papers/ogsa.pdf>).

SSL Secure Sockets Layer: A widespread cryptographic protocol for securing connections on the internet. Uses Public key encryption for certificate-based authentication and symmetric cipher-based traffic encryption.

XML Extensible Markup Language: A text format derived from the Standard Generalized Markup Language (ISO 8879, see <http://www.iso.org>). XML is used

to exchange data on the Web and it is the basis for a variety of languages and protocols (<http://www.w3.org/XML/>).

SOA Service Oriented Architecture: A software architecture that defines the use of software services to support the requirements of business processes and users on a computer network. The underlying paradigm emphasizes the definition of slim and platformindependent communication interfaces in order to achieve loose coupling. The SOA Reference Model provided by the OASIS Committee Specification, can be found at <http://www.oasis-open.org/>.

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. Gompfer, D. Marx (2012), vi,
540 pages
ISBN: 978-3-89336-768-9
URN: urn:nbn:de:0001-2012020208

11. **Periodic Boundary Conditions and the Error-Controlled
Fast Multipole Method**
by I. Kabadshow (2012), v, 126 pages
ISBN: 978-3-89336-770-2
URN: urn:nbn:de:0001-2012020810

12. **Capturing Parallel Performance Dynamics**
by Z. P. Szebenyi (2012), xxi, 192 pages
ISBN: 978-3-89336-798-6
URN: urn:nbn:de:0001-2012062204

13. **Validated force-based modeling of pedestrian dynamics**
by M. Chraïbi (2012), xiv, 112 pages
ISBN: 978-3-89336-799-3
URN: urn:nbn:de:0001-2012062608

14. **Pedestrian fundamental diagrams:
Comparative analysis of experiments in different geometries**
by J. Zhang (2012), xiii, 103 pages
ISBN: 978-3-89336-825-9
URN: urn:nbn:de:0001-2012102405

15. **UNICORE Summit 2012**
Proceedings, 30 - 31 May 2012 | Dresden, Germany
edited by V. Huber, R. Müller-Pfefferkorn, M. Romberg (2012), iv, 143 pages
ISBN: 978-3-89336-829-7
URN: urn:nbn:de:0001-2012111202

16. **Design and Applications of an Interoperability Reference Model
for Production e-Science Infrastructures**
by M. Riedel (2013), x, 270 pages
ISBN: 978-3-89336-861-7
URN: urn:nbn:de:0001-2013031903

17. **Route Choice Modelling and Runtime Optimisation
for Simulation of Building Evacuation**
by A. U. Kemloh Wagoum (2013), xviii, 122 pages
ISBN: 978-3-89336-865-5
URN: urn:nbn:de:0001-2013032608
18. **Dynamik von Personenströmen in Sportstadien**
by S. Burghardt (2013), xi, 115 pages
ISBN: 978-3-89336-879-2
URN: urn:nbn:de:0001-2013060504
19. **Multiscale Modelling Methods for Applications in Materials Science**
Lecture Notes, CECAM Tutorial, 16 - 20 September, Jülich
edited by Ivan Kondov, Godehard Sutmann (2013), iv, 319 pages
ISBN: 978-3-89336-899-0
URN: urn:nbn:de:0001-2013090204

Macroscopic effects in complex materials arise from physical phenomena on multiple length and time scales and therefore properties of such materials can be predicted accurately based on properties of the underlying building blocks. The major benefits of multiscale models are a simpler physical interpretation based on the analysis of sub-models as well as an improved computational scaling making the simulation of very large systems feasible.

This book includes the lecture notes of courses conducted at the CECAM tutorial “Multiscale Modelling Methods for Applications in Materials Science” held at the Jülich Supercomputing Centre from 16 to 20 September 2013. Written by recognized experts the lecture notes complement existing university courses with knowledge and experience gained recently in the field of multiscale materials modelling encompassing theoretical understanding and practical implementation of multiscale models to real-life applications. The book addresses graduate students and young researchers, working in the field of computational materials science, and covers general methodology, tools for implementation of the multiscale modelling paradigm, as well as applications of multiscale modeling techniques. Topics include fields such as coarse graining of polymers and biomolecules, and modelling of organic light-emitting diodes, electrochemical energy storage devices (Li-ion batteries and fuel cells) and energy conversion devices (organic electronics and carbon nanodevices).

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.