Computational Methods to Find and Rank MHC-I Restricted Tumor-Associated Antigens to Improve Therapeutic Efficacy and Tolerability of Antigen-based Cancer Immunotherapy

Computergestützten Methoden, um MHC-I restringierte Tumor assoziierte Antigene zu finden und zu bewerten, welche die therapeutische Anwendbarkeit und Tolerierbarkeit von Antigenbasierter Krebs-Immuntherapie verbessern

Der Naturwissenschaftlichen Fakultät
der
Friedrich-Alexander-Universität
Erlangen-Nürnberg
zur
Erlangung des Doktorgrades

Dr. rer. nat.

vorgelegt von

Christopher Daniel Alfred Lischer

Als Dissertation genehmigt
von der Naturwissenschaftlichen Fakultät
der Friedrich-Alexander-Universität Erlangen-Nürnberg

Tag der mündlichen Prüfung:
20.11.2023

Vorsitzender des Promotionsorgans:
Prof. Dr. Wolfgang Achtziger

Gutachter:
Prof. Dr. Vasily Zaburdaev
Prof. Dr. Julio Vera-González

Science is more than a body of knowledge.

It's a way of thinking. A way of skeptically interrogating the universe with a fine understanding of human fallibility.

- Carl Sagan

Table of contents

1 Summary / Zusammenfassung

1.1 Summary

The development of targeted immunotherapies in the last decade has opened novel treatment modalities for many cancer entities. In particular, antigen-based treatment systems have received significant attention. These methods deliver tumor-derived antigens to the patient's immune system intending to stimulate a specific and lasting immune response. Prominent methods in use for the delivery of antigens are antigen-encoding mRNA- or DNA-laden vectors like lipid nanoparticles or adenoviruses, as well as externally matured autologous dendritic cells or externally expanded autologous T cells. During the application of these immunotherapies, several challenges became apparent. First, the discovery of suitable target genes has proven difficult since the antigens need to be restricted to the tumor, i.e., not found or, at most, very lowly expressed in the rest of the body. Secondly, antigen loss from the tumor under pressure by the immune system is a repeated occurrence and must be mitigated to ensure long-lasting therapeutic efficacy. Finally, unavoidable off-target effects must be limited in their severity. This thesis aimed to develop novel computational tools and algorithms to address and overcome the above-presented issues.

In the first part of this project, we created a pipeline based on next-generation sequencing data to select overly expressed genes in a tumor model which are not or only minimally expressed in survival-critical tissues. As a feasibility study, we predicted antigens against metastatic melanoma and found 35 candidate genes. We predicted all possible peptides with a length of 9 to 12 amino acids and their corresponding binding affinity to different HLA Class I alleles. Using a multivariate score, we ranked all derived peptides and their allele-specific epitopes and provided them to the community in an online database. With our algorithm being free of prior knowledge and based only on primary data, we deemed the selection of the well-known metastatic melanoma marker *MAGE-A3* as validation of our approach. In addition, our tolerability evaluation effectively filtered out a known *MAGE-A3*-derived antigen that had caused severe side effects in a clinical trial. In the second part of this project, we set out to improve several aspects of our pipeline. First, to implement a generalizable prediction procedure; second, to evaluate the biological relevance of the antigen for the tumor and third, to perform experimental validation of the efficacy of our candidates. We adapted our prediction system by integrating a machine learning model that evaluates both binding and immunogenic activity probability in a generalized manner. We also developed a network model that tries to gauge an antigen's relevance for the tumor to reduce the chances of antigen loss. We implemented this approach with data derived from metastasized primary uveal melanoma and found a set of 22 candidate genes. Several experiments with autologous T-cells were performed for validation, showing that our predicted peptides elicited an immune response in an *in vitro* setting for some of our healthy donors. Further, a cytotoxicity assay showed that the peptide-stimulated T-cells killed the uveal melanoma cell line 92.1 in an antigen-specific

fashion. Using *in silico* and *in vitro* methods, we strived to discover novel tumor antigens and to provide a decision support system to facilitate applicability.

1.2 Zusammenfassung

Die Entwicklung zielgerichteter Immuntherapien im letzten Jahrzehnt hat neue Behandlungsmöglichkeiten für viele Krebsarten eröffnet. Insbesondere Antigen-basierte Behandlungssysteme haben große Aufmerksamkeit erfahren. Bei diesen Methoden wird dem Patienten ein gegen den Tumor gerichtetes Antigen verabreicht, mit dem Ziel, eine spezifische und dauerhafte Immunantwort zu stimulieren. Prominente Methoden für die Verabreichung von Antigenen sind Antigen-kodierende mRNA- oder DNA-beladene Vektorsysteme wie Lipid-Nanopartikel oder Adenoviren sowie extern maturierte autologe dendritische Zellen oder extern expandierte autologe T-Zellen. Bei der Anwendung dieser Immuntherapien traten mehrere Herausforderungen zutage. Erstens hat sich die Entdeckung geeigneter Zielgene als schwierig erwiesen, da sie auf den Tumor beschränkt sein müssen, also im übrigen Körper nicht oder nur in geringem Maße exprimiert sein dürfen. Zweitens besteht die Möglichkeit, dass Antigene unter dem Druck des Immunsystems auf den Tumor verloren gehen. Dies muss verhindert werden, um eine langanhaltende Wirkung der Therapie zu erzielen. Schließlich kann es zu *Off-Target*-Effekten kommen, deren Schwere unter Kontrolle bleiben muss. Im ersten Teil dieses Projektes haben wir eine bioinformatische Pipeline entwickelt, welche basierend auf *Next-Generation Sequencing* Daten Gene in einem Modelltumor auswählt, die im Tumor überexprimiert sind, aber in überlebenswichtigen gesunden Geweben kaum oder gar nicht exprimiert wurden. Im Rahmen einer Studie haben wir Antigene gegen metastasierte kutane Melanome vorhergesagt und 35 Kandidatengene gefunden. Wir sagten alle möglichen Peptide mit einer Länge von 9 bis 12 Aminosäuren und ihre Bindungsaffinitäten zu verschiedenen HLA-Allelen der Klasse I voraus. Anhand eines mehrdimensionalen Bewertungssystems ordneten wir alle abgeleiteten Peptide und ihre allelspezifischen Epitope und stellten sie in einer Online-Datenbank anderen Wissenschaftlern zur Verfügung. Da unser Algorithmus kein etabliertes Wissen nutzt und nur auf Primärdaten basiert, sahen wir das Auftauchen des bekannten Melanom-Markers *MAGE-A3* als Bestätigung unseres Ansatzes an. Zudem entfernte unsere Verträglichkeitsbewertung ein bekanntes, von *MAGE-A3* abgeleitetes Melanom-Antigen, das in klinischen Studien schwere Nebenwirkungen verursacht hat. Im zweiten Teil dieses Projekts und als Erweiterung dieser Methodik haben wir versucht, mehrere Abschnitte unserer Pipeline zu verbessern. Wir haben deshalb erstens das Voraussagemodell generalisiert, zweitens die biologische Relevanz des Antigens für das Tumorüberleben abgeschätzt und drittens die Wirksamkeit unserer Kandidaten experimentell validiert. Zu diesem Zweck haben wir unser Vorhersagesystem angepasst. Wir implementierten ein auf maschinellem Lernen basierendes Modell, das sowohl die generalisierte Bindung als auch die immunogene Aktivität bewertet. Zudem wurde ein Netzwerkmodell erstellt, um die Relevanz eines Antigens für das Tumorwachstum zu beurteilen und die Wahrscheinlichkeit eines Antigenverlusts zu verringern.

Wir haben dieses neue System mit Daten von primären Aderhautmelanomen getestet, die bereits Metastasen gebildet hatten. Zur Validierung wurden mehrere Experimente mit autologen T-Zellen durchgeführt, die zeigten, dass die von uns vorhergesagten Peptide bei einigen unserer gesunden Spender *in vitro* eine Immunantwort hervorriefen. Darüber hinaus konnte ein Zytotoxizitätstest zeigen, dass die Peptid-stimulierten T-Zellen die Aderhautmelanom-Zelllinie 92.1 antigenspezifisch abtöteten. Insgesamt konnten wir mit Hilfe von *in silico* und *in vitro* Methoden neue Tumorantigene entdecken und eine Methodik zur Entscheidungsfindung bereitzustellen.

2 Introduction

The past decade has seen an increase in therapeutic options for patients suffering from cancer. Novel immunotherapies like the so-called immune checkpoint blockade (ICB) have dramatically increased patients' survival time, especially in metastatic cutaneous melanoma (MCM). Response rates to these novel treatments, however, remain modest. For ICB, long-term response rates range from 10% to 44%, depending on the study and clinical circumstances (Robert *et al.*, 2015; Wolchok *et al.*, 2017). While prognoses have seen remarkable improvement, many patients still have minimal options for treatment once the tumor returns or fails to respond.

Additionally, in other melanocyte-derived malignancies, like the most common cancer of the eye, uveal melanoma (UM), the prognosis is bleak, and response rates are low, especially after the tumor has spread to distant metastases (Wessely *et al.*, 2020). While more options for treatment have been developed over the past years, there is still an urgent need for adjuvant or monotherapies that can open therapeutic routes for patients and doctors. Without a doubt, therapies that actively target tumor-specific antigens through the patient's immune system will play a significant role in filling this gap. With various delivery systems for tumor-associated antigens existing today, like vaccine-based approaches, the task is to identify and validate novel tumor-associated antigens (TAAs) for their efficacy and safety. Every element of the immune system is relevant for the rational design and selection of TAAs for therapy, ranging from phagocytes which take up antigens, to lymphocytes which create lasting immunity and mediate the elimination of the targeted tumor. Hence, in the following, a brief introduction to the mechanisms and cell types of the immune system is provided while focusing on relevant elements for targeted immunotherapy.

2.1 A primer on the human immune system

All higher organisms, especially mammals and by the nature of things, including humans, have evolved complex defense systems to counter pathogens, toxins, or physical damage (Yuan *et al.*, 2014). These systems, comprised of physical, chemical, and cellular elements, are generally referred to as the immune system and are characterized by complex regulatory networks and biological processes that can respond to a challenge from the external environment and internal dysfunction. The primary goal of this system is to distinguish the self, the host's body, from the non-self, a foreign challenge. Under the current understanding, the human immune system is separated into two discrete categories. They are defined as innate immunity and adaptive immunity. Innate immunity includes passive mechanisms and tools generally considered germline-encoded and invariant over time and between individuals. These passive tools contain elements like skin, mucosa, and other anatomical features to prevent or hinder infection, as well as active elements ranging from specific cell types, soluble factors like small molecules or proteins, and molecular sensors. Once activated, the different

components of the innate immunity trigger an acute state of inflammation (Janeway and Medzhitov, 2002; Chaplin, 2010). This complex process can cause a wide range of changes in an organism, from local effects like swelling to system-wide effects like increased body temperature. With the release of small signaling molecules or cytokines and local effects in the cellular environment, the innate immunity provides the link to the activation of adaptive immunity (Chaplin, 2010; Gasteiger *et al.*, 2017). Two core abilities define the adaptive immunity. The first ability is to react specifically to novel antigens through somatic recombination processes whose resulting receptors are not germline-encoded. These receptors are created and improved to combat specific challenges. The second core ability is the generation of memory of encountered challenges or antigens. In its reaction time, the adaptive immunity is comparatively slow at first exposure, usually taking days to weeks to mount a response. In contrast, innate immunity reacts with a far shorter delay, with local effects occurring within minutes of the challenge (Marshall *et al.*, 2018).

The B and T lymphocytes are the two main cell types involved in the adaptive immune response. These two cell types provide the functional backbone for the adaptive response regarding memory and specificity. B cells produce specific antibodies while T cells mediate several immune activities, e.g., reinforcement to B cell activation or directly killing infected or aberrant cells (Bonilla and Oettgen, 2010; Jain and Pasare, 2017). Additionally, while providing this highly specific response, T cells also eradicate dangerous cell populations within the body by patrolling and checking cells for atypical internal products. This feature is made possible by a system called antigen presentation, which allows the display of internal products of the cell to the T cells. Presentation of internal products is crucial in clearing cells that may have been infected with intracellular pathogens like viruses, bacteria, or protozoa or may have suffered cancerous mutations (Stenger *et al.*, 1998; Huster, Stemberger and Busch, 2006; Jhunjhunwala, Hammer and Delamarre, 2021). While the two categories are semantically distinct, innate and adaptive immunity are highly interdependent. For example, many cells of the innate immunity are required in the specific activation of T and B cells. They are thus essential for creating a targeted response and an immunological memory.

2.1.1 Innate Immunity and its cellular components

Evolutionarily speaking, the innate immune system is a highly conserved defense system in which some hallmark elements are shared between all multicellular life (Buchmann, 2014). Innate immunity in humans encompasses many tools, including physical barriers like the skin or mucosa on exposed tissue like the respiratory tract. Its principal task is mounting a quick response with a short reaction time through molecular pattern recognition receptors (PRRs) to situations like infection, tissue damage, or other stresses. The reaction to these effects can be mediated in many several different ways. Firstly, an immediate response can occur through soluble elements secreted by specialized cell types, like antimicrobial peptides, complement factors, cytokines, or chemokines. Cell-dependent effector functions may occur like phagocytosis or cytotoxicity

(Lubbers *et al.*, 2017). While many cells, also structural cells of the epithelium, can be considered part of the cells of the innate immunity, effector cell populations play a more direct role in host defense. These populations are derived from the hematopoietic system, involved in the generation of the cellular components of the blood, and may be characterized by their maturation sites. Some finish their development in the myeloid tissue, while others do in the lymphoid. Generally, innate immunity's myeloid effector cells lack adaptive elements tailored to the challenge but are generally equipped with tools to fight infection unspecific (Gasteiger *et al.*, 2017). A major subset of these effector cells are the granulocytes, which feature the name-giving granules in their cytoplasm. These granules generally contain broad-spectrum defense elements like defensins, hydrolases and harsh chemical properties like low pH and reactive oxygen species. The four most prominent granulocytes are basophils, eosinophils, neutrophils, and mast cells, with the most common population being neutrophils (Breedveld *et al.*, 2017). While basophils, eosinophils, and mast cells primarily secret pro-inflammatory and anti-pathogen factors into the environment, neutrophils have another critical function. They are part of the effector cell population deemed phagocytes capable of taking up and digesting pathogens from the environment. This ability allows them to act as processing cells for pathogen-derived products and shuttle them to other cells of the immune system via different mechanisms like antigen presentation or through entering programmed cell death (apoptosis) themselves and exposing other phagocytes to their contents (Abdallah *et al.*, 2011; Vono *et al.*, 2017). The uptake and processing of external or internal proteins by phagocytes is an essential link between innate and adaptive immunity, so crucial that there are two cell populations that are considered professional antigen-presenting cells (APC) in the innate immunity. These cell populations are macrophages (MΦ) and dendritic cells (DCs).

As the name suggests, MΦ are relatively large phagocytes and are found widely distributed through many tissues in the body. They play an essential role in the host's defense and homeostasis management by clearing out dead or dying cells and processing fragments derived from them in the phagocytosis process. If pathogens have infected cells, pieces from these will also be processed and presented (Ovchinnikov, 2008). While MΦ have a broad range of functions, from maintaining tissue integrity to managing metabolism, DCs are far more specialized in cross-linking the different parts of the immune system (Wynn, Chawla and Pollard, 2013). DCs, named for their filament-like structure, are deemed the sentinel of the immune system with necessary functionality in bridging the innate and the adaptive immunity. While peripheral MΦ, once engaged, will stay in an area and perform mainly phagocytic activities, DCs will preferentially migrate after being stimulated through the uptake of foreign antigens or external signals like cytokines. The stimulated DCs will drain in far greater numbers to the lymph nodes of the area and proceed there to activate T cells and B cells responsible for mounting targeted response (Tamoutounour *et al.*, 2013). Additionally, while DCs are also considered professional phagocytes like MΦ and Neutrophils, they are far less aggressive in the digestion of material they

have taken up. Thus, they can conserve more information from potential threats for the further activation of specialized components of the immune response (Savina and Amigorena, 2007).

To this end, they feature another characteristic capability called cross-presentation, a process in which DCs can present exogenous antigens through a protein complex generally reserved for presenting internal antigens (Gutiérrez-Martínez *et al.*, 2015; Embgenbroich and Burgdorf, 2018). Innate immunity, especially with its phagocytes and APCs, is essential in hosting a defense and is vital in eradicating threats, including cancer. Since cancers can show a disturbed gene expression and sometimes also a high occurrence of mutations in protein-coding genes, it is upon the APC elements of the innate immunity to take up these aberrant gene products and present them to the adaptive immunity (An *et al.*, 2015; Kang *et al.*, 2020). In cancer therapy, different approaches have been under research to exploit APCs to deliver TAAs to the patient's immune system and stimulate an appropriate response by vaccination-based methods or autologous transfer of APC (El Ansary *et al.*, 2013; Elias A. T. Koch *et al.*, 2022). However, a significant hurdle has been in identifying immunogenic and safe TAAs that could be used as a therapeutic option (Feola *et al.*, 2020).

2.1.2 Adaptive immunity and its hallmark features

With APCs and DCs especially being able to take up and present elements of a destroyed pathogen or cell to other cells, it is in the lymph nodes (LN) where they fulfill one of their primary roles. They form a bridge between innate and adaptive immunity (Bonneau *et al.*, 2006; den Haan, Arens and van Zelm, 2014). With their specialized surface molecules, called the major histocompatibility complex (MHC), APCs can present fragments to one of the two significant cell populations of the adaptive immunity, the T lymphocytes. The other considerable population of the adaptive immunity, B lymphocytes, are APCs themselves and feature antigen-specific receptors that can be activated upon encountering peripheral antigens (Yuseff *et al.*, 2013). These two groups characterize and facilitate two hallmark elements of adaptive immunity – antigen specificity and memory. The first hallmark element, antigen specificity, is a feature that allows cells to produce target-specific antibodies in the case of B cells or receptors in the case of T cells. Importantly, this occurs without needing to encode millions of specific receptors as single genes in the genome. This feat is achieved through the generation of diversity by somatic V(D)J recombination. This remarkable process of diversity generation allows for the exons of the regions that code for the antigen-specific site of an antibody (AB), or antigen specific receptors to be generated by repeated breaks and reassembly of DNA during the development of the lymphocytes. These assembled exons are derived from the blocks called V (Variable), D (Diversity) and J (Joining) elements which function by a cut and paste mechanism with some degrees of freedom in the spacer regions between the fragments. Through this combinatorial system of using different V, D or J blocks while also allowing for small insertions or deletions in the nucleotide sequence, it is possible to generate a large repertoire of antigen recognition for ABs or T cell receptors (TCR) (Roth, 2015). While there are additional

mechanisms for AB diversity creation in B cells, for T cells, the primary avenue is VDJ recombination. The human immune system consists of an estimated 10^{11} T cells and each expressing one TCR only, it is still possible to generate a large spectrum of immunological coverage (Clark *et al.*, 1999).

The second hallmark of the adaptive immunity is memory generation. While recent studies have shown that the innate immunity also undergoes a habituation or training process, no currently established cell types of the innate immunity are known to impart this memory (Netea *et al.*, 2020). In contrast, the adaptive immunity features dedicated subpopulations that are specifically committed to impart long-term storage of the memory of a pathogen encounter and its response for its primary cell types. After encountering strong enough stimulation through exposure to their specific antigen, B cells can proliferate and differentiate into different subpopulations. Given an immunological challenge and the activation of unexposed or naïve B cells, one population of the progeny of the B cell will become either a plasma cell, specialized in producing large quantities of AB and thus directly supporting the current immune reaction. In contrast, the other part of the population will turn into memory B cells that circulate through the body in a hibernation state and show decades-long lifespans. These memory cells require repeated exposure to the antigen to be reactivated and thus provide long-lasting immunity (Taylor, Jenkins and Pape, 2012; Seifert and Küppers, 2016).

Like B cells, T cells can form long-lasting memory populations. Although their exact development path is still unclear, two models are currently being discussed regarding how memory populations of T cells are formed. The first model, the circular or on-off model, poses that once a naïve T cell has been exposed to its antigen, it differentiates into its effector phenotype and clears the infection or cancerous cell population. After clearance and the subsiding of an acute phase of inflammation, a proportion of the effector cells die due to programmed cell death (apoptosis). At the same time, another set differentiates into a memory T cell phenotype. Should the antigen be reencountered, these memory T cells will regain their effector phenotype and proliferate again (Youngblood, Hale and Ahmed, 2013). Notably, the cyclic nature of this model would require the re- and dedifferentiation of these cell populations, a process highly debated and yet to be observed in the non-stem cell, somatic cell populations (Henning, Klebanoff and Restifo, 2018). In an alternative model, deemed the linear differentiation model, memory T cells do not derive from an effector population but directly from a naïve T cell population. The central assumption of this model poses that a gradual process, given continuous and antigenic signaling, turns naïve T cells into memory T cells and subsequently and terminally into effector T cells. This linear path depends on a consistent and increasing antigenic signaling environment that triggers the differentiation of the naïve T cells into memory T cells and, finally, effector T cells (Restifo and Gattinoni, 2013). Both models have shown evidence, and both ways to generate the memory in the T cell population may be true depending on circumstances and tissue (Henning, Roychoudhuri and Restifo, 2018).

Memory and antigen specificity form the foundation for long-lasting and effective immunity in the human immune system. Both aspects, however, also create challenges for the rational design of therapies. The most obvious one is antigen specificity. Short of designing the antigen-specific receptors *de novo*, which has been done for both antibodies and TCRs, it is difficult to find and evaluate possible antigens because of the sheer breadth of possible antigen-receptor diversity. Designing and deploying an artificial TCR or AB is a lengthy process, and in the context of cancer, antigen loss of the tumor is a constant problem, rendering a newly designed TCR or AB useless. While there has been tremendous success and benefit to the patient in tumors with well-defined antigens like the CD19 or CD20 surface molecules in some forms of leukemia, the search for similar markers or antigens remains elusive for other tumor entities (Brentjens *et al.*, 2003, 2013; Smith, 2003; Casan *et al.*, 2018).

Additionally, continuous high antigen signaling, usually necessary to form memory T cells, can lead to a phenotype of T cells with a yet-to-be-established clear lineage. Christened an exhausted T cell characterized by hypo functionality and reduced effector function, these T cells appear to be impaired in their cytotoxic abilities (Blank *et al.*, 2019). Parallels have been observed between memory T cell populations and the exhausted T cell phenotype since similar mechanisms seem to be involved in developing these cell states (Pauken *et al.*, 2016; Yao *et al.*, 2019). Chronic exposure to antigens may trigger this inert state of T cells to limit damage during chronic infection. For cancer, a long-lasting immune challenge, this creates problems for the therapy design around defined antigens. Naturally, therapy against one antigen may produce this exhausted phenotype by introducing high quantities during treatment (Alfei *et al.*, 2019; McLane, Abdel-Hakeem and Wherry, 2019). Hence more research is required to offer a wide range of different antigens circumventing or mitigating the issue of exhaustion by optimizing their selection methodology to avoid or minimize the overall presence of the antigen in other cell populations or tissues of the body.

2.1 Antigen presentation through MHC Class I and Class II

Antigen presentation is an essential part of the immune system. With T cells helping to destroy pathogens and supporting B cells in the generation and adaption of ABs, it is necessary to have a system capable of presenting pieces of immunobiological information, antigens, to the cellular outside. Especially since T cells, as a large effector population of the cellular immune response, cannot directly interact with the intracellular environment nor perform phagocytosis. Hence, a system that can sample and present the intracellular environment to the extracellular space has evolved.

Phagocytes and APCs use this system to take up, process, and present targetable information about a challenge to elements of the adaptive immunity. This molecular-level self / non-self-discrimination system is managed by the major histocompatibility complex (MHC), a name derived from its discovery in tissue rejection after transplantation (Allen, 1955; Cunningham, 1977). Encoded by one of the most polymorphous loci in the human genome, the Human Leucocyte antigen (HLA) locus, the MHC Class I (MHC-I) and MHC Class II (MHC-II) proteins are responsible for presenting endogenously produced antigens and exogenously acquired ones, respectively. Since MHC-I is tasked with presenting intracellular products for self-recognition, it is expressed on all nucleated cells, while MHC-II is only found on professional antigen-presenting cells and foremost on DCs (Ting and Trowsdale, 2002; Li and Raghavan, 2010). Especially with MHC-I being so ubiquitously expressed with its task being specifically signaling self or non-self to effector T cells, it lends itself to be exploited for therapeutical purposes. In searching for restricted genes or transcripts expressed by a tumor, we can attempt to design adjuvant or mono therapies around antigenic peptides derived from these genes, which can preferentially bind these MHC molecules to stimulate T cells against our intended target artificially.

2.1.1 Papers, please – the immunological self-identification process through MHC-I

Under homeostasis, cells present autologous peptides bound to MHC-I to patrolling T cell populations which are, under healthy conditions, tolerant to these self-antigens. These antigens are short, 8 to 12-AA-long peptides non-covalently bound to the MHC-I molecule. They need to be processed and transported to the surface to arrive there. For the protein fragment to be displayed on MHC-I and for the immune system to interact with it, the source protein must undergo several processing steps. The first step is the ubiquitin-proteasome pathway which degrades proteins into peptide fragments (Michalek *et al.*, 1993). These fragments are, in part, further digested and destroyed by peptidases, with some surviving this process. The surviving peptides can be shuttled into the endoplasmatic reticulum (ER) by potentially having a favorable affinity towards a transporter protein complex. This transporter protein is the mediator of the first formal step in loading MHC-I with peptides and is aptly called the transporter associated with antigen processing (TAP). Together with the chaperones Tapasin and Calreticulin, the isomerase ERp57 and the empty MHC-I

molecule, TAP forms the peptide loading complex (PLC) in the ER. Tapasin especially plays a vital role in loading stored MHC-I molecules in the ER since it keeps the empty MHC-I molecule stable in a sterical state that allows for the peptides to bind into the dedicated binding groove. Together with ERp57, Tapasin manages the loading of MHC-I with different peptides (Wearsch and Cresswell, 2007; Garstka *et al.*, 2015). During the loading, multiple peptides can bind non-covalently to the MHC-I molecule until a peptide with a sufficiently high affinity is bound. If the peptides are too long, two peptidases (ERAP1/2) can perform N-terminal trimming of the peptides to improve their binding characteristics (Chang *et al.*, 2005). After a bound peptide stabilizes the complex, it is transported to the cell surface through the Golgi, where it can present the endogenously produced peptide (**Figure 1**). Although this pathway shows a straightforward way for an internal protein product to be presented to the immune system, specialized cells like DCs can present antigens on MHC-I which have been taken up from the extracellular environment in a process deemed cross-presentation (Joffre *et al.*, 2012). Although some details of the cross-presentation mechanism are not entirely understood, two main models have been proposed. One pathway suggested is the phagosome-to-cytosol pathway, in which internalized material in a DC enters the cytosol through export from the phagosomal compartment, a vesicle containing low pH and hydrolytic enzymes (Ackerman *et al.*, 2003; Palmowski *et al.*, 2006). Once in the cytosol, the antigen or peptide can follow the classical path through proteasome degradation and loading in the ER. Alternatively, the vacuolar pathway allows the loading of MHC-I directly in the phagosome after pathogens are degraded. The loading requires the shuttling of ER-derived components like the PLC to the phagosome, which has been demonstrated under some conditions, but the mechanism is still under investigation (Nair-Gupta *et al.*, 2014; Blander, 2018).

Since, from a therapeutical perspective, we must bring antigens somehow to the attention of the effector cells, antigen presentation and cross-presentation by APCs and DCs are important processes. However, they also create difficulties in developing predictive computer models and therapies. Especially predicting the probability of presentation of any given peptide has proven complex due to the many cellular elements involved. Computational models that factor in some of these variables, like the TAP affinity of peptides, have been developed but have yet to produce robust predictions which can be reliably validated, although significant progress has been made (Bhasin, Lata and Raghava, 2007; O'Donnell, Rubinsteyn and Laserson, 2020). Most notably, the lack of training data and missing pieces in the molecular understanding of how antigen presentation and cross-presentation work are persisting issues in constructing accurate, predictive computational models. Thus, current approaches try to model and predict a peptide's chemical affinity to the MHC-I molecule as a generalizable and measurable feature.

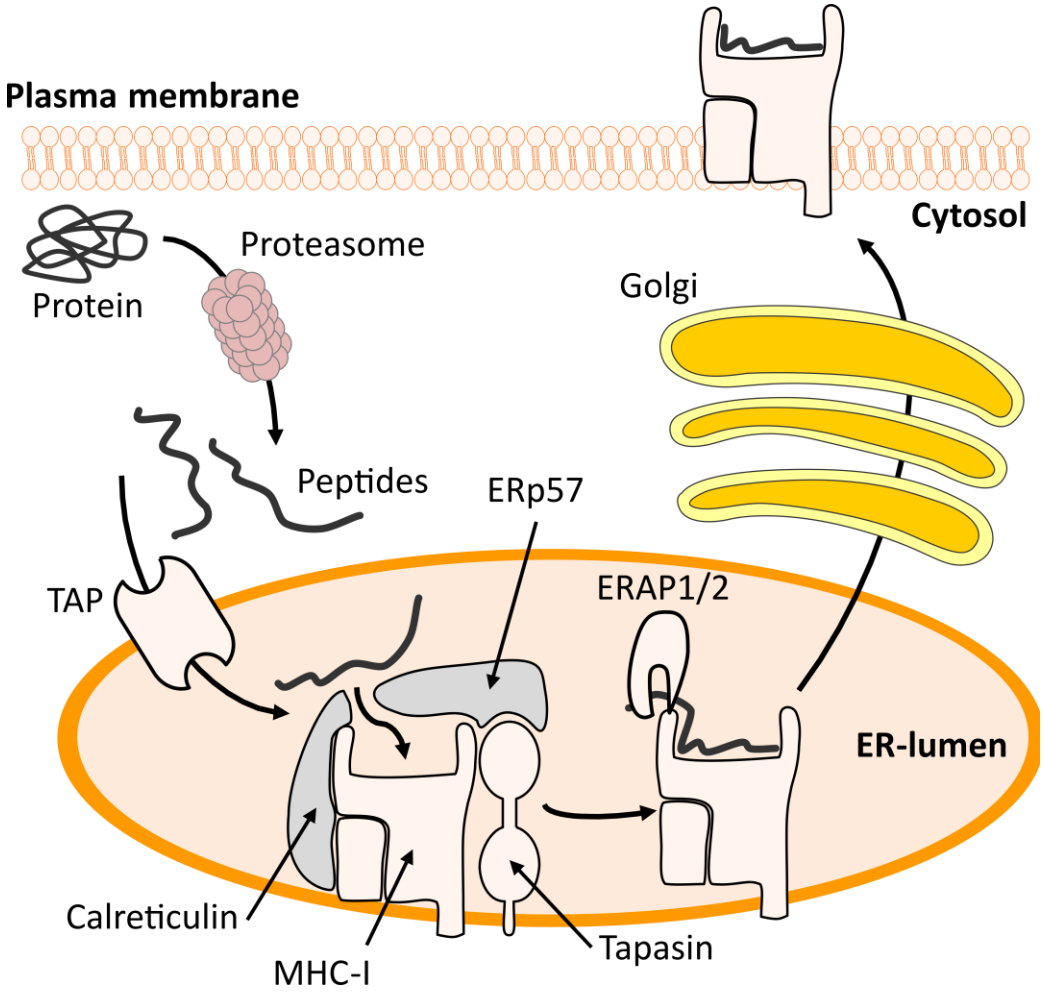


Figure 1: ***Abstracted illustration of the MHC-I processing pathway****. Intracellular proteins are degraded in the cytosol by the Proteasome. Through the TAP transporter, they are translocated into the ER lumen and loaded onto the empty MHC-I molecule. The chaperones Calreticulin and Tapasin stabilize the peptide loading complex, while the isomerase ERp57 aids in peptide loading. If the peptide is too long, trimming through the aminopeptidases ERAP 1 and ERAP2 may occur. Once the complex is stable, it is exported to the plasma membrane through the secretory pathway via the Golgi to the cell's surface to present intracellular products to T cells.*

2.1.2 HLA Class I Polymorphisms- Defense in diversity, scorn for the modeler

The self/non-self-presentation machinery of the immune system has to handle a lot of possible variabilities. Viruses, bacteria, or cancer may all produce different protein variations and, subsequently, peptide subsets that need to be bound and presented to effector T cells reliably for an effective immune response. Evolution has created a highly complex system with human leukocyte antigen (HLA) complex to solve this problem. This set of genes encoded on chromosome 6 is one of the most polymorphic regions in the genome, with an unparalleled number of alleles known. Three major genes, HLA Class I A, B, and C, which are co-dominantly expressed, provide the immune system with a high degree of flexibility to present a large variety of peptides. The IPD-IMGT/HLA database currently holds over 24,000 alleles known for these three subtypes (Robinson *et al.*, 2020). A high degree of polymorphism is predominantly found in the exons coding for the peptide binding pocket, which explains the increased flexibility of these molecules in accepting different peptides (Solberg *et al.*, 2008; Buhler and Sanchez-Mazas, 2011).

This feature is deemed the heterozygous advantage model and was formulated by Doherty and Zinkernagel. It is built on the idea that diversity in these loci confers evolutionary fitness advantage for the individuals in contrast to a homozygous genotype. It is proposed that this results from co-evolution with a high variety of pathogens (Doherty and Zinkernagel, 1975; Meyer and Thomson, 2001). While this complex system allows our immune system to be adaptable to a wide variety of challenges, it creates a significant degree of variables needed to be considered when trying to model binding properties or probabilities of the peptides *in silico*. It has been shown that different HLA alleles have other length preferences, with, for example, HLA-A*01:01[1] preferring peptides of length 10 to 15 while HLA-A*02:01 preferentially bindings 8 to 12 *in vitro*. Through elution of naturally presented peptides, however, it has been determined that most peptides found on MHC-I are of the lengths 9 to 12 AA. Thus, predictive models usually only cover this length space (Trolle *et al.*, 2016).

Additionally, different positions within the peptide are weighted more importantly, and T cells prefer different distributions of AAs bound for a given allele. Adding to this, T cells may not recognize other positions, especially terminal ones in the peptide, since they act as anchor residues to the binding groove of the MHC-I molecule, with some data suggesting that they may yet be recognized by T cells (Calis *et al.*, 2013; Guillaume *et al.*, 2018; Zajonc, 2020). Accounting for all these dimensions immensely increases the number of candidate epitopes (the combination of peptide and MHC-allele). Hence, modeling binding and immune activity is an ongoing research field trying to establish reliable computational models to predict the binding of a peptide to any known HLA allele *in silico* and to help find immunogenic candidates for experimental validation.

[1] HLA nomenclature is a specific system which establishes notation rules for these polymorphic loci. Designating an allele, A*01:01 means that this allele belongs to the group of alleles that encode for the A1 serological antigen with :01 referring to a specific, unique HLA protein. If alleles differ in these digits, this means that they differ in at least one coding nucleotide change between them.

2.2 Cytotoxic T lymphocytes their role in immunity and immunotherapy

T lymphocytes, originating from the bone marrow, migrate to the thymus, where they undergo a maturation process before they enter the periphery. After exiting the thymus, T cells are naïve, meaning they have not yet encountered their cognate antigen through their TCR. Two distinct subsets of naïve T cells are released into circulation, characterized by their defining surface markers. The CD4+ subset, also classified as T helper cells, is mainly responsible for managing and boosting immunity by activating B cells. The other subgroup comprises the cytotoxic CD8+ T cells (CTLs), which play a more direct role in eradicating an infection or immunological challenge. As the name suggests, they can induce apoptosis in cells through different mechanisms if they present the appropriate antigen through the MHC-I surface receptors to the T cell, provided there are secondary activation signals. This ability makes them a highly studied cell type for immunotherapeutic purposes since they aim to employ their antigen-dependent cytotoxic ability against cancer or other diseases (Waldman, Fritz and Lenardo, 2020).

2.2.1 Cellular hunters – Target identification and destruction by T cells

T cells form their TCR during early development in the thymus from randomly combined V(D)J-recombination events. This recombination event leads to a unique TCR for each developing lymphocyte that designates them as a specific T cell clone. While the TCR may be antigen-specific in the context of an ongoing immune response, its binding behavior, with respect to all possible antigens one T cell could encounter, is considered degenerate. This means a T cell can be cross-reactive to achieve a high immunity coverage against foreign antigens. Hence, different TCRs can bind the same antigen, while one TCR may recognize many antigens (Sewell, 2012; Wooldridge *et al.*, 2012).

Fundamentally, once a CTL has been activated by encountering an APC-presenting peptide that its TCR recognizes in the context of MHC Class I, it patrols the periphery outside the lymphatic system for its target cells. These could be cells infected by viruses or cancerous cells expressing mutated or damaged genes. First contact between a CTL and a potential target cell occurs antigen-independent and is mediated by adhesion molecules like LFA-1 on the CTL and ICAM-1/2 on the subject cell (Bierer and Burakoff, 1988; Harjunpää *et al.*, 2019). Should the cell present the antigen on its MHC Class I molecule, the interaction between the target and the CTL increases in strength. It thus extends the contact time between the two cells forming what is generally considered an immunological synapse (Xie, Tato and Davis, 2013). The contact surface between the cells increases through cytoskeletal rearrangements, and more TCRs engage peptide-MHC (pMHC) complexes (Huppa and Davis, 2003). The TCRs recognize non-self-peptides bound on MHC-I through their TCR variable domains. Other molecules like the MHC-I specific CD8 receptor act as a co-stimulatory signal transduction molecule and aid in the final confirmation of the target and subsequent activation of the CTL.

If the CTL was naïve, additional co-stimulatory signaling from an APC through CD28 is necessary to fully activate the CTL (Zumerle, Molon and Viola, 2017). After activation, CTLs clonally expand, meaning that a single CTL with a single TCR proliferates, increasing the amount of antigen-specific CTLs in the body to fight off a threat. Additionally, activated CTLs show increased production of several cytokines and chemokines like Interleukin-2 (IL-2) and Interferon-γ (IFN-γ) (Tomiyama, Matsuda and Takiguchi, 2002). Both of these cytokines can stimulate further CTL differentiation and activation, thus promoting the acute immune response (Castro *et al.*, 2018; Ross and Cantrell, 2018).

Commonly, IFN-γ is used as a marker for *in vitro* confirmation of antigen-specific CTL activation (Schoenborn and Wilson, 2007). Once there are enough peripheral stimuli, the killing machinery is engaged. With close spatial proximity, the CTL begins to secret targeted cell death-inducing granules filled with granzymes, perforin, cathepsin C, and granulysin, which can fuse with the membrane of the targeted cell. Through the action of perforin, pores form in the target cell's membrane through which pro-apoptotic proteases can diffuse (Trapani, 1995). Once in the target cell's cytoplasm, they trigger apoptosis programs, like the caspase-3 cascade, which in turn activates DNA-digesting enzymes. Releasing these apoptotic factors leads to the fragmentation of the cells and potential intra-cellular pathogen debris, and subsequent cell death. The remnants of apoptotic cells are quickly taken up by APCs, which can increase the immune response, thus creating self-sustaining and reinforcing loops (**Figure 2**). While more options exist for the T cell to kill its target, the fundamental aspects remain constant (Gordy and He, 2012). A target is identified, confirmed, and eliminated. This hunting-like approach to antigen-specific killing of target cells by a mobile cell population is particularly interesting for clinical research.

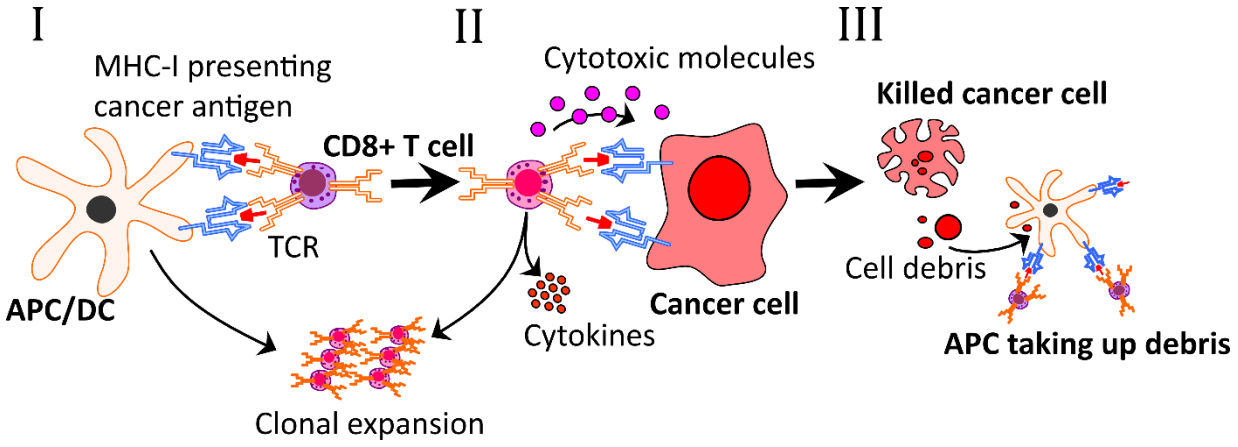


Figure 2: ***Simplified process of antigen-specific cancer cell killing by a cytotoxic CD8+ T cell.*** *(I) Antigen-presenting cell (APC) or Dendritic cell (DC) presents cancer-derived antigen (red molecule) on its MHC-I receptor (blue receptor) to a CD8+ T cell. In the process, if the T cell's TCR (orange receptor) binds the antigen-MHC-I complex, the T cell gets activated and can clonally expand. (II) The activated CD8+ T cells migrate to the tumor site where they can recognize cancer cells antigen-dependent. Through the secretion of cytotoxic molecules, cancer cells are eliminated. (III) Cancer cells are killed through CD8+ T cells. Upon death, the cells leave behind debris which in turn can be taken up by APCs. The APCs can, then, again, stimulate a further reaction from the immune system or stimulate additional T cell clones to react.*

Therapeutically exploiting the T cell population has been discussed in some form or another for the past 50 years and has roots going back even further (Oiseth and Aziz, 2017). Recent exploitations of the cytotoxic potential of T cells have been made with the advent of ICB treatment, an AB-based method to blockade molecules acting as anti-apoptotic signaling molecules strongly expressed by tumors (Favero *et al.*, 2015). Another approach is extracting T cells from a tumor biopsy, stimulating them to expand clonally, and re-transfusing them into the patient. This therapy, called autologous T-cell transfer (ATC), has also shown promising results (Rosenberg *et al.*, 1994; Zacharakis *et al.*, 2018). Circumventing the MHC restriction of T cells, artificially constructed antigen receptors can be introduced into the T cell, making it respond to surface antigens, like ABs. This method, called chimeric antigen receptor (CAR) T cell therapy, has yielded significant results in non-solid tumors. Anti-cancer vaccines or the autologous transfer of APCs are also heavily studied areas of anti-cancer therapy (Raskov *et al.*, 2021).

All methods but checkpoint blockade therapy rely on the presence of tumor-specific T-cell antigens. While ATC can be naïve of antigens, it has been shown that expanding the T cells specifically against tumor antigens yields a better response (Zacharakis *et al.*, 2018). By definition, CAR T cells require a tumor-restricted antigen expressed on the surface of tumor cells, which is not ubiquitously found in the healthy tissue, or the loss of the healthy cells is tolerable and not survival critical. Vaccines and autologous APC transfer rely heavily on prior knowledge of T cell antigens for the vaccine design or APC loading (Van Der Bruggen *et al.*, 1991). Taken together, the study of T cell antigens derived from tumor-restricted gene expression is essential to leverage the cytotoxic potential of T cells.

2.2.2 Central and peripheral tolerance – Managing cytotoxic cells

As we have established, CTLs can kill cells through the recognition of antigen-specific small, generally 9 to 12 amino acid-long peptides bound to the MHC-I molecule. It is immediately apparent that combined with a large population of cytotoxic cells and the nature of the random generation of their antigen receptor, there exists a potential for unintended autoreactivity, which may cause damage in healthy tissue. Hence, a system that efficiently trains CTLs to discriminate between self-antigens and non-self-antigens and select out those T cells that carry a self-reactive receptor must exist. This system is multipronged and generally separated into the central element, an integral part of the maturation of T cells in the thymus, and the peripheral element, a redundancy and *post hoc* method to tamper CTL activity. During development, T cells in the thymus must first demonstrate that their generated TCR can bind to the host's MHC-I alleles, ensuring that base functionality has been established (Davis *et al.*, 1998; Takaba and Takayanagi, 2017). This initial testing process includes an affinity-based selection determining if a T cell will differentiate into a CD8+ CTL or a CD4+ T helper cell by either binding MHC-I or MHC-II preferentially (Anderson and Takahama, 2012). After it has been established that T cells can bind to an MHC receptor, specialized cells and APCs in the thymus will

present self-antigens to the nascent T cells. Medullary thymic epithelial cells (mTECs) will express almost all encoded peripheral genes by a process deemed promiscuous gene expression. This includes genes, which under normal homeostasis, are restricted to highly specialized tissues like the muscle-enriched gene *MYLK2* or the thyroid gland-restricted thyroid stimulating hormone receptor (*TSHR*) (Gabrielsen *et al.*, 2019). These genes are thus known as tissue-restricted antigens (TRA).

However, not every individual mTEC will express all TRAs, but each TRA will only be expressed and presented by a small percentage of the collective with numbers ranging from 1 to 20% (Peterson, Org and Rebane, 2008; Klein *et al.*, 2014; Gabrielsen *et al.*, 2019). Once a CTL encounters a self-antigen on an mTEC and binds the encountered MHC-I self-peptide combination too strongly, it will be clonally deleted through apoptosis. Thus, the escape of a strongly self-reactive T cell clone into the periphery is averted. This system ensures that the randomly generated TCRs, which show a high affinity to self-antigens, are not circulating through the body. However, as with most biological systems, this process is stochastic, restricted by spatial interactions in the thymus, and simply imperfect (Klein *et al.*, 2009).

Autoimmunity induced by self-reactive T cells is a common pathological occurrence, like in type 1 diabetes or inflammatory bowel disease (Kappeler and Mueller, 2000; Pugliese, 2017). In order to manage escaped autoreactive CTLs, several elements are in place roughly characterized by the term peripheral tolerance. Regulatory T cells (T regs), a subset of CD4+ T helper cells, are immune suppressive and actively tamper autoreactive T cells through the excretion of inhibitory cytokines to resolve an ongoing inflammation or maintain self-tolerance during acute inflammation (Kearley *et al.*, 2005). Additionally, suppose a CTL repeatedly encounters its specific antigen without adequate co-stimulatory signals. In that case, the CTL will enter a long-term hypo responsiveness called anergy, characterized by suppressed effector function even when encountering its antigen. This anergic state has been compared and likened to the exhaustion in activated T cells after chronic exposure to antigens in the immune response to cancer (Crespo *et al.*, 2013; Tabana *et al.*, 2021). Immune checkpoint receptors like programmed death receptor 1 (PD-1) and the cytotoxic T lymphocyte-associated protein 4 (CTLA-4) act as off-switches for activated CTLs and can induce an anergy state (Xing and Hogquist, 2012; Syn *et al.*, 2017). While these molecules have been extensively exploited in therapy, dysfunction of CTLs is still an ongoing issue and might be antigen-dependent. Since most tumor-associated antigens (TAA) are, by definition, self-antigens, we must assume that the central tolerance may screen out a large portion of CTLs which may provide therapeutical efficacy. Hence, we also need to consider how much peripheral presence exists, e.g., expression in a TAA's non-tumor tissue, and factor this into our selection methodology. Selection systems need to determine the degree of potential cross-reactivity of a TAA with other tissues, judging what is tolerable regarding autoimmunity and the potential to induce peripheral anergy while balancing it with anti-tumor immunogenicity.

2.3 A perspective on Bioinformatics in antigen based targeted immunotherapy

With the many degrees of variability to be included in the study and the development of targeted immunotherapies, iterative approaches are not feasible anymore. Combined with the availability of high throughput quantitative technologies like transcriptome sequencing, it is apparent that manual experimental approaches may not be sufficient anymore to discover novel targets robustly. With immunotherapies like cancer vaccines or adoptive T-cell transfer aiming for quick turnaround treatment schedules, the accurate prediction of efficacious and tolerable antigens poses a complex and vital task for bioinformaticians (Bulik-Sullivan *et al.*, 2019; Ooki, Shinozaki and Yamaguchi, 2021).

Assuming a median protein length of 375 AAs and with a range of 9 to 12 AAs that preferentially bind MHC-I, we arrive at an average set of 1462 peptides per protein coding sequence in the genome (Brocchieri and Karlin, 2005). Without making prior assumptions, each of these peptides may have the same potential to be a good binder to MHC-I or might elicit an immune response under optimal conditions. While 1462 peptides may still be in the range of batch-testing procedures, each patient can express six different HLA alleles in a worst-case scenario, increasing the candidate number to 8772. This number assumes one gene of interest. In a more realistic scenario, we expect several potential candidate genes, which would increase this number five or ten-fold. Given time, one might be able to test several hundred or thousand antigens for their efficacy.

However, suppose relative haste is required in a clinical setting. In that case, this is not feasible, and manual curation is generally not viable if a rational design is to be followed.

To help remedy this problem and to provide possible off-the-self-treatment options from a large pool of candidates, bioinformatics and *in silico* methods can be employed to rationalize antigen candidate selection. Several key aspects must be addressed for deterministic target design and discovery. First, a data basis must be established regarding which genes may characterize the tumor and differentiate it from other tissues. While many technologies are available today, a robust and cost-effective methodology is high-throughput transcriptomics or RNA Sequencing (RNA-Seq). Provided with isolated RNA from a sample, RNA-Seq uses Next Generation Sequencing (NGS) to quantify transcriptome-wide gene expression of the sample. Briefly, messenger RNA molecules (mRNA) are isolated through their poly-Adenine tail and are reverse transcribed to DNA (cDNA). The cDNA is then fragmented into 200 base pair long molecules, and the nucleotide sequence is established through a process called sequencing by synthesis, which generates reads from the cDNA fragments. By allocating or mapping these reads, to their gene of origin and counting, we can determine the abundance of an mRNA and make assumptions about the expression of the gene (Stark, Grzelak and Hadfield, 2019).

Having established a data basis, we need to address MHC binding probabilities since, as previously mentioned, we can derive many peptides from one protein-coding gene. Different approaches to predict allele-specific binding have been proposed over the years. Some models work by integrating experimentally

determined physiochemical characteristics of peptides into mathematical models, like quantitative matrices, support vector machines, or artificial neural nets (Bhasin and Raghava, 2004). Other models have used peptide motifs, secondary structure prediction, chain flexibility estimations, and solvent accessibility, to name some prominent features used (Novotny *et al.*, 1986; Alix, 1999; Reboul *et al.*, 2012). Machine learning models have unquestionably reigned supreme in recent years, with prominent predictors like MHC-Flurry and NetMHCpan demonstrating pan allele prediction capabilities (Stranzl *et al.*, 2010; O'Donnell, Rubinsteyn and Laserson, 2020). Both solutions also offer predictions of elements of the antigen binding machinery like TAP affinity or immunogenic activity. While significant advances have been made, the overall low power in predicting an MHC-I-restricted antigen's actual presence and biological activity remains a complex issue to address (Bassani-Sternberg *et al.*, 2015). Additionally, it has recently been shown that there are still many issues with epitope prediction, like the lack of adequately curated training data to improve or develop predictive computational models (Prachar *et al.*, 2020).

Further, the existence of an allele-specific binder candidate in a set of predicted epitopes leads to the third issue bioinformatics needs to address. Even if only 1% of potential peptides are binders, the number is extensive, considering six alleles and millions of peptides. Thus, decision support or ranking is a challenging problem but an integral part of the tool suite of bioinformatics. By establishing a rank order of candidates, peptides can be shortlisted for experimental testing that may lead to further understanding of the biological mechanisms leading to their success or failure in eliciting the desired response. Optimally, one would establish a feed-forward loop in which information about their immunogenicity features is re-integrated into the models. A commonly used ranking method is the chain probability approach, in which probabilities are used in a product to arrive at a conservative estimate or rank of a peptide. Conservative, because of the nature of probabilities only being in the range of 0 to 1, including more variables can only keep the overall probability estimate equal or decrease it for the event of interest. For example, one can multiply allele-specific affinity predictions with immunogenicity predictions to refine probabilities of efficacy. Using this method, one can provide a comprehensive, although potentially too strict, ranking for multi-variable selection problems and narrow the field of candidates. There are many more issues bioinformatics or data-driven approaches can address in immunotherapy, like immune evasion modeling through biological pathway analysis, patient response prediction through liquid biopsy, and many more (Fattore *et al.*, 2021; Ischenko *et al.*, 2021). However, all these approaches must be firmly anchored into a biological framework and provide concrete claims that can be tested and validated. Hence *in silico* screening methods should focus, when faced with high dimensional datasets, on producing clearly defined and ranked candidates that, together with experimental experts, can be translated into experimental or clinical validation settings. This outlines a goal we tried to achieve during this project by providing many options for manual input and freedom of selection for users like clinicians or biomedical experts.

2.4 Cutaneous and uveal melanoma – oncology and therapy

Melanoma is a type of cancer that develops from the cells producing pigment, so-called melanocytes. While it is most commonly found as a cancer of the melanocytes in the skin (cutaneous melanoma; CM), it can also stem from the pigmented cells located within the uvea, which are responsible for giving eye coloration (uveal melanoma; UM). In the case of CM, recently, a trend has been observed that while population-wide frequencies have been rising, survival rates have improved over the past decade (Henley *et al.*, 2020). This increase in survival rates is attributed to the development of ICB treatment, which changed the therapy paradigm for the disease. To put this into perspective, therapy for metastatic cutaneous melanoma (MCM) was limited in the not-too-distant past, with a 10-year survival rate of around 10% (Balch *et al.*, 2009). With the development of ICB treatment, especially those targeting PD-1 and CTLA4, these rates could be doubled by showing 10-year survival rates of 22% (Schadendorf *et al.*, 2015). While these are very promising treatment options, patient response rates remain limited. Current data shows that the overall response rate is in the range of 37% to 45%, while complete response demonstrates percentages in the 13 to 19% range. With increasing disease frequencies, alternative or complementary treatments are necessary to supplement existing approaches (Hamid *et al.*, 2019; Curti and Faries, 2021).

The situation is far bleaker in the case of UM. Although it has a low population-wide occurrence, it is the most common ocular cancer in adults and has very limited treatment modalities. Standard approaches include radiation therapy and enucleation of the affected eye, which offers a good prognosis for up to 50% of the patients while the other half develops distant metastasis predominantly in the liver (Kujala, Mäkitie and Kivelä, 2003; Weis *et al.*, 2016). After metastasis, overall survival rates drop to 13.4 months, with the prognosis not significantly changed for the last 30 years. With treatment protocols derived from MCM, ICB treatment has been under clinical investigation for UM for several years with mixed results. Depending on the treatment, overall response rates were quite diverse, ranging from 0% to 25%, with limited improvements in overall survival (Wessely *et al.*, 2020). The reasons for these discrepancies are a multitude. Data suggests that MCM is more prone to generate antigens due to the high mutational burden, while the mutations in UM are comparatively few (Mallet *et al.*, 2014). Additionally, the expression of checkpoint molecules is low in UM, making it not an optimal candidate for ICB treatment. The need for additional treatment modalities in both these tumors is high.

Thus, treatment options like vaccination-based approaches and autologous transfer of immune cells, be it APCs or CTLs, are under investigation (**Figure 3**)(Schuler-Thurner *et al.*, 2015; Bol *et al.*, 2016; Maurer, Butterfield and Vujanovic, 2019; Nathan *et al.*, 2021). These therapy avenues are all necessitating the discovery and validation of tumor-restricted antigens (TAAs) that can be used to design vaccines, discover antigen-specific TCRs, or may be employed to expand naïve populations of immune cells. The discovery of these antigens and their reliable prediction is a problem we set out to address in this project.

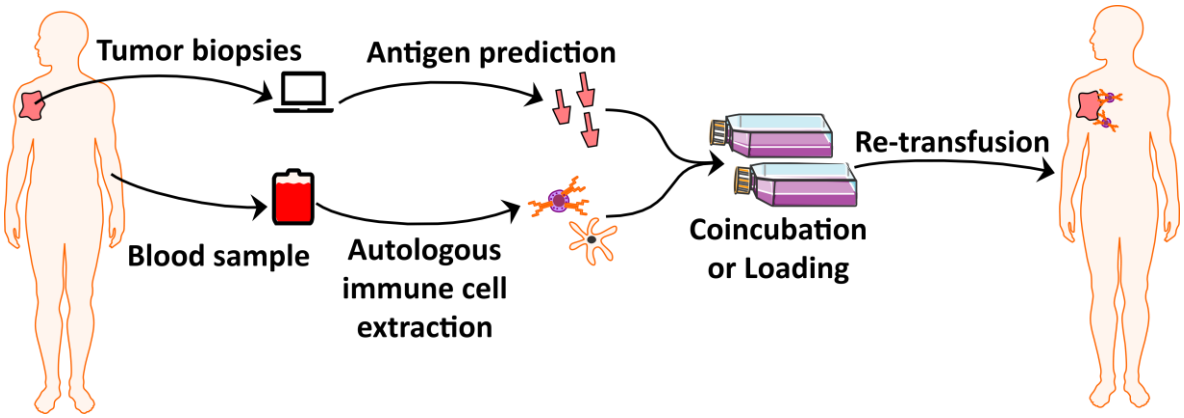


*Figure 3: **Workflow of an example approach to antigen-based immunotherapy.** First, a tumor or multiple tumors are excised or biopsied. From this material, antigen predictions are performed using, for example, transcriptomics measurements. In parallel, a blood sample from the patient is collected, and the autologous immune cells are isolated. Once high-confidence antigen candidates have been established, they may be used to either stimulate autologous T cells or loaded onto autologous DCs. Afterward, the cells are getting re-transfused with the intention of fighting the cancer using the primed and stimulated autologous immune cells.*

2.5 Tumor-associated antigens – threading the immunological needle

The concept of tumor-associated antigens (TAAs) is based on the observation that some genes are heavily overexpressed in tumors compared to healthy cell populations. By extension, this leads to the presentation of peptides derived from these heavily overexpressed genes on MHC-I and the subsequent possibility of CTLs to mount an immune response against these antigens under the right conditions (Boon and Van der Bruggen, 1996; Sadozai *et al.*, 2017). Generally, the assumption is that a gene not mutated during tumorigenesis is still a self-antigen and is thus covered by the central tolerance. However, in some cases, genes are sequestered from the immune system due to their expression being limited to either very early developmental states (oncofetal antigens) or highly specialized cell populations like male germ cells (cancer-testis antigens, CTAs). It has been shown that cancers can express these genes in large quantities, and it is assumed that they convey essential steps during malignant transformation by providing many necessary features like motility, colonization abilities, and unrestricted proliferation (Gjerstorff, Andersen and Ditzel, 2015). One of the first TAAs discovered is the melanoma-associated antigen 1 (*MAGE-A1*). It was found by characterizing T cell clones of melanoma patients with a favorable disease course, showing that these antigens are immunogenic (Van Der Bruggen *et al.*, 1991). These characteristics and their specificity to the tumors make them valuable targets of interest for therapy. Hence, efforts have been made to find and characterize more TAAs through modern high-throughput technologies like transcriptomics or proteomics, and especially for CTAs, databases have been created to curate the findings (Almeida *et al.*, 2009; Olsen *et al.*, 2017; Koşaloğlu-Yalçın *et al.*, 2021). Many targets for different cancer types have been discovered over the recent years, with the question manifesting: How do we know which antigens to further investigate for therapy? The amount of TAAs we can now find far exceeds screening capabilities, with one database like TANTIGEN holding 292 different TAAs and more than 1000 tumor peptides, of which many are not yet characterized in terms of their immunogenic potential (Olsen *et al.*, 2017).

Several obstacles are intrinsic when using self-antigens. One acute problem is the possible lack of reactive lymphocytes for the antigen. Even though the antigens may not be screened by central tolerance, the possibility exists that few or no T cell clones are generated against these antigens. Thus, poor immunogenicity is an ongoing problem (Higgins, Bernstein and Hodge, 2009). If antigen-specific T cell clones exist and are not yet rejecting the tumor, exogenously applied co-stimulation, like cytokine administration or ICB treatment, to overcome the tolerance might induce severe autoimmune side effects with drastic consequences. Indeed, it has been observed that targeting *MART1* in melanoma patients has caused autoimmunity while treating against *MAGE-A3* has caused severe neurotoxicity in a recent trial (Morgan *et al.*, 2013; Chodon *et al.*, 2014). Hence, using TAAs in therapy is threading the immunological needle by balancing possible toxicities with a lack of immunity. With our project, we hope to contribute to a rational design strategy that effectively allows the screening of TAAs concerning their potential efficacy and autoimmunity.

2.6 Aims

With an ongoing need for anti-cancer therapies, the growing availability of sequencing data, and the expansion of vector systems able to deliver antigens directly to a patient, targeted antigen-based immunotherapies can play a significant role in cancer therapy. As of May 2022, over 500 trials against many different neoplasms are registered under clinicaltrials.gov and labeled with "Tumor-associated antigens | Cancer." However, there is a distinct lack of reproducible and rational workflows to discover new antigen candidates comprehensively.

Hence, we aimed to create a pipeline that integrates various first principal data sources to perform predictions to fill this gap (**Figure 4**). We integrated transcriptomics and histology to predict self-tolerant, immunogenic anti-cancer antigen candidates for translation into *in vitro* testing or clinical trials. In detail, we describe the development of a pipeline constructed to identify TAAs with limited to no expression in the healthy periphery outside the tumor for peptide or antigen-based tumor immunotherapy. Our shortlisted antigens and their source genes provide several predicted favorable characteristics for use in therapy. Our predicted antigens are non-mutated, making them applicable to large cohorts of patients. They are further expected to be self-tolerant, producing possibly fewer autoimmune reactions, and are ranked with multi-variable scores, making their selection easier for clinical or *in vitro* validation. All antigens and their meta-information, like expression, have been aggregated into a comprehensive database that provides functionality for selecting user-defined antigen lists for trial design or *in vitro* testing.

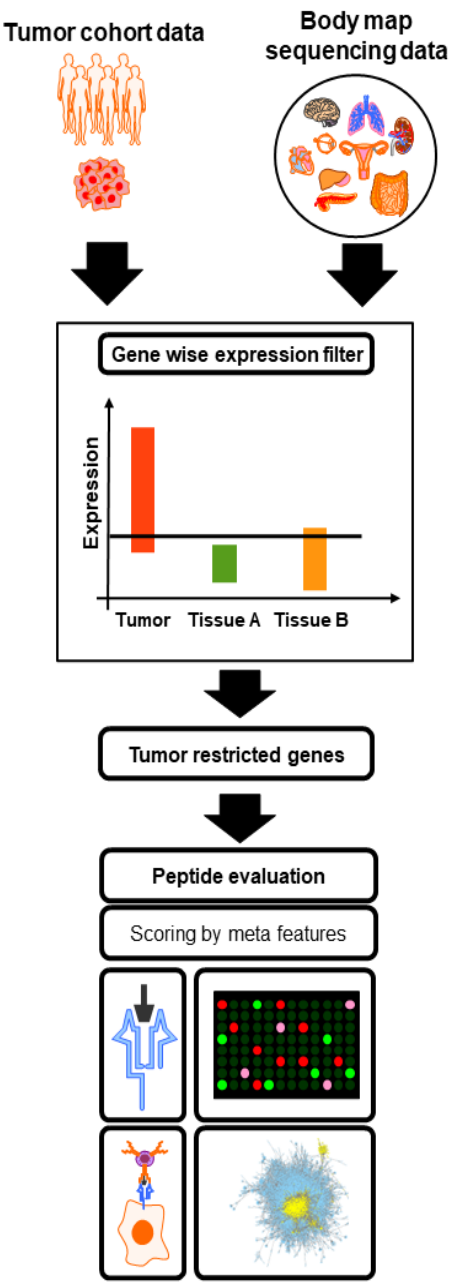


Figure 4: ***Abstracted illustration of the core concepts of the pipeline.*** *To discover novel tumor-associated antigens, we use patient cohort sequencing data to compare it with healthy tissue by a sequential filtering procedure. Finally, we implement several evaluation criteria to help select candidates for application.*

3 Materials and Methods

3.1 Acquisition, generation, and processing of transcriptomic data

During this project, several transcriptomic datasets were used that were either obtained from publicly available repositories or generated from in-house patient biopsies. In the project part relating to MCM, two paired-end sequencing cohorts of publicly available melanoma biopsies were downloaded from the Gene Expression Omnibus repository. There were 27 pre-treatment MCM metastasis samples from GSE78220 (Liu, Beyer and Aebersold, 2016), excluding GSM2069836 and all five samples from GSE96619 (Garcia-Diaz *et al.*, 2017) for a total of 31 metastatic cutaneous melanoma samples. These samples compose our melanoma sample cohort. In the project part relating to UM, transcriptomic data was generated from primary tumor material obtained from in-house patient biopsies. These samples were obtained and processed by scientific and clinical collaborators in the group of Experimental Immunotherapy and the Group of RNA-based Immunotherapy at the Dermatology Department, University Hospital Erlangen. In accordance with current regulatory and ethics standards within the context of the clinical trial registration NCT01983748, informed consent was obtained, and uveal melanoma (UM) biopsies were taken after enucleation from the afflicted eye of 14 patients. Samples were stored and persevered for further processing in RNA*later*™ (ThermoFisher Scientific, Waltham, MA, USA). The RNA was extracted with RNeasy Mini kits (Qiagen, Hilden, Germany) according to the manufacturer's protocol and shipped to a commercial sequencing provider, where sequencing and primary quality control of the RNA were performed (**Table 1**). Transcriptomic sequencing was performed by a commercial service provider (CeGaT Tübingen). For all mentioned samples, data was stored in the FASTQ format and was further processed using in-house pipelines (Peter J.A. Cock *et al.*, 2009).

Table 1: ***Results of entry quality control at sequencing facility using Bioanalyzer RNA Nano****. Listed are the sample identifiers, the RNA concentration, the RNA integrity number, and the sample volume.*

Sample	*RNA ng/µl*	*RIN*	*Volume [µl]*
S727Nr1	103	9	13
S727Nr2	122	10	13
S727Nr3	168	10	13
S727Nr4	107	10	13
S727Nr5	147	10	13
S727Nr6	111	9	13
S727Nr7	123	9	13
S727Nr8	127	10	13
S727Nr9	90	10	13
S727Nr10	111	9	13
S727Nr11	107	9	13
S727Nr12	139	9	13
S727Nr13	104	9	13
S727Nr14	118	10	13

3.1.1 Primary quality control

For the 31 MCM samples, primary quality control (QC) was performed using FASTQC, checking for possible technical adapter contamination and overall basecall (Phred) quality (Andrews *et al.*, 2012). GSM2069836 was subsequently excluded for overall unfavorable Phred scores (Bonfield and Staden, 1995). If necessary, adapters were removed, and reads were quality-trimmed using *BBduk,* part of the *BBTools* suite (Bushnell, 2014). After all primary QC was concluded, samples were processed using in-house software pipelines.

For just the 14 primary UM samples, primary QC, including adapter removal and read quality filtering and trimming, was performed by the commercial service provider. Before further processing and analysis using in-house pipelines, delivered FASTQ files were assessed for overall Phred score. (Bonfield and Staden, 1995; Peter J.A. Cock *et al.*, 2009).

```
# Exemplary usage of the bbduk function included in the bbmap package.
Additional options allow for quality trimming and read length filtering
as necessary. Adapter references for the most used illumine sequencing
adapters are included in the bbmap resource package.

#!/usr/bin/bash

ADAPTERS="/home/lischecr/permanent/App/bbmap/resources/adapters.fa"

bbduk.sh in1="Sample1_1.fq.gz‟ \

in2="Sample1_2.fq.gz‟ \

out1="Sample1_1.qc.fq.gz" \

out2="Sample1_2.qc.fq.gz" \

ref="$ADAPTERS" \

overwrite=true
```

3.1.2 Read alignment

All samples were aligned using the short-read, splice-aware mapper *STAR* (Dobin *et al.*, 2013) in version 2.2.1. As a reference, the nucleotide sequence of the human genome assembly GRCh38 in fasta (.fa) format was used in combination with the comprehensive GENCODE annotation set (Schneider *et al.*, 2017; Frankish *et al.*, 2019). Before mapping the samples, the genome reference was indexed using STAR's inbuilt indexing function using the following default command:

```
#!/usr/bin/bash
GENOME="/home/_common/NGS_mapping/human-9606/hg38-indices/"
STAR --runThreadN 20 \
--runMode genomeGenerate \
--genomeDir $GENOME \
--genomeFastaFiles "$GENOME/hg38.fa" \
--sjdbOverhang 99
```

After the genome index was created, we applied STAR to map the reads to the reference using the following command and options derived from the recommendations in the documentation:

```
#!/usr/bin/bash
GENOME="/home/_common/NGS_mapping/human-9606/hg38-indices/"
ANNOTATION="/home/lischecr/permanent/Src/hg38-annotation/gencode.v28.an-
notation.gtf"
STAR --genomeDir "$GENOME" \
--sjdbGTFfile "$ANNOTATION" \
--runThreadN 20 \
--readFilesIn "Sample1_1.qc.fq.gz" "Sample1_2.qc.fq.gz" \
--readFilesCommand zcat \
--outFileNamePrefix "Sample1" \
--outSAMstrandField intronMotif \
--outSAMtype BAM SortedByCoordinate \
--alignIntronMin 20
--alignIntronMax 500000 \
--alignMatesGapMax 1000000 \
--sjdbOverhang 99 \
--outFilterMultimapNmax 20 \
--outFilterMismatchNoverLmax 0.04 \
--outFilterIntronMotifs RemoveNoncanonical
```

3.1.3 Quantification of transcript abundance

After mapping the raw read files, the output of STAR was stored in the binary alignment map (BAM) format (Li *et al.*, 2009) and sorted in ascending order by chromosomal coordinates for further processing in the quantification pipeline. We used two tools to generate relative transcript expression abundance estimates from the BAM files. First, for the MCM samples, we applied the tool *Cufflinks* in version 2.2.1 to generate transcript-level fragments per kilobase o transcript per million fragments mapped (FPKM), a common unit to utilized to measure transcript or gene-level expression, using the same GENCODE annotation set as previously used for mapping (Mortazavi *et al.*, 2008; Trapnell *et al.*, 2012; Conesa *et al.*, 2016).

Since many published databases do not report their expression in FPKM but in the transcripts per million (TPM) metric, which is considered more comparable and robust, we transformed FPKM to TPM using the formula:

$$TPM_i = \frac{FPKM_i}{sum(FPKM)} \cdot 10^6 \ (\boldsymbol{I})$$

With the TPM of a $Gene_i$ being defined as the FPKM of $Gene_i$ divided by the sum of all FPKMs in the sample scaled by a million. In our subsequent work on the UM cohort, we applied the more modern quantification solution *StringTie* in version 2.1.5 to our BAM files (Pertea *et al.*, 2015). *StringTie*, by default, returns the TPM metric as well as FPKM; hence no transformation was necessary.

The commands used for *Cufflinks* and *StringTie*:

```
#!/usr/bin/bash


# Command used for all cutaneous melanoma samples:
cufflinks -p 20 \
-o "Sample1.FPKM.csv" \
-G " gencode.v28.annotation.gtf" \
"Sample1.bam"


# Command used for all uvea melanoma samples:
stringtie "Sample1.bam" \
-o "Sample1.gtf" \
-G "gencode.v28.annotation.gtf" \
-p 20
```

3.1.4 Externally processed transcriptomics data

To bolster our data basis on the rare cancer UM, we downloaded the UM transcriptomics dataset from a previous study with 80 primary UM samples (Robertson *et al.*, 2017). Due to patient data protection rules, public access to FASTQ and BAM files was restricted, and only non-sequence level data was freely available. Thus, we acquired the gene-level FPKM and transformed FPKM to TPM, as described in section 3.1.3.

3.2 Databases and annotation sets

3.2.1 Selection of protein-coding genes from transcriptomics data

To confirm that the expressed and selected genes in the tumor transcriptomics data would indeed be able to produce targetable MHC-I restricted antigens, it was necessary to control that they were annotated as having any protein product. To this end, we used several databases and annotation sets throughout our pipeline to verify the presence of a protein product. For the first iteration of our selection system, selecting targets for MCM, we applied the API provided by *biomaRt* to link a gene's unique identifier to its presence in different databases (Durinck *et al.*, 2009). We used all genes found by our processing pipeline for the MCM samples and ensured that we would only have protein-coding genes in our dataset. By applying *biomaRt,* an R package for database cross-linking and data mining, we determined if the genes were present in the consensus coding sequence database (CCDS), the Human Protein Atlas (HPA), or the Ensemble genome annotation database (Pruitt *et al.*, 2009; Uhlén *et al.*, 2015; Howe *et al.*, 2021). If a gene was found to be annotated as protein-coding in any of these databases, we included it in downstream analysis.

For the UM analysis of our project, searching for new targetable antigens for primary uveal melanoma, we forwent using *biomaRt* and implemented a new validation mechanism. The replacement of *biomaRt* was done to improve performance and usability since biomaRt queries require an internet connection and depend on the online status and user load of the Ensemble servers while also being comparatively slow for large bulk queries. We constructed an internal database from the human reference proteome for the human reference genome assembly GRCh38 provided by the Ensemble database project release 94 (Howe *et al.*, 2021). The database consisted of a key-value pair construct, with the key being a gene’s unique identifier and the value being a list of all the protein sequences stored in the reference proteome, allowing us to input a gene identifier and retrieve all its associated protein products.

3.2.2 Reference Expression Databases

To integrate and develop a tolerance filter for putative TAAs, we downloaded Human Protein Atlas (HPA)'s normal tissue immunohistochemistry database, which features a table of genes by their ensemble gene identifier and a given tissue (Uhlén *et al.*, 2015). It further quantifies the protein profile by tissue microarray into qualitative levels of detection (high, medium, low, and not detected) for 58 tissues and provides protein abundance data for 13207 genes. We retrieved HPA Version 18, which is based on the Ensemble release version 88.38. The tissues included in the database are as follows, adrenal gland, appendix, bone marrow, breast, bronchus, caudate, cerebellum, cerebral cortex, cervix uterine, choroid plexus, colon, duodenum, two samples of the endometrium, epididymis, esophagus, eye, fallopian tube, gallbladder, hair, heart muscle, hippocampus, hypothalamus, kidney, lactating breast, liver, lung, lymph node, nasopharynx, oral mucosa, ovary, pancreas, parathyroid gland, pituitary gland, placenta, prostate, rectum, retina, salivary gland, seminal vesicle, skeletal muscle, three samples skin, small intestine, smooth muscle, two samples soft tissue, sole of the foot, spleen, two samples of stomach tissue, testis, thymus, thyroid gland, tonsils, urinary bladder, and vagina (Uhlén *et al.*, 2015). As a secondary level of evidence for the presence of a protein in each healthy tissue, we downloaded an RNA sequencing-based body map from the Genotype-Tissue Expression (GTEx) project, which provides TPMs for known genes over the entire human body (Carithers *et al.*, 2015). This dataset contains 11688 individual sequencing runs of 57 healthy tissues and cell lines, ranging from 5 to 564 samples per tissue. We removed all cell line-derived samples for a final set of 51 tissues. Using manual curation, we classified a subset of tissues in the database as "critical tissue", meaning that in these tissues, side effects are less or not tolerable due to their importance for survival. For an overview of the tissues found in the GTEx expression database, official release 7, and their classification, see **Table 2.**

Table 2: ***List of GTEx recorded tissues used in our analysis****. If the tissue was deemed survival critical, it is marked as such in the third column. The link to the original repository is included.*

Tissue	GTEx Portal Link	Critical Tissue
Adipose.Subcutaneous	https://gtexportal.org/home/tissue/Adipose_Subcutaneous	
Adipose.Visceral_Omentum	https://gtexportal.org/home/tissue/Adipose_Visceral_Omentum	
AdrenalGland	https://gtexportal.org/home/tissue/Adrenal_Gland	x
Artery.Aorta	https://gtexportal.org/home/tissue/Artery_Aorta	
Artery.Coronary	https://gtexportal.org/home/tissue/Artery_Coronary	x
Artery.Tibial	https://gtexportal.org/home/tissue/Artery_Tibial	
Bladder	https://gtexportal.org/home/tissue/Bladder	
Brain.Amygdala	https://gtexportal.org/home/tissue/Brain_Amygdala	x
Brain.Anteriorcingulatecortex_BA24	https://gtexportal.org/home/tissue/Brain_Anterior_cingulate_cortex_BA24	x
Brain.Caudate_basalganglia	https://gtexportal.org/home/tissue/Brain_Caudate_basal_ganglia	x
Brain.CerebellarHemisphere	https://gtexportal.org/home/tissue/Brain_Cerebellar_Hemisphere	x
Brain.Cerebellum	https://gtexportal.org/home/tissue/Brain_Cerebellum	x
Brain.Cortex	https://gtexportal.org/home/tissue/Brain_Cortex	x

Brain.FrontalCortex_BA9	https://gtexportal.org/home/tissue/Brain_Frontal_Cortex_BA9	x
Brain.Hippocampus	https://gtexportal.org/home/tissue/Brain_Hippocampus	x
Brain.Hypothalamus	https://gtexportal.org/home/tissue/Brain_Hypothalamus	x
Brain.Nucleusaccumbens_basalganglia	https://gtexportal.org/home/tissue/Brain_Nucleusaccumbens_basal ganglia	x
Brain.Putamen_basalganglia	https://gtexportal.org/home/tissue/Brain_Putamen_basal_ganglia	x
Brain.Spinalcord_cervicalc.1	https://gtexportal.org/home/tissue/Brain_Spinal_cord_cervical_c-1	x
Brain.Substantianigra	https://gtexportal.org/home/tissue/Brain_Substantianigra	x
Breast.MammaryTissue	https://gtexportal.org/home/tissue/Breast_Mammary_Tissue	
Cervix.Ectocervix	https://gtexportal.org/home/tissue/Cervix_Ectocervix	
Cervix.Endocervix	https://gtexportal.org/home/tissue/Cervix_Endocervix	
Colon.Sigmoid	https://gtexportal.org/home/tissue/Colon_Sigmoid	x
Colon.Transverse	https://gtexportal.org/home/tissue/Colon_Transverse	x
Esophagus.GastroesophagealJunction	https://gtexportal.org/home/tissue/Esophagus_Gastroesophageal_J unction	x
Esophagus.Mucosa	https://gtexportal.org/home/tissue/Esophagus_Mucosa	x
Esophagus.Muscularis	https://gtexportal.org/home/tissue/Esophagus_Muscularis	x
FallopianTube	https://gtexportal.org/home/tissue/Fallopian_Tube	
Heart.AtrialAppendage	https://gtexportal.org/home/tissue/Heart_Atrial_Appendage	x
Heart.LeftVentricle	https://gtexportal.org/home/tissue/Heart_Left_Ventricle	x
Kidney.Cortex	https://gtexportal.org/home/tissue/Kidney_Cortex	x
Liver	https://gtexportal.org/home/tissue/Liver	x
Lung	https://gtexportal.org/home/tissue/Lung	x
MinorSalivaryGland	https://gtexportal.org/home/tissue/Minor_Salivary_Gland	
Muscle.Skeletal	https://gtexportal.org/home/tissue/Muscle_Skeletal	
Nerve.Tibial	https://gtexportal.org/home/tissue/Nerve_Tibial	
Ovary	https://gtexportal.org/home/tissue/Ovary	
Pancreas	https://gtexportal.org/home/tissue/Pancreas	x
Pituitary	https://gtexportal.org/home/tissue/Pituitary	
Prostate	https://gtexportal.org/home/tissue/Prostate	
Skin.NotSunExposed_Suprapubic	https://gtexportal.org/home/tissue/Skin_Not_Sun_Exposed_Suprap ubic	
Skin.SunExposed_Lowerleg	https://gtexportal.org/home/tissue/Skin_Sun_Exposed_Lower_leg	
Small intestine.TerminalIleum	https://gtexportal.org/home/tissue/Small_Intestine_Terminal_Ileum	x
Spleen	https://gtexportal.org/home/tissue/Spleen	
Stomach	https://gtexportal.org/home/tissue/Stomach	x
Testis	https://gtexportal.org/home/tissue/Testis	
Thyroid	https://gtexportal.org/home/tissue/Thyroid	
Uterus	https://gtexportal.org/home/tissue/Uterus	
Vagina	https://gtexportal.org/home/tissue/Vagina	
WholeBlood	https://gtexportal.org/home/tissue/Whole_Blood	x

3.2.3 Curation of known melanoma antigens from additional sources

By querying databases like CAPED (Vigneron *et al.*, 2013), TANTIGEN (Olsen *et al.*, 2017), and CTPedia (Almeida *et al.*, 2009), as well as literature research, we compiled a list of known melanoma antigens as reference points (**Table 3**).

Table 3: ***Manually curated melanoma-associated antigens reported in the literature and public databases.*** *We curated a list of known peptides that have been shown to generate an immune response in different studies involving metastatic melanoma. If no preferentially bound allele was provided for the peptide, the allele filed holds "NA" for stating not available.*

Gene Symbol	Peptide Sequence	Allele	Source	PMID
CDKN2A	AVCPWTWLR	A*11:01	Huang,2004	15128789
CLPP	ILDKVLVHL	A*02:01	Corbière, 2011	21216894
CSNK1A1	GLFGDIYLA	A*02:01	Robbins, 2013	23644516
CTAG1B	MSLQRQFLR	NA	Wang, 1996	8642255
CTAG1B	LSLLMWITQC	A*02:01	Robbins, 2016	25538264
CTAG1B	SLLMWITQC	A*02:01	Gibney, 2015	25524312
CTAG1B	SLLMWITQCFL	A*02:01	Nicholaou,2011	21698545
DCT	ANDPIFVVL	C*08:02	Castelli, 1999	9973437
DCT	LLGPGRPYR	A*31:01	Wang, 1996	8976176
DCT	LLGPGRPYR	A*33:01	Wang, 1998	9551926
DCT	TLDSQVMSL	A*02:01	Noppen, 2000	10861482
DCT	SVYDFFVWL	A*02:01	Parkhurst, 1998	9809996
FOLH1	GLPSIPVHPV	A*02:01	Weber, 2013	21760528
GAS7	SLADEAEVYL	A*02:01	Robbins, 2013	23644516
GPR143	LYSACFWWL	A*24:02	Touloukian, 2003	12538723
HAUS3	ILNAMIAKI	A*02:01	Robbins, 2013	23644516
MLANA	AAGIGILTV	A*02:01	Kawakami, 1994	7516411
MLANA	AEEAAGIGIL	B*45:01	Schneider, 1998	9455808
MLANA	AEEAAGIGILT	B*45:01	Schneider, 1998	9455808
MLANA	EAAGIGILTV	B*35:01	Benlalam, 2003	14634146
MLANA	EAAGIGILTV	A*02:01	Fleischauer, 1996	8752930
MLANA	RNGYRALMDKS	C*07:01	Larrieu, 2008	18097665
MLANA	ILTVILGVL	A*02:01	Castelli, 1995	7807017
PMEL	ALLAVGATK	A*03:01	Skipper, 1996	8943411
PMEL	ALNFPGSQK	A*03:01	Kawashima, 1998	9797143
PMEL	ALNFPGSQK	A*11:01	Kawashima, 1998	9797143
PMEL	AMLGTHTMEV	A*02:01	Tsai, 1997	9029118
PMEL	LLDGTATLRL	A*02:01	Wang, 1998	9551926
PMEL	HTMEVTVYHR	A*68:01	Sensi, 2002	12135425
PMEL	IALNFPGSQK	A*03:01	Kawashima, 1998	9797143
PMEL	LPHSSSHWL	B*35:01	Vigneron, 2005	15713214
PMEL	MLGTHTMEV	A*02:01	Tsai, 1997	9029118
PMEL	RLMKQDFSV	A*02:01	Kawakami, 1998	9862734
PMEL	RLPRIFCSC	A*02:01	Kawakami, 1998	9862734

PMEL	RSYVPLAHR	A*32:01	Michaux, 2014	24453253
PMEL	RTKQLYPEW	A*32:01	Vigneron, 2004	15001714
PMEL	SEIWRDIDFD	NA	Brichard, 1996	8566071
PMEL	SLADTNSLAV	A*02:01	Tsai, 1997	9029118
PMEL	ITDQVPFSV	A*02:01	Kawakami, 1995	7706734
PMEL	KTWGQYWQV	A*02:01	Kawakami, 1995	7706734
PMEL	LHHAFVDSIF	NA	Lennerz, 2005	16247014
PMEL	YMDGTMSQV	NA	Skipper, 1996	8627164
PMEL	IYMDGTADFSF	NA	Dalet, 2011	21670269
PMEL	SSPGCQPPA	B*07:02	Lennerz, 2005	16247014
PMEL	VLYRYGSFSV	A*02:01	Kawakami, 1995	7706734
PMEL	VPLDCVLYRY	B*35:01	Benlalam, 2003	14634146
PMEL	VYFFLPDHL	A*24:02	Robbins, 1997	9200467
PMEL	YLEPGPVTA	A*02:01	Cox, 1994	7513441
PMEL	LIYRRRLMK	A*03:01	Kawakami, 1998	9862734
PPP1R3B	YTDFHCQYV	A*01:01	Robbins, 2013	23644516
PRAME	ISPEKEEQYIA	A*02:01	Weber, 2013	21760528
PRAME	SLLQHLIGL	A*02:01	Weber, 2013	21760528
RAB38	VLHWDPETV	A*02:01	Walton, 2006	17114498
TYR	AFLPWHRLF	A*24:02	Kang, 1995	7543520
TYR	CLLWSFQTSA	A*02:01	Riley, 2001	11394498
TYR	LPSSADVEF	B*35:01	Morel, 1999	10597191
TYR	MLLAVLYCL	A*02:01	Woelfel,1994	8125142
TYR	QCSGNFMGF	A*26:01	Lennerz, 2005	16247014
TYR	KCDICTDEY	A*01:01	Kittlesen, 1998	9498746
TYR	TPRLPSSADVEF	B*35:01	Benlalam, 2003	14634146
TYR	SSDYVIPIGTY	A*01:01	Kawakami, 1998	9862734
TYR	YMDGTMSQVA	A*02:01	Powell, 2008	17056585

3.3 Selection of candidate genes

The core of our developed methodology is constructed through a multi-level and multi-variable evaluation step that selects candidate genes from transcriptomics data for any tumor model of interest. The final goal of this framework is to provide a set of genes and their derived MHC-I restricted epitopes that allow for precise targeting of the tumor entity while minimizing potential side effects or off-site damage. Each step is designed to provide additional levels of evidence to support the final selection of target genes, of which then allele-specific epitopes are generated and deposited in a database. The following sections will describe the core selection mechanism in detail, while an overview of the methodology is provided in **Figure 5**.

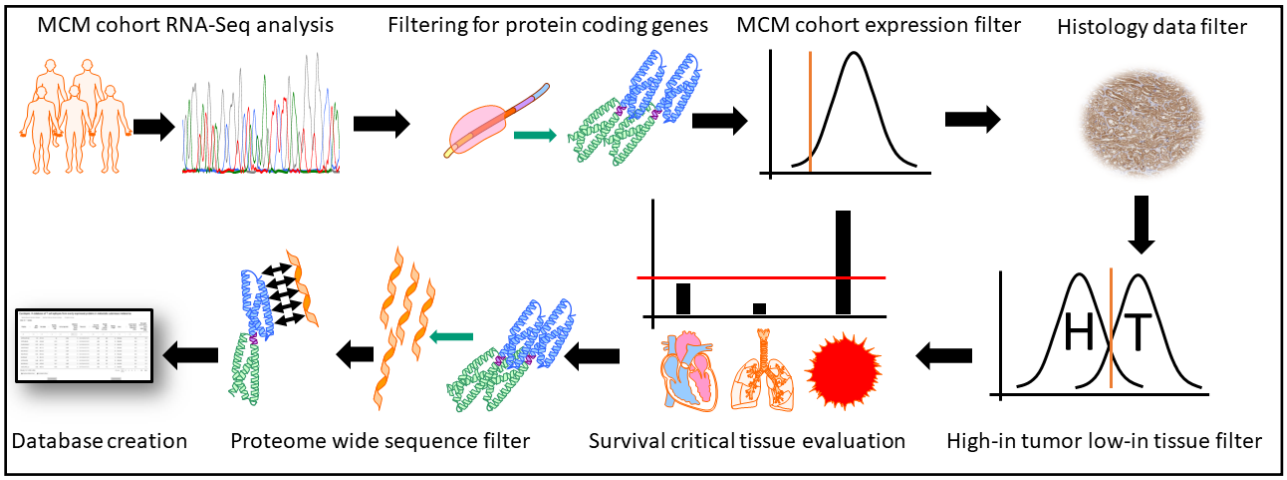


Figure 5: ***Abstraction of the epitope generation and selection process.*** *Tumor RNA-Seq from a melanoma patient cohort was analyzed and filtered only to contain protein-coding genes. All genes too lowly expressed in 90% of the melanomas were excluded. Next, all genes were filtered against histochemical evidence available in the Human Proteome Atlas. If present in any tissue, the gene was removed. We then selected genes that showed a high-in-tumor, low-in-tissue phenotype. The expression of the genes in a curated list of tissues deemed critical for survival was evaluated, and genes were separated into two tolerability sets. The predicted epitopes for those genes were finally filtered against the available proteome and added to the database if they did not occur in any other sequence.*

3.3.1 Determining overly expressed genes in tumor models against healthy tissue

The fundamental concept of our selection methodology was to establish a set of protein-coding genes, which were, to a degree, restricted in expression to the tumor model in question. We thus constructed a pipeline that would select genes, given tumor and tissue expression data, for a high-in-tumor, low-in-tissue-phenotype. To this end, we used the previously processed MCM cohort, consisting of 31 samples, and the 80 primary UM cohort. We calculated the mean, median, 99th, 90th, and 10th percentile TPM values for all genes for the MCM, UM, and GTEx samples. We then kept all genes from the tumor cohorts that showed sufficient expression, which we defined as having a 10^{th} percentile expression larger than 1. To filter candidate genes for a high-in-tumor, low-in-tissue expression phenotype, we selected genes through a transcriptomic filter so that 90% of the tumor samples showed a higher expression level in TPM than 90% of the tissue samples. Accordingly, this was formally defined as the requirement for the 10^{th} percentile of tumor expression

to be greater than the 90th percentile of tissue expression. We compared all genes in the tumor data against all 51 tissues in the GTEx data set.

Additionally, using the input from medical and scientific collaborators, we annotated the 51 tissues in the GTEx database with their survival criticality (**Table 2**), thereby judging which tissues off-site targets may be tolerated and, conversely, in which they are to be avoided. In parallel to the transcriptomics filter, we relied on histological evidence provided by HPA. We retrieved the database and filtered our gene set by removing every gene annotated to have any expression level of the protein above "Not detected" in the database. In the case of our MCM cohort, the candidate set of genes filtered through all three filtering steps (protein coding, GTEx, and HPA filtering) was further classified into two sets of expected tolerability. We defined a gene as superior tolerable if its expression was less than 10 TPM in all our GTEx-derived survival-critical tissues. If this condition was not true for all critical tissues, the gene was allocated to the enhanced tolerance set of genes since the gene shows residual expression in some survival critical tissues. This additional filtering into two discrete tolerance sets was not performed for the UM cohort's candidate genes. In that case, all genes derived from the previous filtering steps showed no residual expression in healthy tissue and were thus all deemed superior tolerance and highly tumor restricted.

3.4 Processing of candidate genes and their derived peptides

3.4.1 Peptide k-mer extraction and post-hoc screening

Once genes had passed through the filtering steps, extracting peptides of length k ranging from 9 to 12 amino acids from a gene's annotated protein product was necessary. While MHC I molecules can bind peptides with lengths outside this range, it covers the preferentially bound lengths of most alleles and thus simplifies the generation of peptide k-mers (Trolle *et al.*, 2016). To this end, a fasta file of the human proteome was used to retrieve all annotated protein sequences from our gene candidates. We further implemented a simple k-mer extraction algorithm that would take as input a protein's complete AA sequence and return all overlapping k-mer peptides of length 9 to 12 AAs. Given all extracted peptides from our gene set, we screened each peptide against the complementary part of the human proteome (e.g., all non-selected genes), excluding peptides with literal sequence matches.

3.4.2 Decision support ranking of peptides

Since one can derive many peptides of lengths 9 to 12 from a single protein sequence and given that there is a high allele variability in the HLA locus with peptides binding one or more alleles, it is necessary to provide a ranking system to facilitate candidate selection for clinical or experimental use. To this end, a crucial part of our selection pipeline is the decision support system implemented for specific MHC-I-restricted tumor antigens. In the first phase of our project, we implemented a score based on several variables derived from the peptide's source gene or the peptide-allele combination. Further, we extended this score in the second phase using a machine learning and network approach.

3.4.3 Implementation of a continuous multivariate score for MCM

While developing our system on the MCM cohort, we implemented the generalized Predicted Immuno-Efficacy score (gPIE), which evaluates parameters to judge if an epitope is a valid candidate for targeting MCM. The gPIE is constructed from the predicted binding affinity between peptide and HLA allele (f_1), the predicted immunogenicity (f_2), our cohort's median transcript expression of a peptide's transcript of origin (f_3), and an expression index comparing the gene of origin's expression in MCM with its maximum expression in healthy tissue (f_4). Formally, the gPIE is defined as follows:

$$\begin{aligned} gPIE_{Epitope} = 100 \cdot\ & f_1\left(BindAff_{Epitope}\right) \\ \cdot\ & f_2\left(Immunogen_{Epitope}\right) \\ \cdot\ & f_3\left(MelTranscExp_{Peptide}\right) \\ \cdot\ & f_4\left(GeneExpIndex_{Peptide}\right) \end{aligned} \qquad (\boldsymbol{II})$$

The gPIE has a value range from 0 to 100 and is to be interpreted in a higher-is-better manner. Each score element is normalized and restricted to unit distances 0 to 1 and will be explained in more detail in the following subsections.

3.4.3.1 Allele-specific binding affinity prediction

Since a significant determinant for a peptide's efficacy as a targetable antigen is its binding affinity to an MHC-I allele, we included an established allele-specific affinity prediction in our scoring system that predicts the binding affinity. Generally, the binding affinity is expressed as the half-maximal inhibitory concentration (IC50). In general terms, IC50 measures the concentration at which the ligand occupies half of all binding pockets on a given target in the absence of competition. To predict the binding affinity for a given peptide, we first curated a set of 36 HLA alleles (**Table 4**) that are well-characterized and common in the European population (Sanchez-Mazas *et al.*, 2017). We predicted the binding affinity for each allele using netMHCpan 4.0, a machine learning-based affinity predictor commonly used in the field (Jurtz *et al.*, 2017). After prediction, we filtered all peptides with a binding affinity of less than 500 nM as a cut-off value. To integrate the affinity value into the score, we performed a min-max normalization based on the observed total range of all predicted affinities, with a scale inversion due to lower IC50 values representing higher binding affinities. The function f_1 generates a normalized IC50 value per epitope, and the procedure is described in **Equation III** with $BindAff_{Epitope}$ denoting the predicted IC50 of a given epitope to be normalized, $\max(BindAff)$ the maximum and $\min(BindAff)$ the minimum of the set of all predicted IC50 values across all alleles.

Table 4: ***HLA alleles used in the affinity prediction****.*

For each allele and peptide combination, the corresponding IC50 was predicted.

HLA-A	*HLA-B*	*HLA-C*
01:01	07:02	01:02
02:01	08:01	02:02
03:01	13:02	05:01
11:01	15:01	06:02
24:02	18:01	07:01
25:01	27:02	08:02
26:01	35:01	12:03
29:02	41:01	15:02
31:01	44:02	16:01
32:01	44:03	
33:01	45:01	
68:01	49:01	
	50:01	
	51:01	
	57:01	

$$f_1(BindAff_{Epitope}) = \frac{BindAff_{Epitope} - \max(BindAff)}{\min(BindAff) - \max(BindAff)} \quad (\boldsymbol{III})$$

3.4.3.2 Allele-specific immunogenicity prediction

We used a prediction tool published by the Immune Epitope Database and Analysis Resource (IEDB) to evaluate if an epitope has a high probability of immunological activity. The tool evaluates the contribution and positional relevance of amino acids in a peptide to its immunogenicity (Calis *et al.*, 2013). It leverages measured amino acid preferences of T-cell receptors which are supposed to approximate the likelihood of a peptide being recognized by T cells. Additionally, it uses allele-specific anchor positions in the peptides to mask the contribution of said positions in calculating the score. The tool has been validated using viral 9-mers and was used in version 1.1. For the model's output to be used in our multi-variate score, we applied the min-max normalization function f_2 to constrain the score range to the interval between 0 and 1, as described in Equation IV. Here, $Immunog_{Epitope}$ denotes the predicted immunogenicity value of a given epitope while $\max(Immunog)$ and $\min(Immunog)$ represent the maximum and minimum predicted immunogenicity over all epitopes respectively.

$$f_2\left(Immunog_{Epitope}\right) = \frac{Immunog_{Epitope} - \min(Immunog)}{\max(Immunog) - \min(Immunog)} \quad (\boldsymbol{IV})$$

3.4.3.3 Transcript specific expression

Since the RNA abundance of a gene is a generally accepted surrogate for the abundance of a protein product and thus may have a major impact on peptide availability for presentation, we integrated a measure of expression into the score (Conesa *et al.*, 2016). We used transcript-specific expression measured in TPM to estimate the abundance of a peptide's source gene. Since the value range of TPM goes from 0 to 10^6, we took care to cap the influence of high expression values on the gPIE. Hence, we hypothesized that a TPM of 100 would saturate the availability of a peptide for presentation. Thus, the integration function f_3 for transcript-specific expression into the gPIE score is a piece-wise equation designed so that all normalized expression values were constrained to 0 to 100. In which $TranscExp_{Peptide}$ describes the TPM expression of a transcript that may give rise to the protein and subsequently the peptide.

$$f_3\left(TranscExp_{Peptide}\right) = \begin{cases} \dfrac{TranscExp_{Peptide}}{100}, & TranscExp_{Peptide} < 100 \\ 1, & TranscExp_{Peptide} \geq 100 \end{cases} \quad (\boldsymbol{V})$$

3.4.3.4 Gene expression index

Since our goal is to select peptides that are restricted to the tumor as much as possible, we designed a metric to penalize low differences between a peptide's source gene's expression in the tumor and healthy tissues. This metric, deemed the gene expression index, is generated by the function f_4. It evaluates the 10th percentile of a gene's expression, denoted as $MelGeneExp_{Peptide}$, in the MCM samples compared to its highest 90th percentile expression in healthy tissue, described by the term $TissueGeneExp_{Peptide}$. This comparison has the effect that this parameter gets lower the closer the expression values are, thus penalizing a low expression difference. The gene expression index was calculated as follows:

$$f_4\left(GeneExpInd_{Peptide}\right) = \frac{10^{th} pct\ MelGeneExp_{Peptide}}{10^{th}\ pct\ MelGeneExp_{Peptide} + \max\left(90^{th} pct\ TissueGeneExp_{Peptide}\right)} \quad (\boldsymbol{VI})$$

3.4.4 Extension of the multivariate score into an ensemble model

During the first phase of this project, we derived a multi-parameter decision support ranking for helping to find epitope candidates for the treatment of MCM. We expanded this ranking approach by implementing a more complex scoring system in the second phase. This novel system should evaluate the biological necessity of a candidate antigen to circumvent or estimate antigen-loss likelihood in the tumor and provide a more generalizable prediction of binding and immunogenicity probability.

To this end, we conceived the extended generalized predicted immuno-Efficacy Score (ES). It was designed to be broadly applicable to different tumor entities and thus uses gene-level expression values, which are available through anonymized data portals like The Cancer Genome Atlas Program (TGCA). We implemented this approach to generate new TAA candidates for metastasized primary uveal melanoma (UM) and provide novel therapeutic possibilities for further testing. The fundamental pillars of the ES are the extension of our ranking function by two own-trained machine learning models (ML) and one network model, assessing biological functionality. The two ML models were designed to provide generalized probability predictions for binding to MHC-I and immunogenic activity. Also, a network-based method was implemented that constructs a gene-specific biological indispensability metric to minimize the probability of the tumor losing the antigen due to immune escape. We retained the predicted binding affinity in the ES to supply an allele-specific metric. Formally, we define the ES as follows:

$$ES(P) = consTME\big(gene(P)\big) \cdot \mathrm{consIC50}\big(epitope(P)\big) \cdot Idspx\big(gene(P)\big) \cdot gBP(P) \cdot gAP(P) \ (\boldsymbol{VII})$$

Similarly to the gPIE mentioned before (section 3.4.3), each parameter in the score is constrained to a range between 0 and 1 (also called unit distance) to ensure an unweighted contribution. In this formula, *P* denotes a peptide with *gene(P)* indicating the corresponding gene of origin. Accordingly, an *epitope(P)* denotes the combination of *P* with a specific HLA allele. The individual subfunctions of the ES are the constrained tumor median expression *consTME*, the indispensability index (Idspx), the generalized binding predictor gBP, the generalized activity predictor gAP, and the constrained allele-specific binding affinity *consIC50*. Firstly, the *consTME* of a gene *G* was calculated as follows:

$$consTME(G) = \begin{cases} 1 \ if Expr(G) > 100 \\ \dfrac{1}{100} \cdot Expr(G) otherwise \end{cases} \ (\boldsymbol{VIII})$$

We followed the same previously described (section 3.4.3.3) rationale by capping the upper bound for a gene's TPM at 100, thus ensuring that very high expression alone would not immediately lead to a higher score. The subfunction consIC50 was calculated in a similar manner. The variables *G* and *E* denote *gene(P)* and *epitope(P)*, respectively, with *Expr(G)* denoting *G*'s RNA abundance. *IC50(E)* denotes the allele-specific binding affinity predicted with netMHCpan 4.0 (Jurtz *et al.*, 2017). Binding affinity upper and lower bounds, 2000nM and 30 nM, respectively, were derived from a logistic regression applied to the training dataset of the machine learning model after annotating them with predicted IC50 values.

$$\mathrm{consIC50}(E) = \begin{cases} 0 \; if \; \mathrm{IC50}(E) > 2000nM \\ 1 \; if \; \mathrm{IC50}(E) < 30nM \\ \dfrac{L\big(\mathrm{IC50}(E)\big) - \big(L(2000\ \mathrm{nM})\big)}{L(30\ \mathrm{nM}) - \big(L(2000\ \mathrm{nM})\big)} \text{ with } L(x) = -log_{10}(x) otherwise \end{cases} \quad (\boldsymbol{IX})$$

The bounds were selected to obtain a high positive predictive value and to discard presumptive non-binders reliably. The additional parameters Idspx, gBP, and gAP were derived from a network analysis and machine learning model and are already returning unit distance probabilities.

3.4.4.1 Physiochemical annotation for peptides

With the ES score including two ML models, we first used properties directly derivable from the AA sequence of a peptide kmer as a feature. We selected properties that we deemed suitable encodings of a peptide's biological characteristics or are known to play a role in binding to MHC-I (Altuvia *et al.*, 1994; Huang, Kuhls and Eisenlohr, 2011; Chowell *et al.*, 2015). The features used were molecular weight in Dalton, instability index according to dipeptide occurrence (Guruprasad, Reddy and Pandit, 1990), isoelectric point, grand average of the hydropathy index (GRAVY) according to Kyte and Doolittle (Kyte and Doolittle, 1982) and a polarity score. For all features but the polarity score, we used the Python library 'Biopython' and its included function 'ProtParam' to derive the values (Peter J A Cock *et al.*, 2009). The polarity score was calculated as the average of the AA chain. Raw AA polarity values were derived from Zimmerman et al. (Zimmerman, Eliezer and Simha, 1968) and can be found in **Table 5**.

Table 5: ***Amino acid polarity values derived from literature and used in the manual computation of the polarity score*** **(Zimmerman, Eliezer and Simha, 1968).** *As a reference, the 1-letter symbol and IUPAC names of the amino acids are listed here since in the supplementary code used for the computation, only 1-letter notation is used.*

Aminoacid 1-letter symbol	Aminoacid	Polarity values
A	Alanine	0.00
R	Arginine	52.00
D	Aspartate	49.70
N	Asparagine	3.38
C	Cysteine	1.48
E	Glutamate	49.90
Q	Glutamine	3.53
G	Glycine	0.00
H	Histidine	51.60
L	Leucine	0.13
I	Isoleucine	0.13
K	Lysine	49.50
M	Methionine	1.43
F	Phenylalanine	0.35
P	Proline	1.58
S	Serine	1.67
T	Threonine	1.66
W	Tryptophan	2.10
Y	Tyrosine	1.61
V	Valine	0.13

3.4.4.2 Generalized binding and activity probability prediction

One core element of our novel approach for predicting cancer-associated antigens and their presumptive targetable epitopes was implementing an allele-independent measure of peptide immunogenicity. We constructed two random forest (RF) models to predict which peptides have a high chance of generalized MHC-I binding (gBP) and of eliciting an immune response (gAP) (Breiman, 2001). The RF model was selected for its ability to work well in settings where the number of variables is far larger than the number of observations (Boulesteix *et al.*, 2012). Our two models were implemented in R with the library '*randomForest'* version 4.7-1.1 and designed to accept a peptide's physiochemical properties as features. MHC-I-restricted training peptides for both models were extracted from the MHCBN database 4.0 by selecting only peptides with unambiguous (i.e., yes or no) classification for binding and activity, respectively (Lata, Bhasin and Raghava, 2009). The training set of 3777 entries for binding was supplemented with 201 peptides classified as binders or non-binders through crystallography in the Protein Data Bank (PDB) and literature research to bolster our data basis. Since we expected the distribution of predicted peptides to skew towards non-binders heavily, as most peptides generated by cells will not be presented on MHC-I, we modeled the distribution in the training data such that the distribution of binders to non-binders would be 1:10 (Yewdell, Reits and Neefjes, 2003). Since this led to a reduction in the input size of the training data, we decided to construct an ensemble model, which would repeatedly perform weighted sampling from the total training data. Using 100 iterations of this weighted sampling approach, we trained a model with 10,000 trees in each iteration. Accordingly, for the generalized probability of activity, we performed balanced sampling (active to non-active, 1:1) since, in theory, any peptide should be able to elicit an immune response given a complementary TCR and the right environment. For both models, response was discretized at the decision threshold of 0.5. Predictive performance evaluation was performed by resampling from the entire dataset and comparison to published alternatives (**Figure 6**).

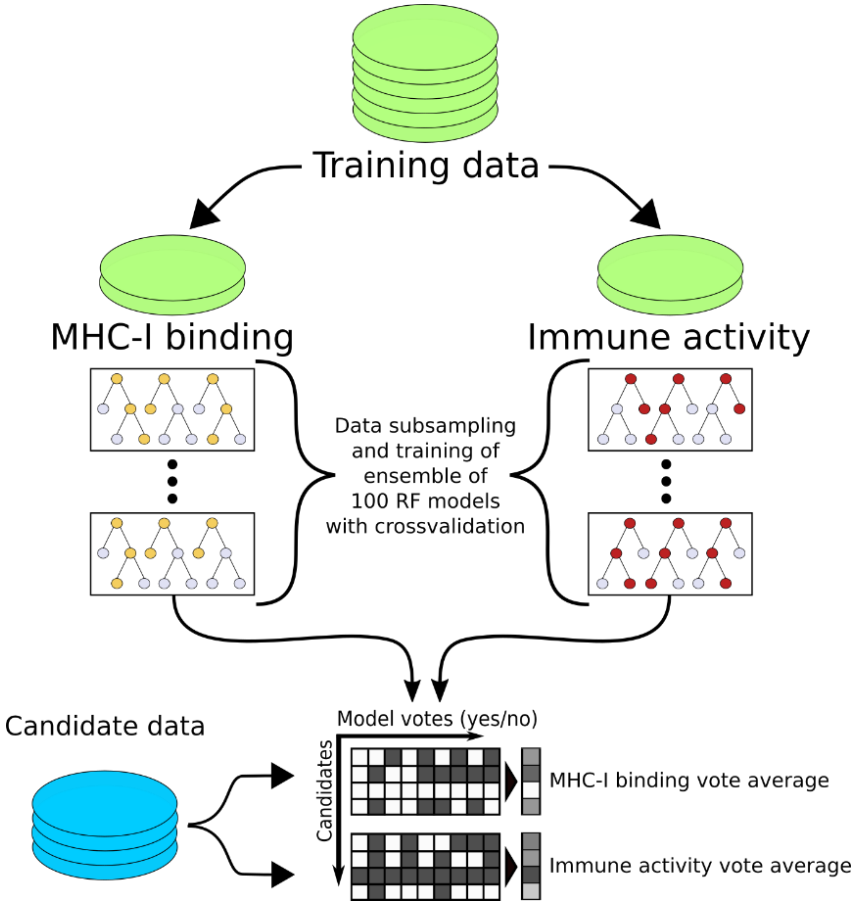


Figure 6: ***Detailed illustration of the ensemble model approach used for generating generalized binding and activity predictions.*** *The training data was used to construct two ensemble random forest models, one for the probability of binding to MHC-I and one for eliciting an immune response. For each condition, we trained 100 models with 10,000 trees each while sampling training data in a weighted manner for binding prediction and a balanced manner for activity. Weighted sampling was done to emulate the heavy skew towards non-binding peptides expected and observed in empirical data. Thus, we applied a 1:10 ratio. This also had the effect of heavily biasing our models towards a high positive predictive value per model at the cost of the type II error rate.*

3.4.4.3 A network model for indispensability estimation

Antigen loss, and the subsequent immune evasion by a tumor, is a problem we tried to address by creating a network model whose primary task is to quantify the indispensability of a gene for the tumor entity. First, a list of 90 cancer-relevant gene ontology (GO) terms (**Table 6**) was gathered through subject-expert level curation. We calculated the sum of associations to these terms for each gene to estimate its intrinsic importance. Additionally, we counted the occurrence of a gene in four cancer biology databases, Oncogene, the Cancer Proteomic Database (http://apoptoproteomics.uio.no/), the Epithelial-Mesenchymal Transition Gene Database, and DriverDBv3 (Liu, Sun and Zhao, 2017; Liu *et al.*, 2020). Occurrence in the GO term list and occurrence in the cancer biology databases was summed to determine an individual gene's importance (*GI*). These calculations were performed for our candidate genes derived from our pipeline and genes annotated in DriverDBv3 to establish a generalized distribution of gene importance independent of the tumor model of interest. Since proteins are generally embedded into a biological network and act within pathways, we implemented a gene's indispensability estimate, which is assumed to be higher when its loss would influence other genes of high biological relevance. Using in-house software, we reconstructed an interaction network from our candidate genes and all genes found in DriverDBv3 and expanded it with direct interaction partners extracted from the databases TRANSFAC, HTRIdb, miRecords, and miRTarBase (Matys *et al.*, 2006; Xiao *et al.*, 2009; Bovolenta, Acencio and Lemke, 2012; Chou *et al.*, 2016).

Each node or gene *G* in the network was assigned a neighborhood importance (*NI*) calculated as the sum of its own and its direct interaction partner's gene importance *(GI)*, defined as:

$$NI_i(G_i) = GI(G_i) + \sum GI(G_j)\ for\ all\ G_j\ \ where\ adjency\ to\ G_i = 1\ (\boldsymbol{X})$$

The underlying assumption is that the higher the NI, the more relevant the gene is for the tumor's survival. We further assumed that beyond some threshold of NI, there is no actual increase of importance for a gene (**Figure 7**). Hence, following this reasoning, all genes whose NI is more significant or equal to this threshold are assumed to be equally important for the cancer cell.

Additionally, for practical purposes in modeling, it is sensible to constrain outlier values so as not to skew the model too harshly in this direction. Accordingly, we transformed our network's empirical distribution of NI values to a distribution with saturation properties. Since Michaelis-Menten functions lend themselves to model saturation characteristics, we derived a Michaelis-Menten type function which defines the threshold of not gaining more importance for a gene as 90% of the maximum. To derive the numerical value of the threshold (*t*), we multiplied two cutoffs derived from the empirical distributions of the variables of NI. The

variables are the node degree (the node's number of neighbors) and the gene importance (the summed-up occurrences in cancer-related databases).

To do this, we first had to set an arbitrary threshold for what we would define as “sufficiently important”. We assumed this threshold to be at five occurrences. Then, we plugged the value five into the gene importance’s cumulative distribution function *F* to obtain the value. After establishing this, we used the empirical cumulative probability *p* to derive the node degree distribution's corresponding *p*-quantile (indicated by function *Q* in **Equation XII**). This value corresponds to the quantile *Q* in the distribution of node degree values at which the gene is also sufficiently important in the gene importance table. The threshold parameter t was calculated as the product of five and *Q*.

$$Threshold\ t\ =\ GI_{Sufficiently\ important\ gene}\ \cdot\ Node\ degree_{\ For\ genes\ sufficiently\ important} \quad \textbf{(XI)}$$

$$K_M = \frac{1-0.9}{0.9} \cdot t = \frac{1-0.9}{0.9} \cdot \Big(5 \cdot Q_{\ Node\ degree}\big(p = F_{GI}(5)\big)\Big) \quad \textbf{(XII)}$$

The interval-constrained neighborhood importance (NI) value of a gene G with saturation behavior was labeled the indispensability index.

$$Indispensability(G) = \frac{Neighborhood\ importance(G)}{K_M + Neighborhood\ importance(G)} \quad \textbf{(XIII)}$$

With this method, we set out to estimate the biological importance value of a gene for the tumor entity and ensure that if the gene is targeted, the possibility of evasion is minimal while the disruption of the cancer phenotype is maximal due to the high potential cost of losing this gene.

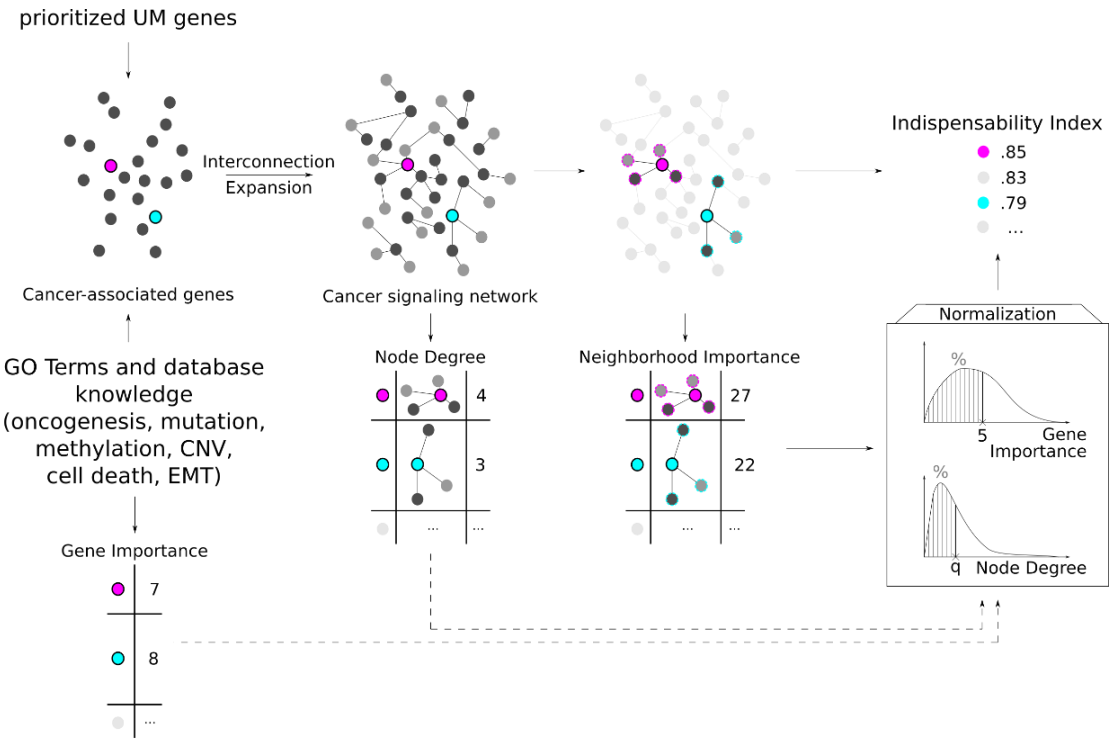


Figure 7: ***Process for the generation of an indispensability estimate for a candidate gene in the context of it getting targeted during therapy.*** *After filtering procedures, a candidate list is supplied to the algorithm. Genes are characterized in terms of their connectivity and importance in an expanded gene signaling network and the sum of their occurrences in cancer-related databases. The neighborhood importance of a gene is the sum of these values for all its direct neighbors.*

Table 6: ***Selected gene ontology (GO) terms from which gene importance values were derived.***

GOID	*TERM*
GO:0001569	branching involved in blood vessel morphogenesis
GO:0001833	inner cell mass cell proliferation
GO:0001834	trophectodermal cell proliferation
GO:0002040	sprouting angiogenesis
GO:0002041	intussusceptive angiogenesis
GO:0002174	mammary stem cell proliferation
GO:0002674	negative regulation of acute inflammatory response
GO:0002677	negative regulation of chronic inflammatory response
GO:0002862	negative regulation of inflammatory response to antigenic stimulus
GO:0002941	synoviocyte proliferation
GO:0003347	epicardial cell to mesenchymal cell transition
GO:0003419	growth plate cartilage chondrocyte proliferation
GO:0006925	inflammatory cell apoptotic process
GO:0008284	positive regulation of cell population proliferation
GO:0008285	negative regulation of cell population proliferation
GO:0008637	apoptotic mitochondrial changes
GO:0010463	mesenchymal cell proliferation
GO:0010657	muscle cell apoptotic process
GO:0010717	regulation of epithelial to mesenchymal transition
GO:0010718	positive regulation of epithelial to mesenchymal transition
GO:0010719	negative regulation of epithelial to mesenchymal transition
GO:0014009	glial cell proliferation
GO:0014029	neural crest formation
GO:0016525	negative regulation of angiogenesis
GO:0033002	muscle cell proliferation
GO:0033028	myeloid cell apoptotic process
GO:0033687	osteoblast proliferation
GO:0034349	glial cell apoptotic process
GO:0035172	hemocyte proliferation
GO:0035492	negative regulation of leukotriene production involved in inflammatory response
GO:0035726	common myeloid progenitor cell proliferation
GO:0035736	cell proliferation involved in compound eye morphogenesis
GO:0035988	chondrocyte proliferation
GO:0036093	germ cell proliferation
GO:0042127	regulation of cell population proliferation
GO:0042981	regulation of apoptotic process
GO:0043065	positive regulation of apoptotic process
GO:0043066	negative regulation of apoptotic process
GO:0043276	anoikis
GO:0044340	canonical Wnt signaling pathway involved in regulation of cell proliferation
GO:0044346	fibroblast apoptotic process
GO:0045765	regulation of angiogenesis

GO:0045766	positive regulation of angiogenesis
GO:0048134	germ-line cyst formation
GO:0048144	fibroblast proliferation
GO:0050673	epithelial cell proliferation
GO:0051402	neuron apoptotic process
GO:0051450	myoblast proliferation
GO:0060055	angiogenesis involved in wound healing
GO:0060266	negative regulation of respiratory burst involved in inflammatory response
GO:0060317	cardiac epithelial to mesenchymal transition
GO:0060722	cell proliferation involved in embryonic placenta development
GO:0060809	mesodermal to mesenchymal transition involved in gastrulation
GO:0060886	clearance of cells from fusion plate by epithelial to mesenchymal transition
GO:0060978	angiogenesis involved in coronary vascular morphogenesis
GO:0061323	cell proliferation involved in heart morphogenesis
GO:0061351	neural precursor cell proliferation
GO:0070341	fat cell proliferation
GO:0070661	leukocyte proliferation
GO:0071335	hair follicle cell proliferation
GO:0071838	cell proliferation in bone marrow
GO:0071839	apoptotic process in bone marrow cell
GO:0071887	leukocyte apoptotic process
GO:0072089	stem cell proliferation
GO:0072104	glomerular capillary formation
GO:0072111	cell proliferation involved in kidney development
GO:0090255	cell proliferation involved in imaginal disc-derived wing morphogenesis
GO:0097152	mesenchymal cell apoptotic process
GO:0097190	apoptotic signaling pathway
GO:0097194	execution phase of apoptosis
GO:0097360	chorionic trophoblast cell proliferation
GO:0106015	negative regulation of inflammatory response to wounding
GO:0140208	apoptotic process in response to mitochondrial fragmentation
GO:0150079	negative regulation of neuroinflammatory response
GO:1900016	negative regulation of cytokine production involved in inflammatory response
GO:1902362	melanocyte apoptotic process
GO:1902489	hepatoblast apoptotic process
GO:1902742	apoptotic process involved in development
GO:1903594	negative regulation of histamine secretion by mast cell
GO:1904019	epithelial cell apoptotic process
GO:1904516	myofibroblast cell apoptotic process
GO:1904606	fat cell apoptotic process
GO:1990009	retinal cell apoptotic process
GO:1990654	sebum secreting cell proliferation
GO:2000793	cell proliferation involved in heart valve development

3.4.5 Aggregation of results into an accessible database

Since we envision that our predictions could be used for further *in vitro* or pre-clinical experiments by thirds parties, we designed and deployed a publicly accessible database, developed in RStudio, a development platform for the language R and using the Shiny framework (RStudio, 2011; Chang *et al.*, 2023). Martin Eberhardt, at the Laboratory of Systems Tumor Immunology at the Department of Dermatology, University Hospital Erlangen, has designed the front end of the database and performed deployment and management of web services and domain services. The databases are available at www.curatopes.com.

3.5 Validation procedures for a subset of selected antigen candidates

3.5.1 Candidate selection for experimental validation

We selected three distinct groups of peptides from our pipeline's output for validation since we had to constrain the number of experiments to perform. The chosen groups were labeled the high efficacy (HE), low efficacy (LE), and alternative predictor (AP), respectively. The high-efficacy (HE) group contained the highest-ranked peptides, which by our score, we deemed to be the best-suited for therapy.

To maximize donor availability for experimental validation, we selected peptides with high efficacy scores for the locally prevalent HLA allele A*02:01 (abbreviated A2) as follows: For each scored peptide, we first assessed its potential to engage bystander alleles, i.e., any of the other 35 considered alleles beyond A2 (**Table 4**). This was done since donors are rarely fully HLA-typed in practice, and we had to account for possible presentation by other alleles, which could add noise to the readout. The probability of binding at least one bystander allele by the peptide, thus forming an epitope, was estimated by interpreting an epitope's (*Ep*) efficacy scores (ES) for a bystander allele, denoted as *b*, as probabilities of success and calculating the probability of at least one success across all bystander alleles, i.e., the non-failure probability P_{NF} according to **Equation XIV.**

$$\mathrm{P}_{NF} = 1 - \prod_{b \in bystander} (1 - p_b), \qquad with\ p_b = \mathrm{ES}(Ep_b) \ (\mathbf{XIV})$$

Further, the maximum bystander efficacy scores per peptide were calculated. Peptides were then ranked by multi-sorting their attributes in the following order: bystander non-failure probability (ascending), maximum bystander efficacy scores (ascending), and A2 efficacy score (descending). This multi-sort step prioritizes discarding peptides with undesirable binding to bystander HLA alleles. In a subsequent step coined Levenshtein filter, we ensured that no two peptides in the final selection were highly similar when considering pure AA sequence similarity. For this, we sequentially discarded peptides whose sequences differed by only one AA substitution, insertion, or deletion (i.e., showed a Levenshtein string distance of 1) from a peptide of higher ES ranking (Levenshtein, 1966). The top 20 peptides from the remaining list made up the HE tier.

Low-efficacy (LE) peptides had minimal efficacy scores across all 36 considered HLA alleles. Since we found many peptides with a score of zero, the LE peptide order was randomized, ensuring an unbiased selection of the many peptides with efficacy scores of zero across all alleles. Then, for each peptide, the product, maximum, and mean of its scores across all 36 alleles were calculated, and the amount of its non-zero scores was counted. A subsequent all-ascending multi-sort on these four features ranked the peptides in order of increasing efficacy profile. After applying a Levenshtein filter described above, the top twenty peptides were selected for the LE tier. The 20 peptides in the alternative-predictor (AP) tier served as theoretically efficacious

counterparts to the HE peptides to examine how our selection pipeline performs compared to established methods. Our methodology chose them from the set of peptides assigned an A2 efficacy score of zero. After applying the Levenshtein filter, the AP tier was filled by selecting the 20 peptides with the closest marginally better IC50 value predictions to the HE tier.

3.5.2 An alternative *in silico* testing through molecular docking

As a complementary *in silico* methodology to cross-evaluate our subset of selected candidate peptides, molecular docking simulations were performed by the Department of Systems Biology and Bioinformatics of the University Rostock in the Gupta Group. Independent structure prediction of the peptides was performed using the *Build and Edit Protein* tool in Discovery Studio 2020 software suite (DS2020) on the amino acid sequence. The generated 3D structure was subjected to geometry optimization using smart minimization algorithm for 5000 steps with the CHARMM force field (Brooks *et al.*, 2009). The minimization cutoff was set to 0.001 root mean square gradient in a Generalized Born implicit solvent model. The 3D structure of the HLA-A*02:01 protein was obtained from Protein Data Bank entry 5YXN, which contains a T-cell receptor in complex with HLA-A*02:01 and a hepatitis-C virus peptide. The HLA-A*02:01 alpha chain was extracted and corrected for possible errors, including missing atoms in incomplete residues, missing loop regions, alternate conformations (disorder), nonstandard atom names, and incorrect protonation state of titratable residues with DS2020's *Prepare Protein*. The 3D conformation of the HLA-A*02:01 alpha chain was subsequently optimized using the same protocol mentioned before. All the epitope poses were further refined using RDOCK, a CHARMM-based procedure for refinement and scoring. For each peptide, the ten best binding poses were generated using ZDOCK and further refined with RDOCK for selecting the best pose by the E_RDock score.

3.5.3 *In vitro* validation using autologous PBMC stimulation

To test our prediction algorithm in an *in vitro* environment, we designed an experimental setup that would test our three selected tiers of peptides in a controllable and blind setting. In accordance with our collaborators, a testing strategy was devised to use donor-derived peripheral blood mononuclear cells (PBMCs) and stimulate them with our peptide groups (**Figure 8**). Each group, HE, LE, and AP, were divided into four pools of five peptides and supplied without disclosure of the tier (blind testing) to our experimental collaborators. *In vitro* validation assays were performed by Dr. Cindy Flamann in the group of PD Heiko Bruns at the Universitätsklinikum Erlangen, Department of Hematology.

Leukapheresis products were obtained from four donors selected for their positive CMV and HLA-A*02:01 status while adhering to current regulatory and ethical standards, including obtaining informed consent. PBMCs were purified by Ficoll gradient centrifugation (800 g, 20 min, 20 °C, break off) and subsequently

cryopreserved in liquid nitrogen at a concentration of 100 million/ml in a freezing medium containing 10% DMSO. After thawing, 1 – 2 million/ml cells were recovered for 18-24 h in serum-free TexMACS™ GMP medium (Miltenyi Biotec, Bergisch-Gladbach, Germany) at 37 °C. Before peptide stimulation, cells were harvested by centrifugation and counted. Batches of 20 million live PBMCs were stimulated per peptide pool or CMV positive control (human PepTivator® CMV pp65, Miltenyi Biotec, Bergisch-Gladbach, Germany) at a total peptide concentration of 1 µg/ml. Stimulation was performed in 20 ml of prewarmed serum-free medium for two hours at 37 °C. Afterward, cells were spun down and washed with a medium to eliminate unbound peptides. Cells were then incubated at an initial concentration of 2 million/ml for nine days at 37 °C in RPMI 1640 medium (Gibco by Life Technologies GmbH, Darmstadt, Germany) supplemented with 1% (v/v) GlutaMAX (Gibco by Life Technologies GmbH, Darmstadt, Germany), 50 IU/ml IL-2 (Aldesleukin, Novartis Pharma GmbH, Nürnberg, Germany) and 1% (v/v) human AB serum (Anprotec, Bruckberg, Germany). During day 5 of incubation, culture volume was increased with fresh RPMI 1640 medium with supplements to a total of 2.5 times the volume on day 0. On day nine after stimulation, culture supernatant was used for IFN-γ ELISA (ELISA MAX™ Deluxe Set, Biolegend, San Diego, USA) and stimulated PBMCs investigated with IFN-γ Secretion Assay (Miltenyi Biotec, Bergisch-Gladbach, Germany) as well as Incucyte® Live Cell Imaging (Sartorius, Göttingen, Germany) according to the manufacturer's instructions. The HLA-A*02:01-positive UM cell line 92.1 (De Waard-Siebinga *et al.*, 1995) was selected as a cytotoxicity target and cultivated in uveal melanoma medium containing RPMI1640 (Gibco by Life Technologies GmbH, Darmstadt, Germany), 2 mM L-glutamine (Gibco by Life Technologies GmbH, Darmstadt, Germany), 10% fetal bovine serum (Merck, Darmstadt, Germany), and 1x Antibiotics-Antimycotics (Gibco by Life Technologies GmbH, Darmstadt, Germany) at 37 °C with 5% CO_2. The 92.1 cells were stained with 0.75 µM Cytolight Green (Sartorius, Göttingen, Germany) in PBS for 20 min at 37 °C before the Cytotox Assay. After two washing cycles, stained 92.1 cells were seeded in a 96-well plate and incubated for 30 min at 37 °C to allow for reattachment. Peptide-stimulated PBMCs were then added in an effector:target-ratio of 4:1 (final volume 200 µl) and the culture medium supplemented with Annexin V Red Dye (Sartorius, Göttingen, Germany) to facilitate ongoing staining of apoptotic cells. Green and red fluorescence channels were recorded once every 60 min for a total of 45 hours, and the colocalization of the green and red area (µm²/Image) was automatically evaluated.

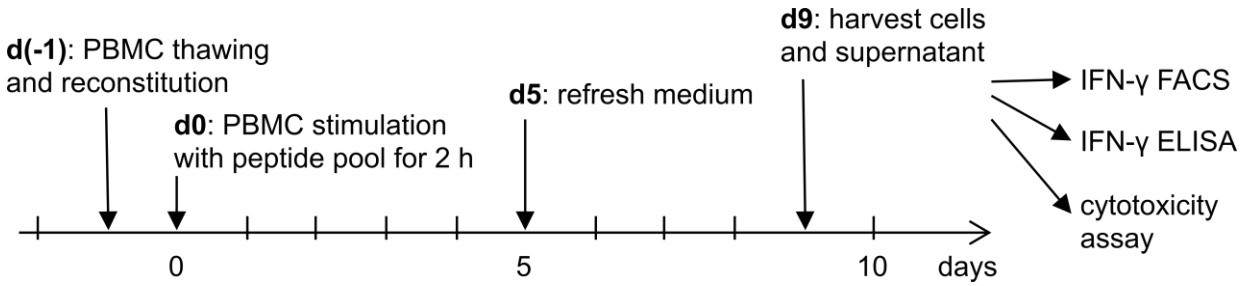


Figure 8: ***Schematic of the experimental procedures conducted for validation.***

4 Results

In this project, we created a computational selection pipeline for novel self-tolerant non-mutated therapeutically viable MHC-I-restricted tumor-associated antigens for metastatic cutaneous melanoma (MCM) and metastasized primary uveal melanoma (UM). To this end, we first conceptualized the underlying self-tolerance model so that gene expression and transcript expression were used to filter the potential list of genes to a set that we would deem immune tolerable in healthy and/or survival-critical tissues. Secondly, since the first iteration of this methodology lacked generalizability and an evaluation of the biological relevance of an antigen, we endeavored to improve these aspects utilizing the second tumor model, UM, as a case study. Since we envision the translation of our antigens into treatment options, we set up *in vitro* validation assays in collaboration with experts in these experimental procedures.

4.1 Tumor-associated antigens in metastatic cutaneous melanoma

In the first phase of this project, we focused on creating a comprehensive filtering procedure that would allow us to find tumor-restricted genes from transcriptomics data quickly. We created this workflow for the tumor model of metastatic cutaneous melanoma (MCM), which, although having improved prognosis through ICB, still needs treatment alternatives for patients non-responsive to ICB.

4.1.1 Transcriptomics-based gene filtering procedures

The first levels in establishing a set of candidate genes for tumor-associated antigens in MCM were comparative filtering procedures to identify genes that were only expressed in the tumor model. We designed our pipeline to be highly restrictive to ensure a high degree of tumor exclusivity, with only a few genes fulfilling the general conditions (**Figure 9**). Using the transcriptomics data extracted from the published sample cohorts GSE78220 and GSE96619, we processed the samples as described in section 3.1 and combined them into one dataset containing 58,368 genes. We applied our transcriptomics-based filtering steps against the 31-sample strong MCM cohort's transcriptomic data to select potential tumor-restricted genes.

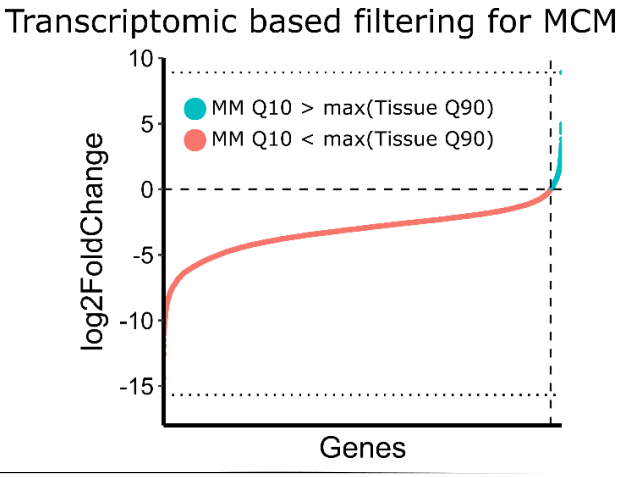


Figure 9: ***Transcriptomics filter for metastatic cutanous melanoma (MCM).*** *To illustrate the restrictiveness of our TPM-based filtering procedure, the log2 fold change of the 10th percentile Tumor expression against the 90th percentile maximum tissue expression is shown. The dotted horizontal lines represent the maximum positive fold change (tumor expresses the gene higher) and maximum negative fold change (tissue expresses the gene higher), respectively, while the dashed horizontal line indicates parity in expression. Only 317 genes show a desirable expression profile. This amount represents less than 1% of the initial 44,334 overall expressed genes.*

In the first step, we removed 24,117 genes from the total set of 58,368 that were not annotated as protein-coding in any of the three databases (CCDS, HPA, Ensembl). We then pruned this set further to remove genes with very low expression in the tumor by discarding genes with a 10^{th} expression percentile smaller than 1; this led to the removal of 13,486 genes and left us with 10,631. Since transcript expression level does not always translate to protein expression levels, we applied a histopathological filter, removing all genes whose protein presence was detected in histological screenings through the Human Protein Atlas (HPA). Hence, we removed another 8,663 genes from this set, concluding these filtering steps with 1,968 genes. Using the Genotype-Tissue Expression (GTEx) sequencing data, we intersected the remaining gene set with genes expressed in healthy tissue. This step generated an overlap between the two datasets to ensure a reference basis for further analysis. This overlap consisted of 1,893 genes that were present in both datasets. As we advanced with these genes, we applied our high-in-tumor, low-in-tissue filter, which removed all genes whose 10^{th} expression percentile in the tumor was smaller than the 90^{th} expression percentile in any healthy tissue, yielding 40 genes. To further narrow down this selection, we decided to group these genes into tolerability sets to reduce the risk of severe autoimmunity. We defined a gene as tolerable in each tissue when 90% of the tissue samples would express the gene at a TPM of less than 10. Genes that were tolerable in all defined critical tissues (**Table 2**) were collected into the superior-tolerance set, while the others were collected into the enhanced tolerance set. Each set was composed of 20 genes (**Figure 10**). Our strict pipeline filters a large set of genes to a condensed version for further investigation. While we described an iterative filtering process here, all individual steps are independent of each other, and changing the order of operations will not yield a different result.

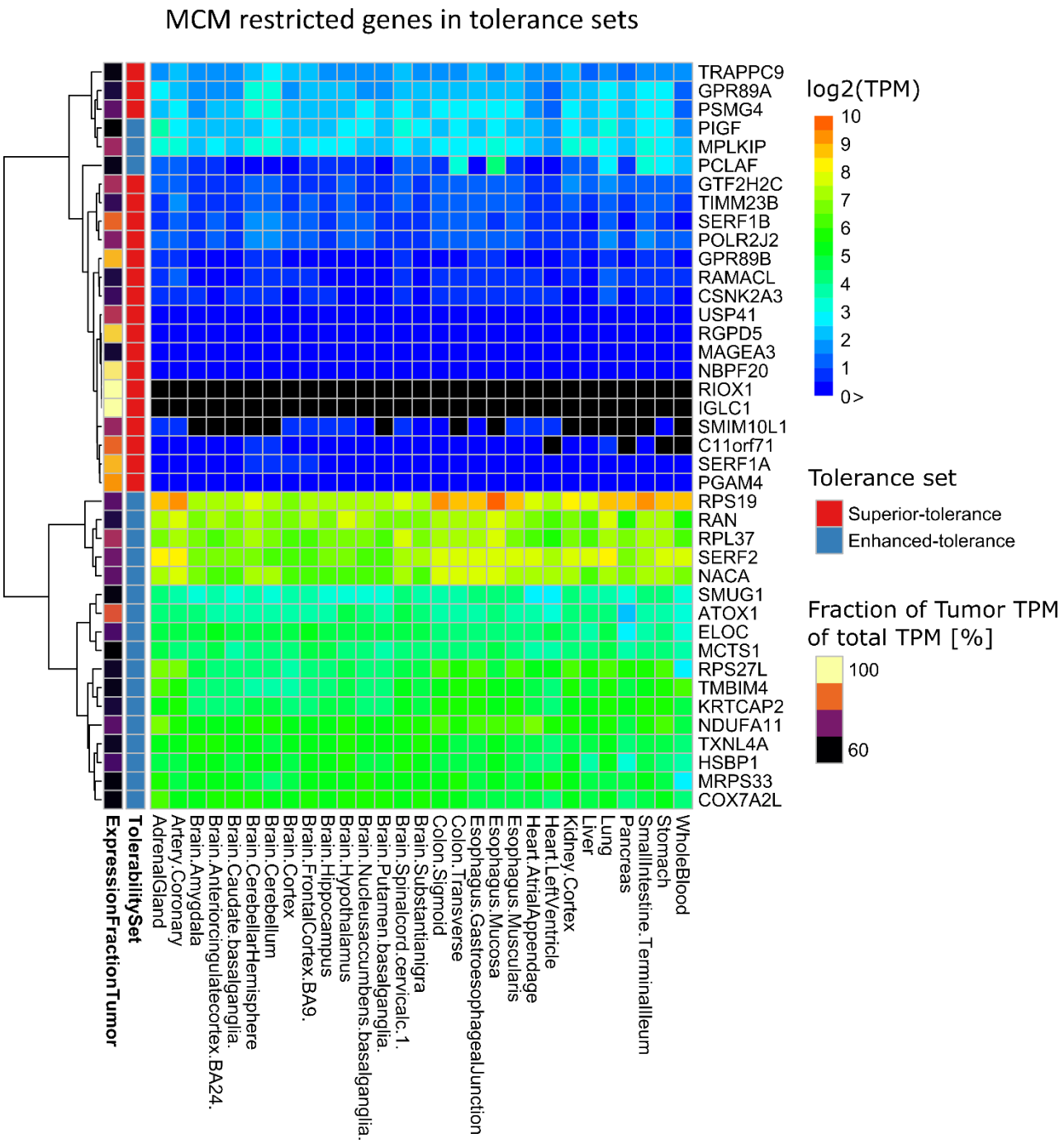


Figure 10: ***Expression of genes in the superior- and enhanced-tolerance sets in the 29 critical tissues****. Expression is shown as log2 of the gene's TPM value with values at exactly zero (before log transformation) in black. Additionally, these values have been excluded from the log transform. Further, the fraction of tumor expression is shown in percent. Genes with an expression of precisely 0 (before log transformation) in peripheral tissue reach a value of 100% here, meaning the tumor contributed exclusively to the expression of this gene.*

4.1.2 Candidate gene-derived peptide sequence-based post-hoc screening

With our predicted tolerant 40 candidate genes established, we continued with peptide-level post-hoc screening (section 3.4.1). Since the antigenic peptide bound to the MHC-I molecule is the primary determinant of CTL engagement, we began sequence level-filtering procedures. Since it is possible for homologies between proteins to lead to identical subsequences originating from different genes, we collected the 165 protein-coding transcripts originating from our 40 candidate genes and extracted peptides of lengths 9 to 12. In total, from all 20 genes in the superior-tolerance set, we generated 31,733 unique peptides. For the Enhanced-tolerance set, we generated 22,793 peptides, for a total of 54,526 peptides. Each peptide was compared in a literal sense, AA by AA, to all known complementary sequences in the human proteome, excluding the protein sequences derived from our selected 40 genes. For the Superior-tolerance set out of 31,733 peptides, 17,670 failed this filtering procedure, while 14,063 passed.

Additionally, five Superior-tolerance genes ultimately failed literal comparison since they showed high redundancy in the proteome. Accordingly, we established that of the 22,973 peptides for the enhanced tolerance set, 20,164 passed filtering procedures. All 20 genes passed this filtering procedure (**Figure 11**). The five superior-tolerance genes removed during peptide-level filtering were *RGPD5*, *SERF1A*, *SERF1B*, *POLR2J2*, and *IGLC1*. Notably, the expression of these genes was not extraordinarily high in the tumor, with 1.77, 7.99, 11.02, 5.90, and 1.17, respectively. However, since expression does not linearly translate into immunogenicity or presence on MHC-I, filtering out genes that show a high sequence identity with other genes is still relevant to avoid potential side effects. Straightforward sequence comparison may thus present an appropriate systematic approach to filter antigen candidates to minimize the chance for cross-reactivity.

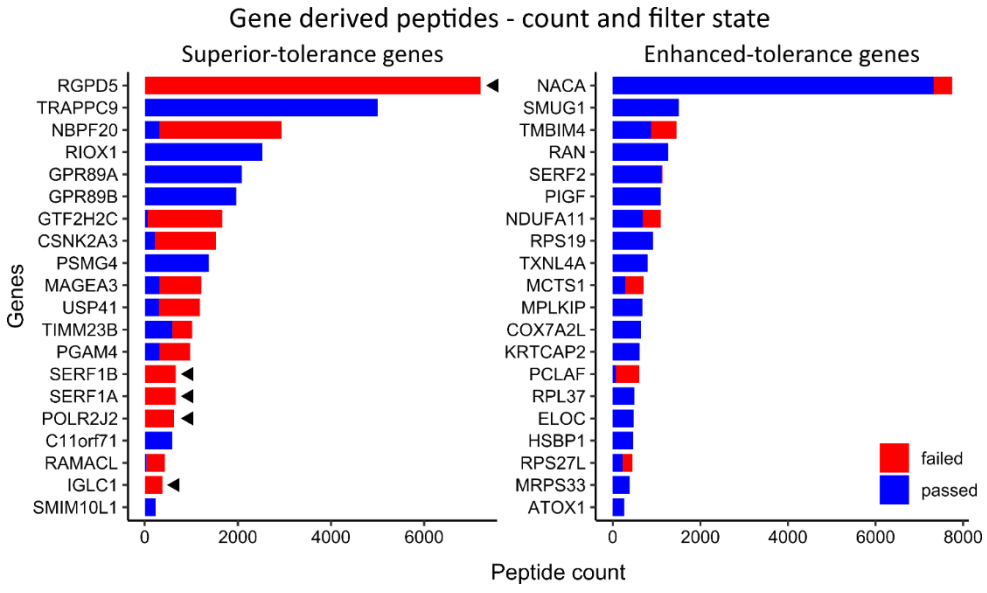


Figure 11: ***Candidate genes are shown with the amount of derived 9- to 12-mers****. Colors indicate the filtering result of peptides regarding the literal comparison to the complementary proteome. Genes that failed this filter because of all their peptides appearing in other known proteins are indicated with a black arrow. Five genes in the superior-tolerance set and no genes in the enhanced-tolerance set failed this procedure.*

4.1.3 Characterization of predicted efficacious peptides

Since the previous processing steps focused on filters regarding the peripheral presence of peptides, meaning anywhere outside the tumor, through either expression or homology, it was necessary to establish their validity as MHC-I-restricted antigenic targets. Hence, we established a ranking system for the short-listed peptides which remained. A total of 34,277 peptides were processed in this step. Using netMHCpan 4.0, we predicted allele-specific binding affinities for 36 HLA alleles as described in section 3.4.3.1, yielding 1,232,172 epitopes. To reduce the candidate list further and bring it closer to the realm of applicability, we used an affinity-based filter, removing all epitopes if their predicted binding affinity between the peptide of interest and the HLA allele was higher than 500 nmol/L, a commonly used cut-off to determine presumed binders (IC50<500 nmol/L) from non-binders (IC50>500 nmol/L)(Zhao and Sher, 2018). Most epitopes were not predicted to be high-affinity binders, and the list was reduced to 10,597 epitopes with 6,397 unique peptides since a peptide may be predicted to bin several alleles. Generally, the distributions of binding affinities over all alleles were unimodal, with the peak in low-affinity ranges, demonstrating that the principal expectation should be that most peptides are predicted to be weak binders (**Figure 12**).

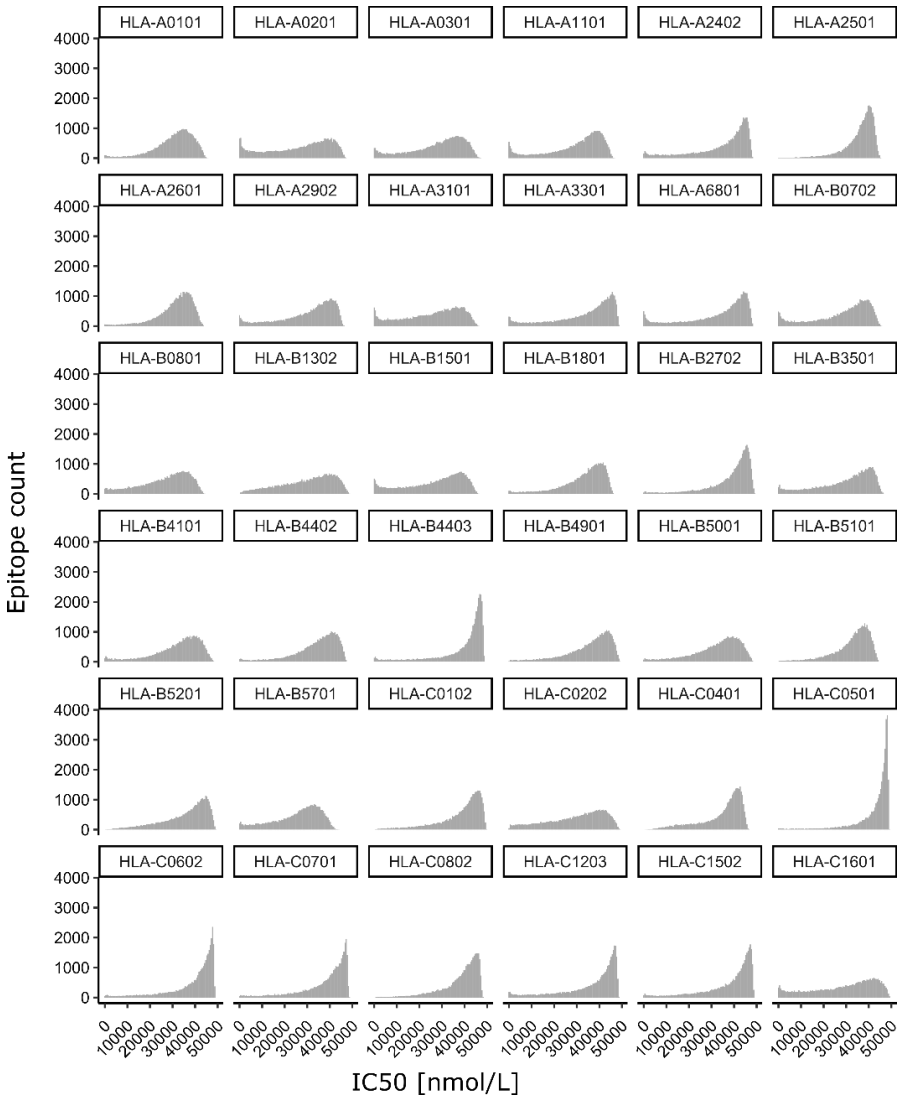


Figure 12: ***Binding affinity distributions per HLA allele for all 34,277 peptides not discarded in the sequence identity filter.*** *All distributions are relatively unimodal and left-tailed, with high counts of epitopes in the low-affinity regions. Some alleles, like HLA-A*02:01 or HLA-A*11:01, showed a secondary peak in the high-affinity regions. Since we try to cover a broad population with our predictions, it is necessary to ensure that we can achieve good coverage of binders and find potential alleles for which we have gaps in our candidates.*

The range of the predicted affinity values was 2.46 to 49,739.3 nmol/L. Since HLA alleles are patient-specific, it is relevant to observe if there are gaps in the predictive modeling of our pipeline where some alleles would not produce valid candidates for possible clinical use. Indeed, while the overall distributions presented as homogenous, we found that the alleles HLA-C*04:01 and HLA-B*52:01 did not have any predicted epitopes with an affinity lower than 500 nmol/L. Generally, Superior-tolerance and Enhanced-tolerance sets produced comparable binders (IC50<500nmol/L). The behavior was likewise observable for the non-binders for each allele (IC50>500nmol/L) (**Figure 13**).

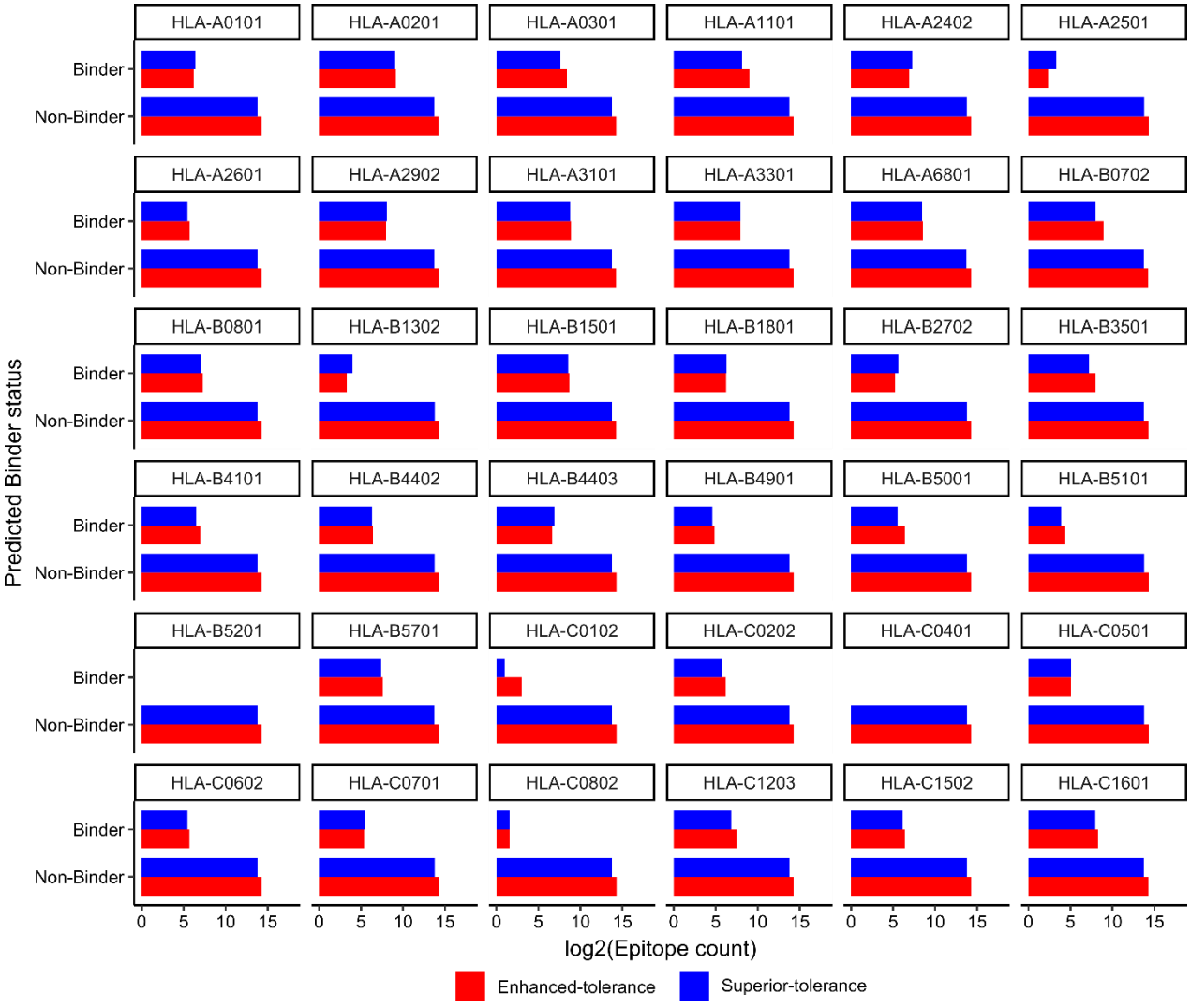


Figure 13: ***Amount of binding (high-affinity) or not-binding (low-affinity) peptides produced per tolerance set and per allele on a log2 scale for ease of comparison.*** *Shown are the predicted binder status for each allele and its associated epitopes. Since the relation of non-binders to binders is heavily skewed towards non-binders, the x-axis of the counts is log2 transformed. Generally, no significant imbalance in the sets was observed. Enhanced- or superior-tolerance sets produced comparable amounts of binders and non-binders. Two alleles, HLA-B*52:01 and HLA-C*04:01, did not give rise to any binders, making them gaps in our predictive pipeline.*

We further investigated if there are peptides that are promiscuous in their affinity characteristics since, for an off-the-shelf therapy option, a broad high-affinity profile would be advantageous from both therapy and manufacturing perspectives. We found that one peptide would at most bind 11 different alleles while most of the peptides were only highly affine to one allele (**Figure 14A**). Additionally, we did not observe any significant clustering of peptides according to alleles, with all high-affinity peptides distributed broadly over the investigated alleles (**Figure 14B**).

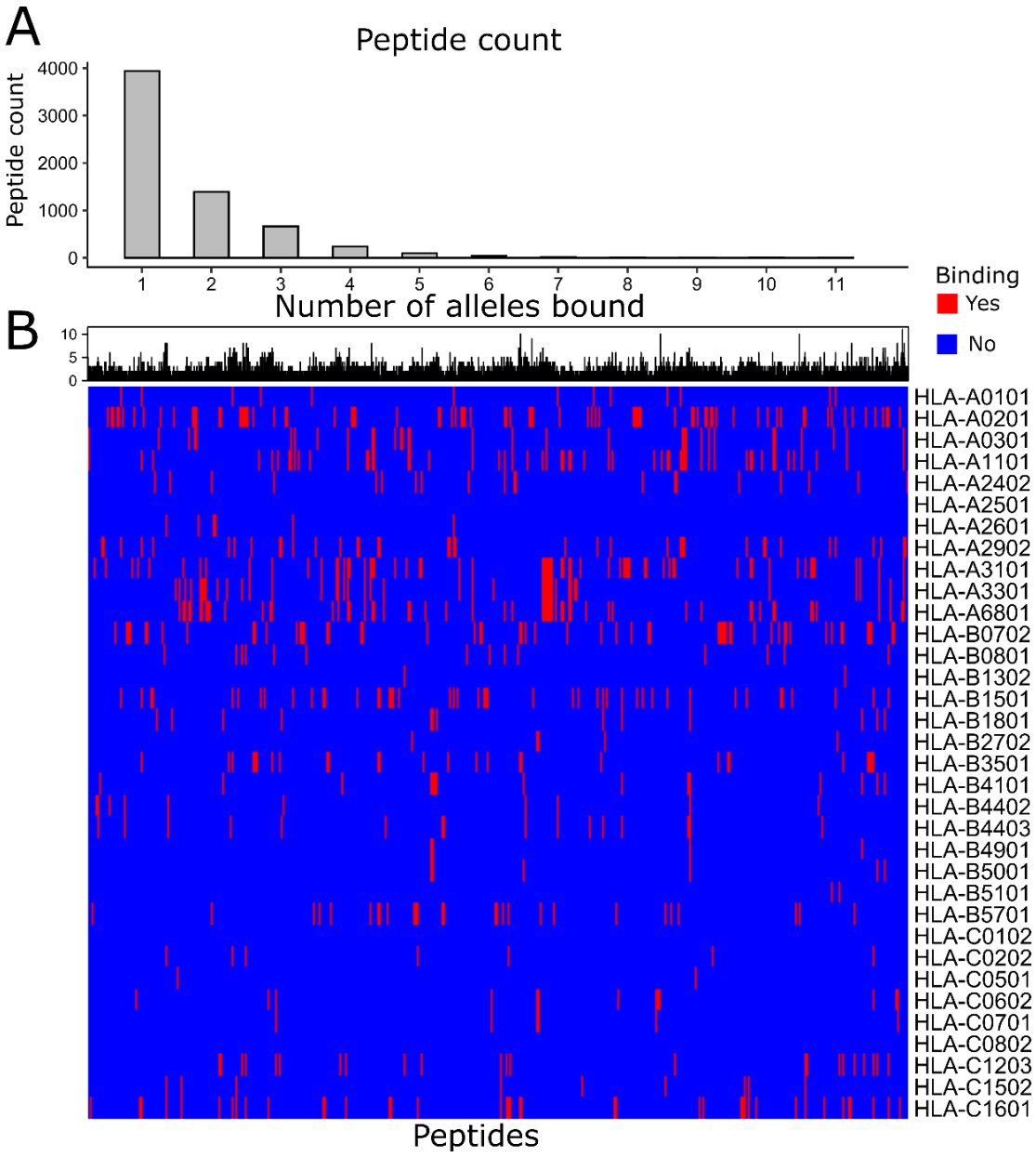


Figure 14: ***(A) The distribution of high-affinity alleles by peptide count.*** *3940 of the 6,397 peptides are only highly affine to one allele. Taking together peptides binding one, two, or three alleles covers 93% of all peptides. The X-axis of Figure A Is also the title of the histogram of Figure B. Both show the number of alleles bound.* ***(B) Binary heat map for the peptides and their binding profiles over the HLA alleles.*** *Peptides are presented on the columns with red indicating an IC50 smaller than 500 nm and blue conversely indicating an affinity greater than 500 nm. Few good general binders exist, with the maximum being 11 bound alleles for the peptide* `YTVENSRVY`*, while there are three peptides that bind ten alleles and one that binds nine. However, we did not find general binders that may be used supertype-wide, for example, in all HLA-A alleles.*

Since no single peptide would be predicted to bind a majority of alleles and with a large set of peptides available to select from, we deemed it necessary to implement an easy-to-understand ranking system to facilitate selection decisions. Hence, we designed a ranking scheme based on a multi-criteria score, as discussed in section 3.4.3. Briefly, the score evaluates several epitope features, which are commonly used in the community, like immunogenicity predictions performed through IEDB (Calis *et al.*, 2013), affinity predictions (IC50) performed through netMHCpan4.0 (Jurtz *et al.*, 2017) and expression characteristics in TPM, and combines them into a single value that is easily interpretable. We named this value the gPIE score and calculated it according to section 3.4.3. The score represents a multiplicative metric that combines all our considered parameters. All predicted efficacious epitopes from our two candidate gene sets considered for the database included the transcript-specific expression for the peptide's source gene into the score. This step from gene-level to transcript-level expression increases the total database size since the same peptide can potentially be produced from several transcripts due to redundant exon usage. This, however, allows a more granular ranking of peptides since the total gene expression is resolved into individual-transcript expression. When computing the gPIE score, it became apparent that while its hypothetical range is 0 to 100, its limits in our data were ranging from 0 to 52.36 The highest values were calculated in the Enhanced-tolerance set while the Superior-tolerance set reached a maximum of 17.3 for its highest-ranking epitope derived from the known melanoma antigen *MAGEA3*. We also included other previously described melanoma antigens in our database **(Table 3)**. We found, however, that most of them did not rank particularly highly with the maximum score being 5.49 **(Figure 15**).

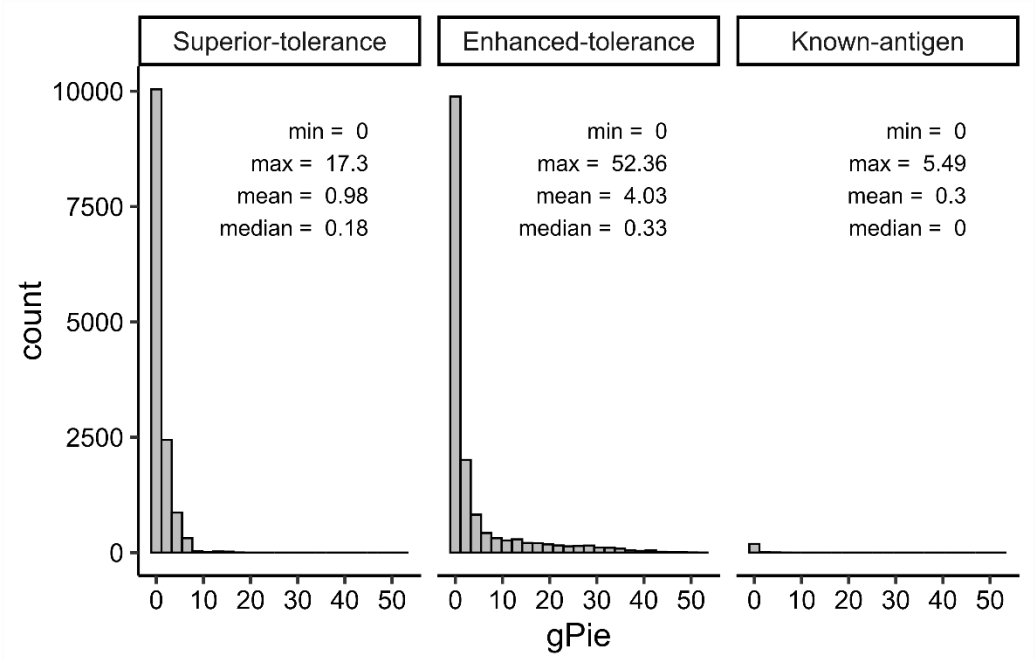


Figure 15: ***Distribution of the gPIE score for all sets in the database.*** *A high quantity of epitopes were assigned a score of zero and are thus not considered particularly efficacious for application in therapy. The highest score was found in the Enhanced-tolerance set, perhaps reinforcing the idea that a balance between autoimmunity risk and anti-tumor immunogenicity must be struck.*

The observation that many epitopes were assigned a score of zero begs the question as to which factor is causative. The nature of a multiplicative score necessitates that the entire score will be zero if one element is zero. An investigation into the dominating cause for zero-scored epitopes showed that, in most cases, normalized transcript expression (**Figure 16, F3**) was the culprit in both the Superior-tolerance and the Enhanced-tolerance set. Surprisingly, the binding affinity for the known antigens was the leading cause for zero-scored epitopes, together with the total contribution of the source gene's expression to the total gene expression in our analyzed MCM cohort (**Figure 16, F1 and F4**). Otherwise, the elements of the gPIE score were distributed evenly and covered almost the entire range, with F1 (normalized IC50) having high-affinity binders with a score of 0.99 and a mean of 0.50.

Similarly, F2 (normalized predicted immunogenicity) showed a broad range with a minimum of 0, a maximum of 0.84, and a mean of 0.62. As mentioned, F3 (normalized transcript expression as observed in our cohort) had the most drastic effect on the zero-scored candidates while still having a high overall range with a very skewed distribution with a maximum of 1 and a mean of 0.03, implying that highly expressed tumor transcripts were still filtered out through other variables. Finally, F4 (the gene expression index) demonstrated a similarly broad range with 0.82 as a maximum and 0.54 as a mean. This shows that highly tumor-restricted transcripts, with an 0.82 expression index (most expression originates from the tumor), also may be filtered out (**Figure 16, F4**).

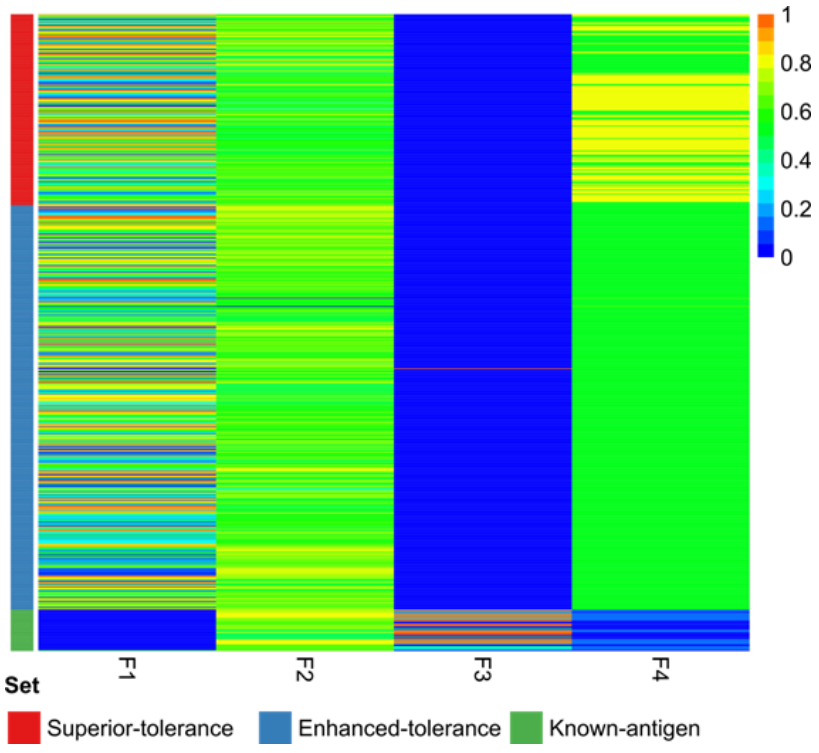


Figure 16: ***Epitopes that were assigned a gPIE score of zero.*** *Heatmap shows the elements of the gPIE score normalized IC50 (F1), normalized predicted immunogenicity (F2), normalized transcript expression (F3), and expression index (F4) for the epitopes that were scored zero in the gPIE annotated as the set from which they originate. Generally, transcript expression was shown to be the most common cause for the Superior- and Enhanced-tolerance set, while for the known antigens, both the expression index and the binding affinity were the cause.*

Accordingly, we investigated the individual elements' contribution to the epitopes' overall score to see if one factor would dominate positively or negatively. To this end, we filtered out all zero-scored elements to see the effects of the other variables on the positively scored epitopes. We did not observe similar trends in the positive score epitopes compared to the removed ones. Transcript expression was still a dominant factor in decreasing the overall score of an epitope (**Figure 17, F3**). While expression was again an important factor in scoring, we made efforts to combine all usually applied metrics to derive a comprehensive score that helps users make quick and parameter-based decisions.

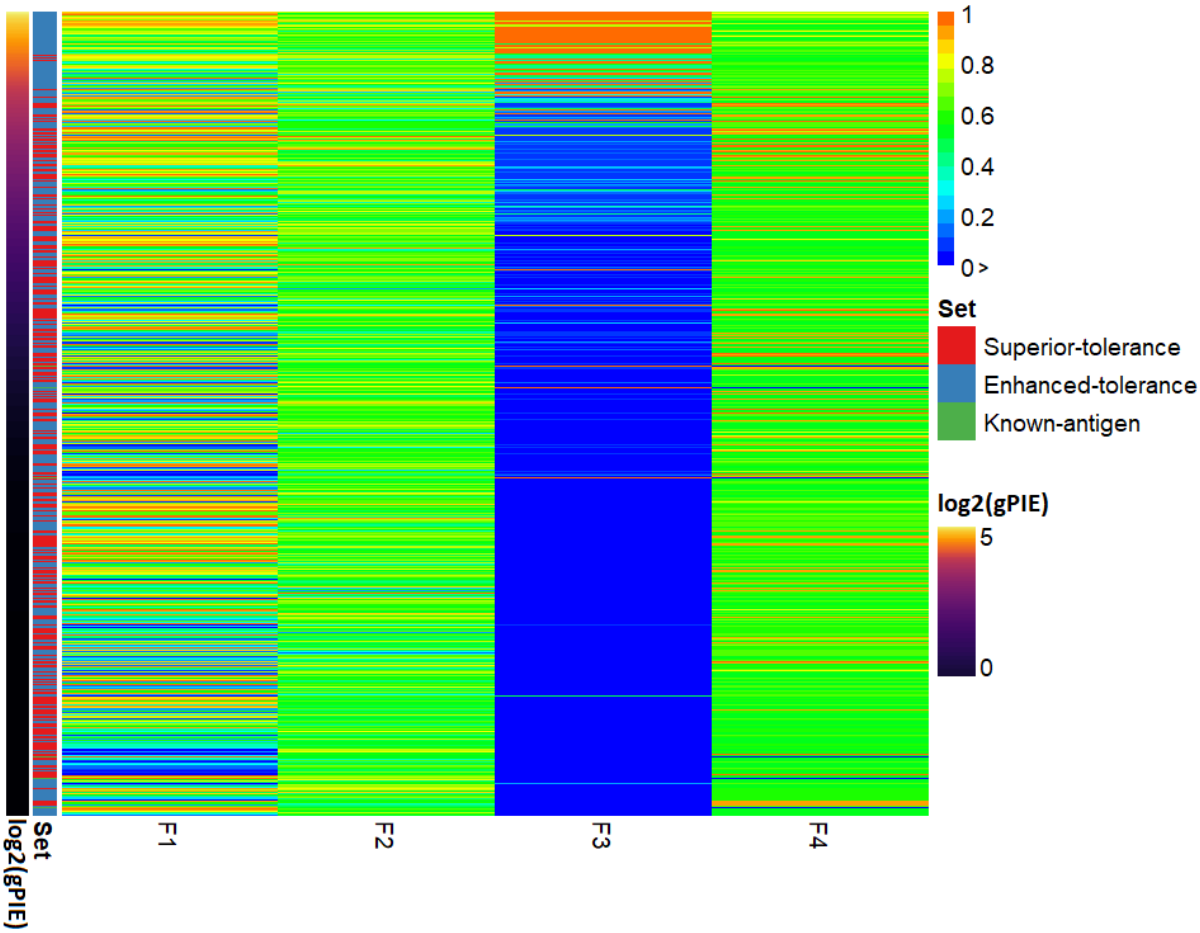


Figure 17: ***Contribution of the individual elements of the gPIE for each epitope scored above 0.*** *The gPIE scale visualization is presented in log2 scale after addition of 1 to each value for ease of interpretation. It is apparent that transcript expression (F3) in the tumor is still a major contributor to the score, while all other factors show a homogenous distribution with no dominating element between them. It is of note that the known antigens are widely distributed between the sets and are hard to make out in the overall map due to the high-class imbalance.*

4.1.4 Curatopes 1.0 – Database design and functionality

We created a database of our scored epitopes available to the public. The landing page of the database, which is available under www.curatopes.com/melanoma, features an overview of the superior-tolerance set and serves as a first introductory portal for further use (**Figure 18**). The overall design is separated into a tutorial document, available to download, a detailed documentation that explains the database's elements, a legal disclaimer in case the database is used in clinical settings, and a link to the peer-reviewed scientific article introducing Curatopes.

One can also download the whole database or select entries using the filtering fields below the table headers. Immediately accessible data columns are a gene of origin with a link to the corresponding Genecards (www.genecards.org) entry, the peptide's AA sequence, its three best-scored HLA alleles, the epitope's gPIE score, and finally, how many collateral tissues we detected expression of the gene of origin.

If needed, additional columns for binding affinity, immunogenicity, transcript expression, tumor 10th percentile expression, maximum tissue 90th percentile expression, peptide start position in the native AA chain, and the peptide length can be made visible by the user but are initially hidden as to not overload the page with information.

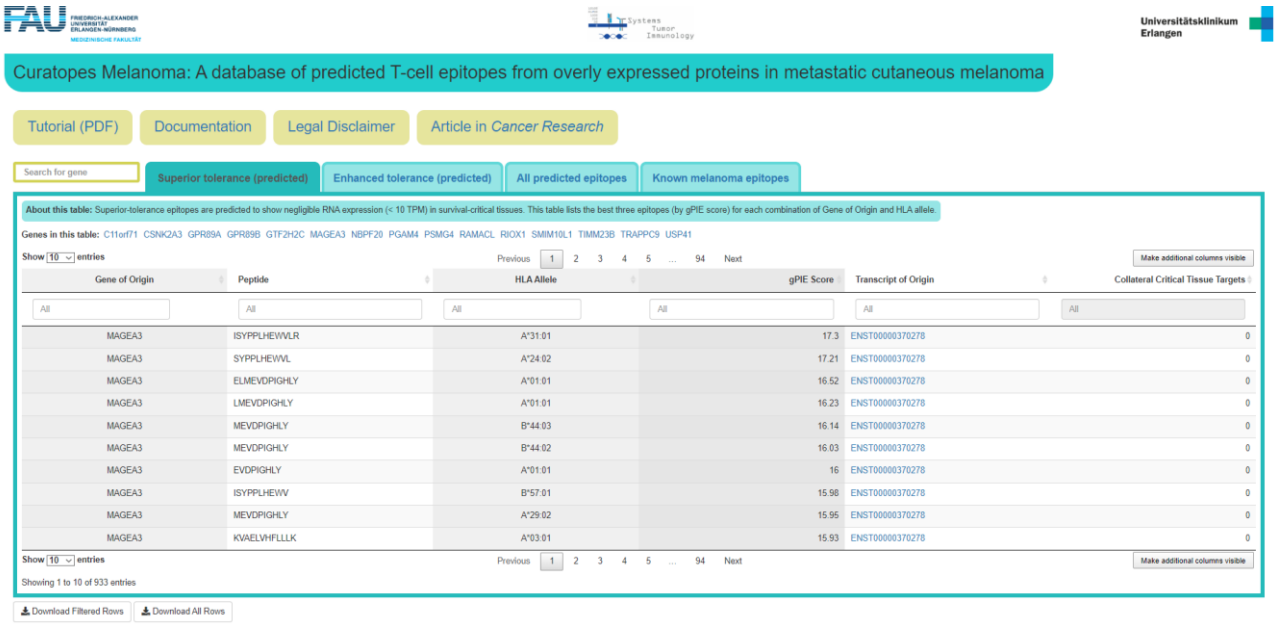


Figure 18: ***Landing page of the Curatopes Melanoma database available to the public.*** *The highest-scoring epitopes in the superior-tolerance set are shown. Additionally, all the additional functionalities can be accessed from here. First, the page offers a tutorial explaining how to query the database and download tables for further use. Detailed documentation on each parameter shown in the table is linked at the top. Since this is predicted data, there is a legal disclaimer in case somebody wants to use peptides or epitopes in clinical settings. Finally, there is a link to the published article covering the database. The fundamental functions to operate on the data are exploring the gene sets, sorting, subset, filtering them as needed, and downloading selected subsets.*

One can access the tutorial document by clicking on the Tutorial button if interested in more detailed information on how to use the database. This will download a short document addressing questions like the analysis's fundamental idea and resulting database. If a more in-depth understanding is needed, users can access complete documentation via the “Documentation” link, which elaborates in greater detail on how the database was curated methodologically and how elements of the gPIE score were calculated (**Figure 19**).

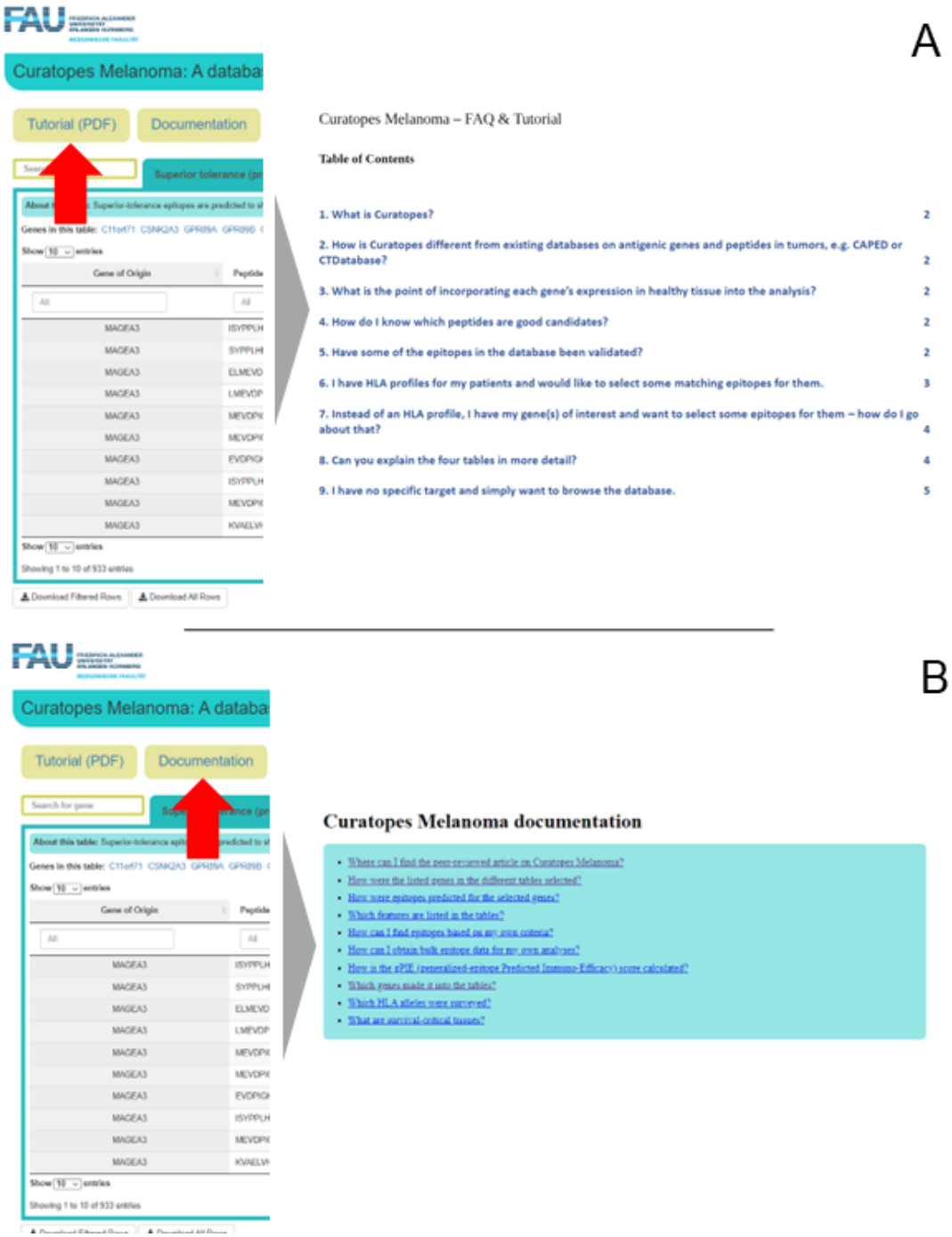


Figure 19: ***Quick access buttons for the tutorial (A) and the documentation (B) documents on the web platform.***

4.2 Tumor-associated antigens in primary uveal melanoma

In the second phase of this work, we implemented an extension of our methodology for the search for TAAs in the ocular cancer uveal melanoma (UM). This phase focused on extending the score by including network modeling and supplementing existing elements with novel prediction systems. Additionally, while including the peripheral tissue filter developed during the first part of this project, we now set out to check the plausibility of our predictions with *in vitro* experiments. The pipeline is built to create a rational and deterministic prediction-to-validation workflow that can be repeated for different tumor entities as necessary (**Figure 20**).

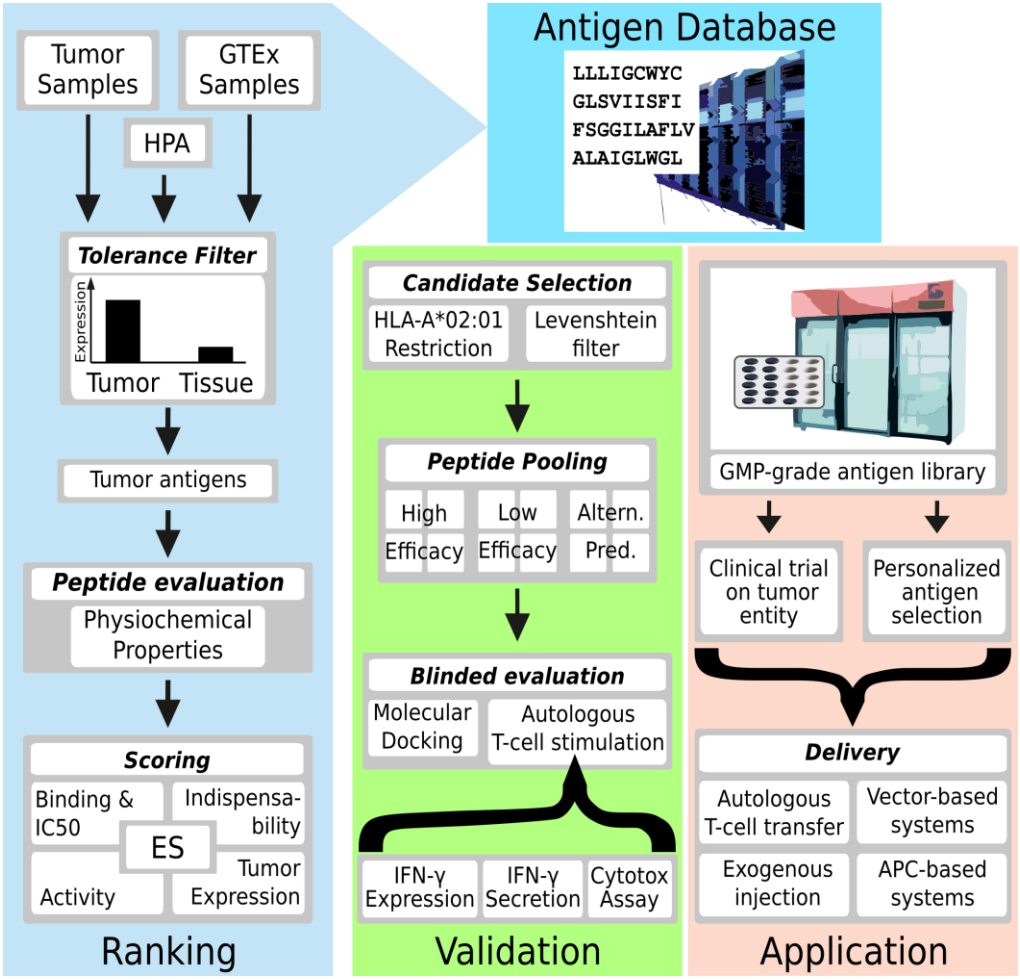


Figure 20: ***Overview of study scope.*** *Our approach can be conceptualized as four interleaved workflows: an in-silico ranking pipeline (blue), a permanent database of ranked candidates available to clinicians (cyan), a validation protocol for proof-of-principle tests (green), and paths to application in the clinics (salmon). During ranking, we evaluated and filtered genes based on their expression profiles to create a database of tumor antigens that we propose as optimized candidates for targeted anti-cancer therapy. To check whether high-ranked peptides can elicit an immune response, we performed blinded in-vitro tests with PBMCs from healthy donors. The immunogenic candidates can then be tested in clinical trials or applied in a personalized setting by way of different delivery systems. GTEx, Genotype-Tissue Expression. HPA, Human Protein Atlas. IC50, binding affinity. Altern. Pred., alternative predictor. GMP, good manufacturing practice. APC, antigen-presenting cell. ES, efficacy score.*

4.2.1 In-house and model cohort transcriptomics

First, we curated a study cohort for uveal melanoma. Through a clinical trial headed by the group of Experimental Immunotherapy at the Universitätsklinikum Erlangen, we had access to 14 primary UM samples sequenced using NGS. In parallel, we downloaded 80 samples of primary UM from a published study (Robertson *et al.*, 2017). In a first exploratory analysis, we investigated how heterogeneous the samples were across cohorts. To this end, we performed principal component analysis on the combined TPMs of our in-house-produced data and the external cohort. It became immediately apparent that the two groups were separated distinctly by their respective processing locations, thus showing a strong batch effect commonly seen for data generated at different institutions (Leek *et al.*, 2010). We decided to leave our in-house generated samples out of the primary workflow and use them later for validation analysis since 80 samples would provide a broader sampling basis from the UM expression landscape without the need to correct batch effects (**Figure 21**).

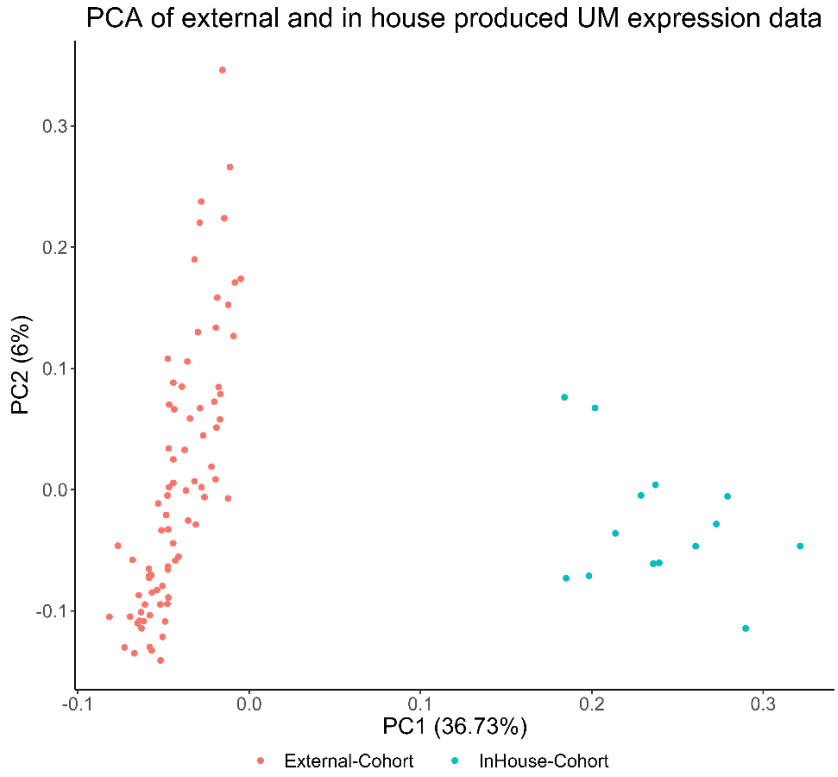


Figure 21: ***Principal component analysis of the in-house-generated expression data and the publicly available dataset.*** *A high degree of variation on the first PC separates the two groups. This may indicate strong technical or processing differences between the samples, which we cannot separate from biological differences.*

4.2.2 Transcriptomics-based filtering procedures

As described in section 3.3, we applied our transcriptomics filter for candidate gene selection to the 80-sample strong UM cohort to identify protein-coding genes with a high-in-tumor, low-in-tissue expression profile from the empirically measured mRNA abundances. We observed again that the filtering procedures developed in the first phase and applied now to UM were quite strict and yielded 22 candidates out of 10,514 genes with a 10^{th}-quantile expression above 1 (**Figure 22 A**). In our in-house dataset, we checked whether the 22 genes were detectable and expressed to see if our candidates generalize to the larger UM patient population. We found that all 22 genes were stably expressed in the independent in-house cohort, albeit with a wide range of expression intensities (**Figure 22 B**), increasing our confidence that these genes are targetable across UM patients in a general manner. This is an essential feature for TAAs since we must avoid overfitting our set of selected genes on a small subset of patients, which is especially difficult for a rare tumor-like UM with limited sampling opportunities. It is of note that known melanocyte-derived antigens like *MLANA*, *TYR*, and *PMEL* could be shown to be stably expressed across our selected UM samples (Rähni *et al.*, 2022). Even though the melanocytes forming these tumors are located in different tissues, these antigen's expression profiles seem relatively stable, at least in a malignant state.

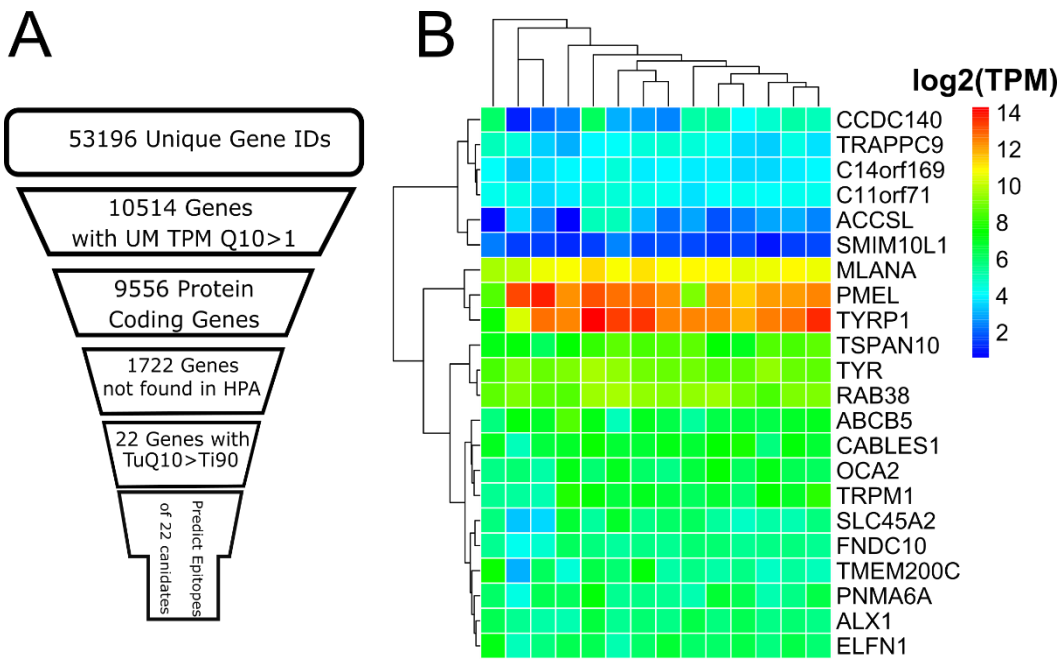


Figure 22: ***Selection and cross-comparison of candidate genes.*** *Selection funnel representing a cascade of* in-silico *filters for genes. Each slice of the funnel lists the feature criterion and the number of genes meeting it. Tumor expression statistics were calculated using a published set of 80 primary UM samples. Ultimately, 22 candidate genes passed all filters.* ***(B)*** *Heat map of gene expression of the 22 candidate genes in an independent set of 14 primary UM biopsies produced in-house. Log2-transformed transcripts per million (TPM) estimates are shown. Stable expression levels of the 22 candidate genes were observed across individuals.*

4.2.3 Generation of an indispensability network index

One element by which we extended the functionality of our pipeline was the evaluation of the biological significance of a gene. To this end, we generated what we deemed the indispensability index (Idspx), a score that should estimate the costliness of a tumor to suppress a particular gene. In the context of targeted immunotherapy, the Idspx measures how difficult it is for cancer to evade the therapy by not expressing the antigen anymore, a phenomenon known as antigen loss. If we take the network perspective on this question, how central is a gene in the context of a broader functional biological network that governs a tumor's biological viability, stability, and transformation? We first used our 22 genes as a seed to generate a candidate network. We queried their immediate connections to other genes according to different interaction databases and thus generated a network consisting of 167 nodes and 349 edges (**Figure 23 A**). In a second step, we grouped our 22 candidates with known oncogenes derived from published resources to generate a background network that calculated how well-connected these oncogenes are (**Figure 23 B**). The background network comprised 20,057 nodes and 290,657 edges (section 3.4.4.3).

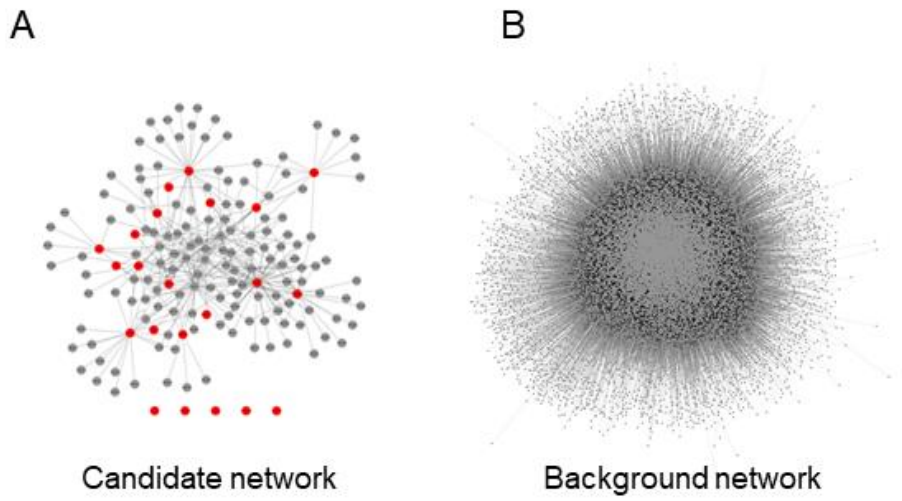


Figure 23: ***Networks were created for the generation of a biological gene importance index.*** *(A) Candidate network estimating the connectivity and the functional distance between candidate genes. (B) Candidate genes are embedded into an extensive background network that contextualizes them to other oncogenes. From this network, we derived metrics like node degree, which were used in the indispensability index.*

From the networks, we generated a node degree (ND) table which listed to how many other genes a single gene was immediately connected to. From cancer-related databases and a curated list of cancer-relevant GO terms, we derived a gene importance (GI) value by counting how often any gene in the network was associated with cancer (**Table 6, section 3.4.4.3)**. We found that the GI values had a wide range, with some outlier genes occurring very often in this scoring setting. For example, *TP53* had the highest GI value of 235, with *TNF* in the second rank with 122. In general, 95% of the genes showed a GI below 15. Our candidate genes had a maximum GI value of 14 for the *TEMEM200C*, while two genes, *C14orf169* and *PNMA6A,* had a GI value of zero, pointing to no known association with cancer. The melanocyte antigen *MLANA* had a GI value of two, the same as the melanoma-associated gene *TYRP1* (**Figure 24 A, Panel 1**).

Having established the GI scores of our genes, we proceeded to investigate the previously calculated node degrees. We interpreted the node degrees as a measure of the biological embeddedness of a particular gene into larger biological pathways and, thus, necessary for tumor functionality.

Overall, the node degree had a wide range of values, with genes like *YBX1* having extremely high node degrees (5991) in the background network. Our candidate genes were generally located in the lower regions of the node degree distribution, considering that the maximum was 43 for *CABLES1*, followed by TYRP1 with 28. Three candidate genes did not have any connection to other nodes: *SCL45A2*, *ACCSL,* and *ALX1* (**Figure 24 A, Panel 2**). By multiplying the adjacency matrix, a binary matrix indicating with binary (0,1) values if a connection (edge) exists between two genes (nodes), with the occurrence vector (GI values), we derived the neighborhood importance (NI). The NI value estimated the importance of a gene within its local community and, by extension, how sensitive its targeting would be in the global tumor network. Again, we found the gene *CABLES1* to have a high score here, with a NI of 76, followed by *TRPM1* with 47 and *TYRP1* with 46.

Interestingly, the candidate's distribution over the three parameters looked somewhat similar (**Figure 24 A**). Using all these metrics, we plugged the values into the formulae described in section 3.4.4.3, **Equations XI** to **XII**. The computed parameter for a value of five in the empirical distribution function of the node degree was 0.675 or the 67.5^{th} percentile. Calculating the corresponding *p*-quantile from the ND distribution, we derived a value of 18, which we deemed "sufficiently connected" by analogy. Plugging this into our equations, we arrive at a K_M parameter of 10 and finally calculate the normalized NI values, now termed the Idspx. The Idspx was constrained to unit distance using a Michaelis-Menten-like function with saturation behavior to make it easier to interpret and compare to our other subfunctions. We calculated this value over all background genes to estimate how highly relevant cancer genes would be evaluated. We found that *YBX1*, *FOXP3*, *TP53,* and *MYC* were all in the top-ranked genes in our Idspx table, with *YBX1* holding the top rank, *FOXP3* ranked second, *MYC* ranking nine, and *TP53* ranking 13. The values for these highly ranked genes were all very near one, with 0.99959, 0.99953, 0.99728, and 0.99939, respectively, for the mentioned genes.

Our candidate genes were widely distributed. *CABLES1,* at rank 679, was the highest evaluated gene with 0.8837, while *C11orf71* ranked lowest with 0.375 at rank 12,928 (**Figure 24B**). We hypothesize that this novel method estimates the biological importance of a particular gene to the tumor entity, in this case, UM, which may be generalizable to other cancers.

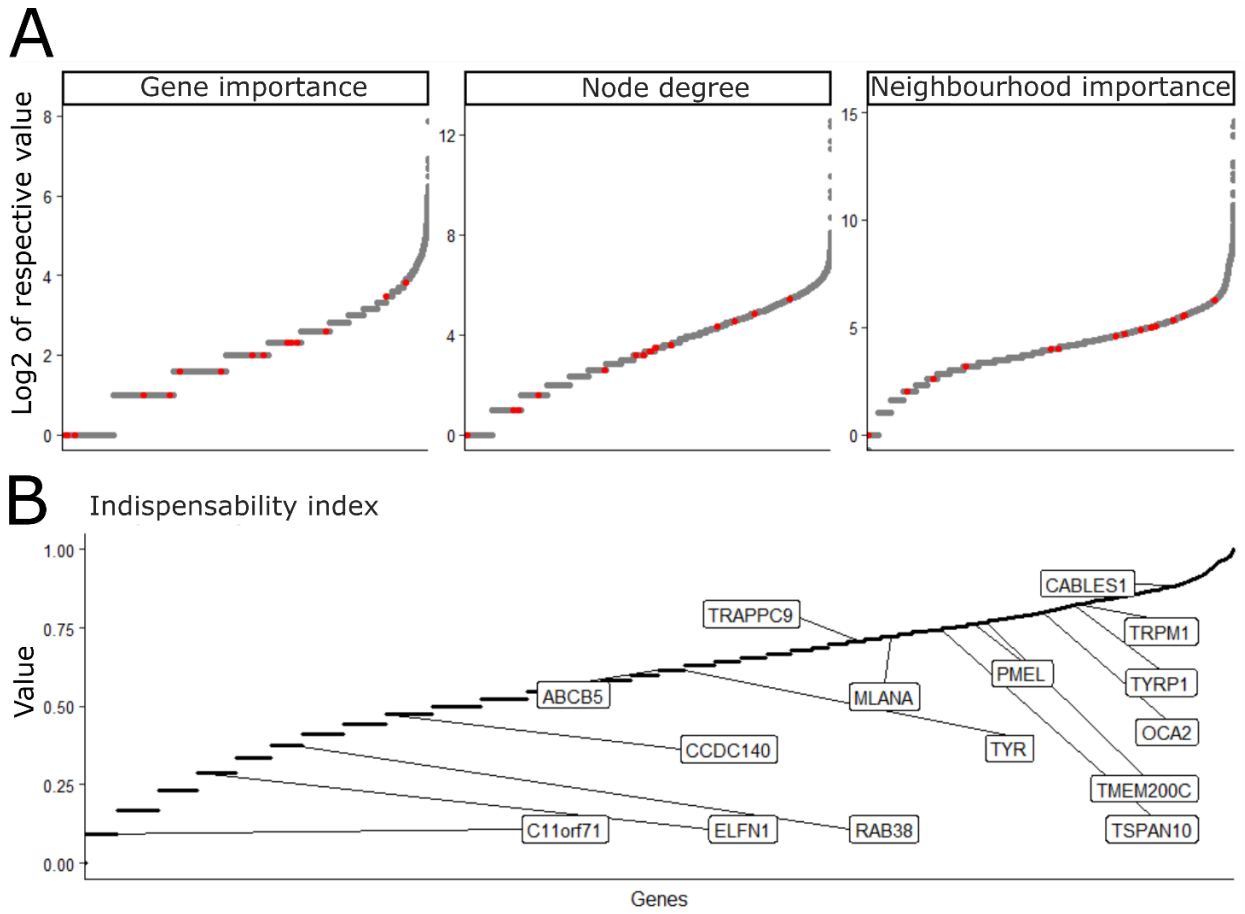


Figure 24: ***Elements of the network score and normalized indispensability index for the selected candidate genes that had non-zero values in either gene importance or node degree.*** *(A) The three panels show the individual ordered elements from which we derived our indispensability index. Red shows our candidate genes, while grey indicates an oncogene from the curated database. The panels primarily show that our genes are rarely located in the extreme value ranges but rather in the lower to mid ranges, giving robust estimates of their biological relevance for UM. (B) Distribution of the indispensability index, which was calculated from the elements shown above.*

4.2.1 Candidate gene-derived peptide sequence-based post-hoc screening

To avoid possible cross-reactivities based on sequence identity, we applied post-hoc peptide level screening to all peptide k-mers derived from the 22 candidate genes **(section 3.4.1)**. From 71 protein-coding transcripts, we extracted 51,374 unique peptides on which post-hoc k-mer-based peptide filtering was performed. During this filtering, 11,343 peptides were removed upon finding cross-matches, while 40,031 were kept. In contrast to our model for MCM, no genes were lost in this filtering step since no candidate showed such high AA sequence homology with another annotated gene that all of its derived 9- to 12-mers were discarded (**Figure 25**).

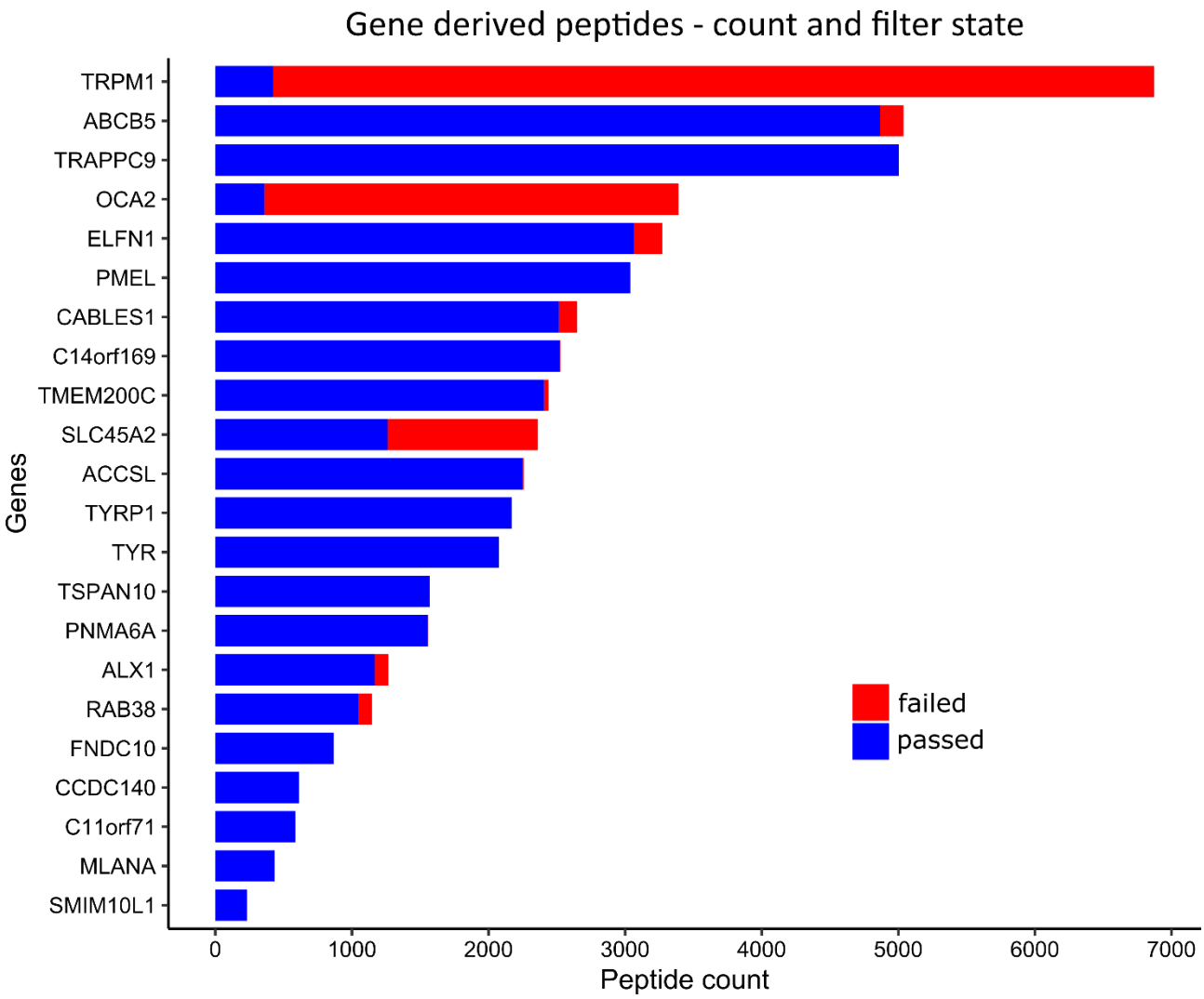


Figure 25: ***Candidate TAA for UM shown with the amount of derived 9 to 12 mers****. Colors indicate the filtering state of peptides regarding the literal comparison to the proteome. No genes failed the sequence-based filter.*

4.2.2 Ensemble model prediction for binding and activity

Having a set of 40,031 peptides with MHC-I-preferred lengths and screened against the proteome, we now implemented a novel generalized prediction model to gauge a peptide's potential efficacy in binding MHC-I and inducing an immunological response. We used the MHCBN database, which includes binding and T-cell reactivity data. We constructed two random forest predictors derived from the peptides' physiochemical properties as features (hydrophobicity, polarity, stability, isoelectric point, and molecular weight)(section 3.4.4.1) (Lata, Bhasin and Raghava, 2009). Since generating all possible peptides from the primary protein sequence yields a large dataset, including peptides that would, under normal physiological circumstances, get discarded during the processing leading up to MHC-I loading, we designed our predictor to handle this highly imbalanced data. We plotted the receiver operation characteristics (ROC) curves to test the overall performance of our two models. A random process, meaning when the true positive rate is equal to the false positive rate at every threshold, lies on the diagonal. Our binding model performed better, having an area under the curve (AUC) of 0.853 compared to the activity model with an AUC of 0.636. However, comparable, already published alternatives for binding prediction and immunogenic activity performed overall similarly but slightly worse than our models. The binding predictor NetMHCpan4.0 showed an AUC of 0.815 for binding prediction, while IEDB's `predict_Immunogenicty` produced an AUC of 0.559 (**Figure 26**) (Calis *et al.*, 2013; Jurtz *et al.*, 2017).

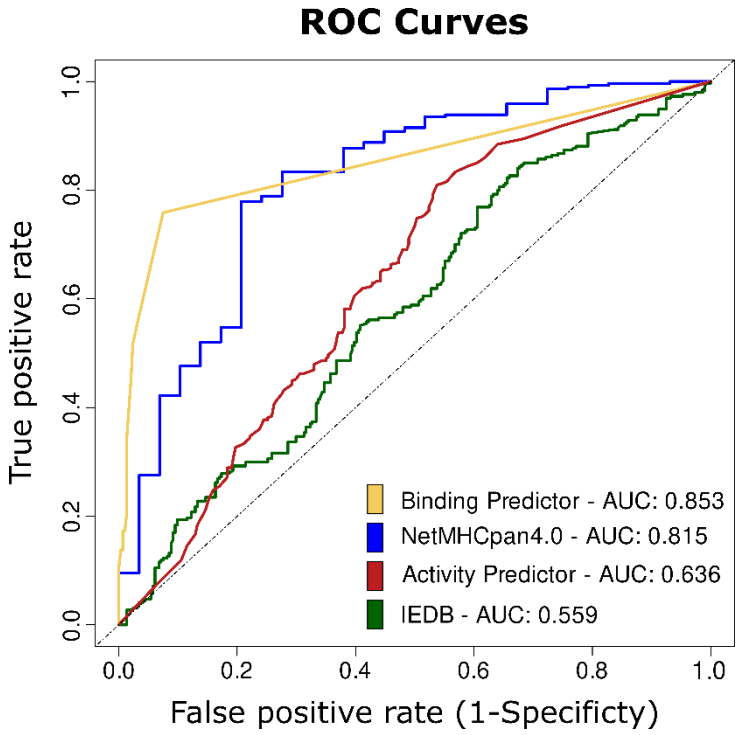


Figure 26: ***Random forest based binding and activity predictor performance benchmarks validated against sampling from the input training compared to the predictive performance of published tools.*** *We tested each model against sampling from the entire training set and found that our models performed slightly better in terms of area under the curve (AUC) compared to two standard tools for immunogenicity prediction (IEDB) or binding prediction (netMHCpan4.0).*

After benchmarking and demonstrating our models' performance, we predicted the generalized binding probability (gBP) and the generalized activity probability (gAP) for all unique peptides. We predicted gBP and gAP for 40,031 unique peptides derived from our 22 candidate genes. After predictions were performed, we first investigated if there was a dependence between the gAP and gBP. We did not find a direct relationship between these two components, with the correlation being 0.07. This suggested that a high binding probability does not translate into a high probability of immunogenic activity and *vice versa* (**Figure 27**).

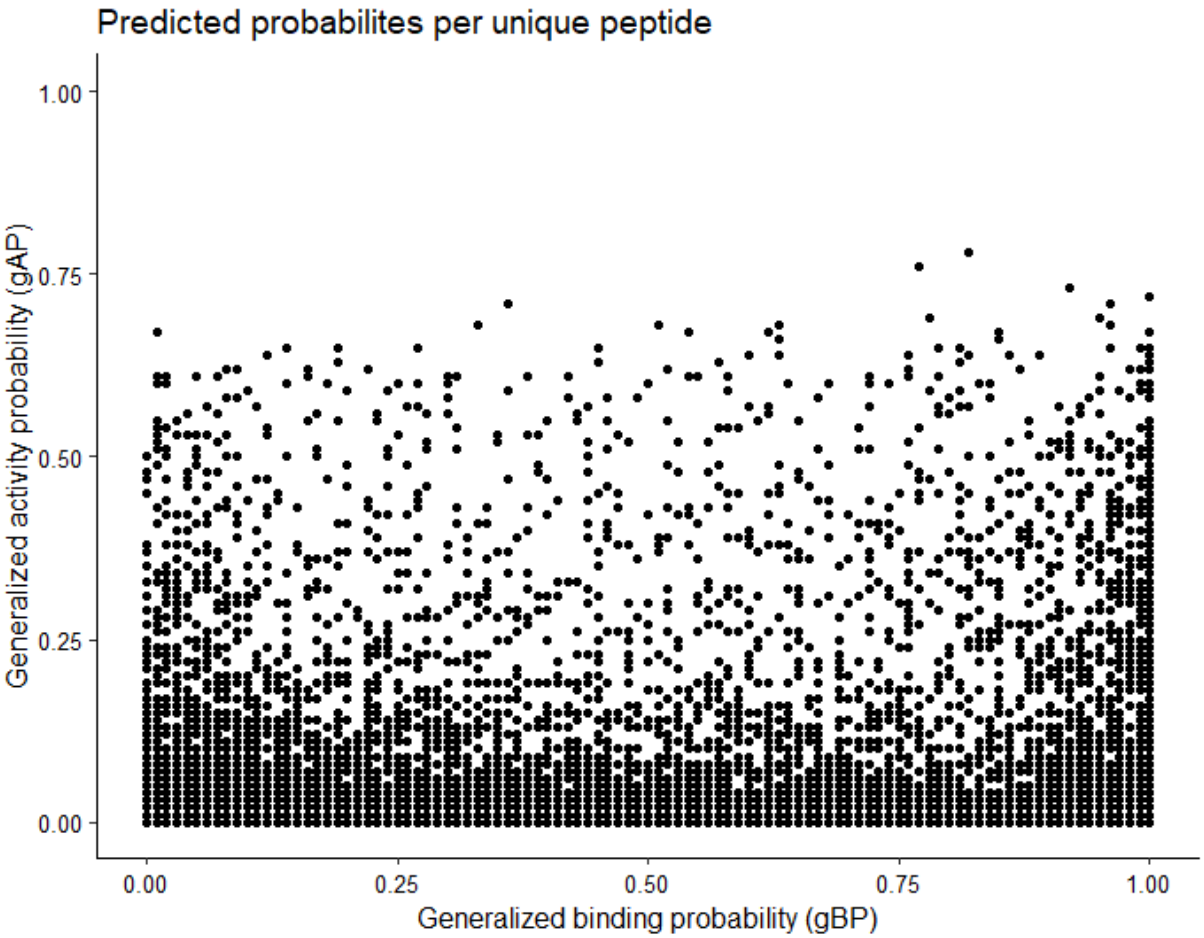


Figure 27: ***Scatter plot of our set's predicted generalized binding and activity probability for each unique peptide.*** *Since predictors were trained without regard to HLA alleles' binding preferences, just sequence derivable features, each unique peptide receives a generalized (cross-allele) probability value for activity (gAP) and binding (gBP). Our predictors did not show an appreciable correlation against each other. High-probability binders could have low activity probability and vice versa.*

4.2.3 Characterization of predicted efficacious peptides

To aggregate our parameters into the efficacy score (ES), we integrated an expression estimate, an affinity estimate, the network model-derived indispensability index, and two probabilities generated by the ML model. Using these parameters, we calculated the ES according to section 3.4.4, **Equation VII**, for all combinations of the 36 alleles and 40031 unique peptides which passed filtering. Hence, we scored approximately 1.4 million epitopes and assessed how the ES distributed over the epitopes. The range of the ES was designed to be between 0, the worst score, and 100, the best score. Most epitopes received a score of 0, with only 47408 epitopes receiving a non-zero score. This translates roughly to a ratio of 1 in 30 or 33 in 1000 epitopes having a positive, non-zero score. While the ES is unitless and dimensionless, it can be interpreted as a probability for the epitope to be efficacious. Investigating the zero-score epitopes further, we found that two genes, *FNDC10* and *SMIM10L1*, were assigned scores of 0 for all their 31176 and 8352 epitopes, respectively. In both cases, the gene's Idspx was 0, causing the ES for these two genes' peptides to be 0. For illustrative purposes, we will only consider predicted epitopes with an ES of at least 1 in the subsequent distribution. This reduces the total set drastically by only considering 1534 epitopes derived from 17 genes (**Figure 28**).

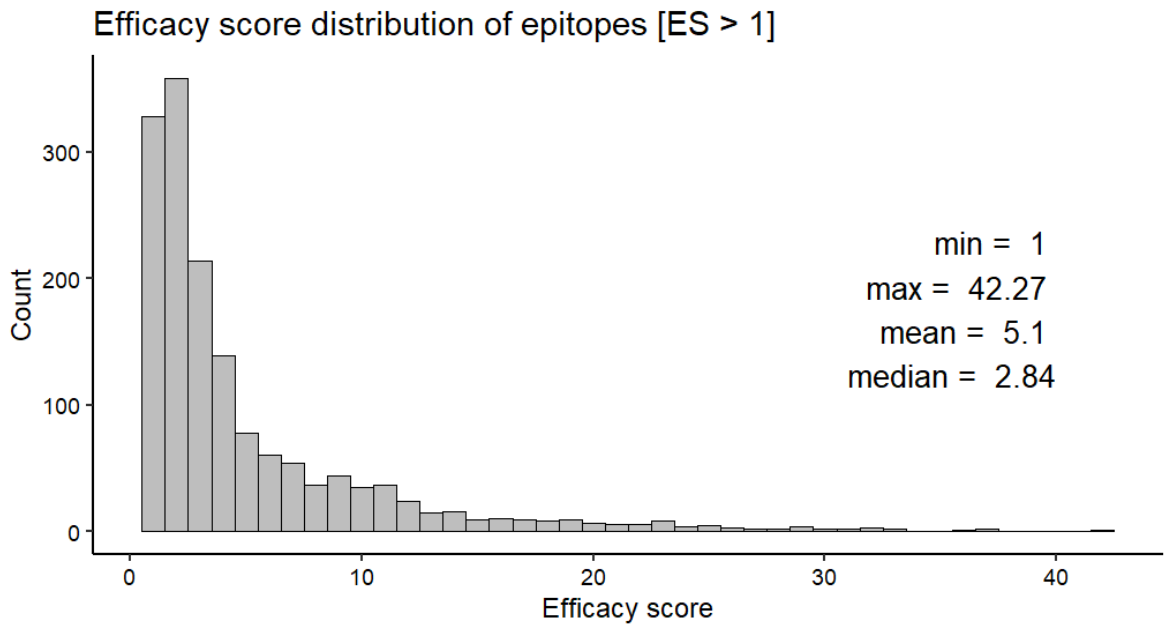


Figure 28: ***Distribution of efficacy scores of epitopes with a score higher than 1.*** *Most epitopes still score comparatively low in our ES score. Percentile-wise, 99% of the considered epitopes we scored below an ES of 28.84. The minimum was set to 1 to reduce the number of close-to-zero epitopes skewing the distribution too strongly. The maximum ES was 42.27. The mean and median were both in low one-digit percent ranges.*

Since a principal goal is to provide both individual epitopes and tumor-specific genes for further analysis, we investigated how the ES was distributed over the 17 remaining (above 0 ES scoring) candidate genes. We also explored where the top-scoring epitopes originated from in terms of source genes and potential overall relevance for UM. When considering all genes, we found that most candidates offered a broad range of potential epitopes to select from. However, *MLANA*, a known immunogenic MCM gene, had a relatively low score range (1 to 15.81) compared to the other prominent MCM antigens like *PMEL* (1 to 31.05) or *TYR* (1 to 36.55)(**Figure 29**)**.** Few of our discovered genes were directly relatable to UM. Recently the *TMEM200C* was identified to be a potential marker for progression in UM (Ness *et al.*, 2021). While it scored in the lower ranges in this analysis, it might present a worthy target for further investigation.

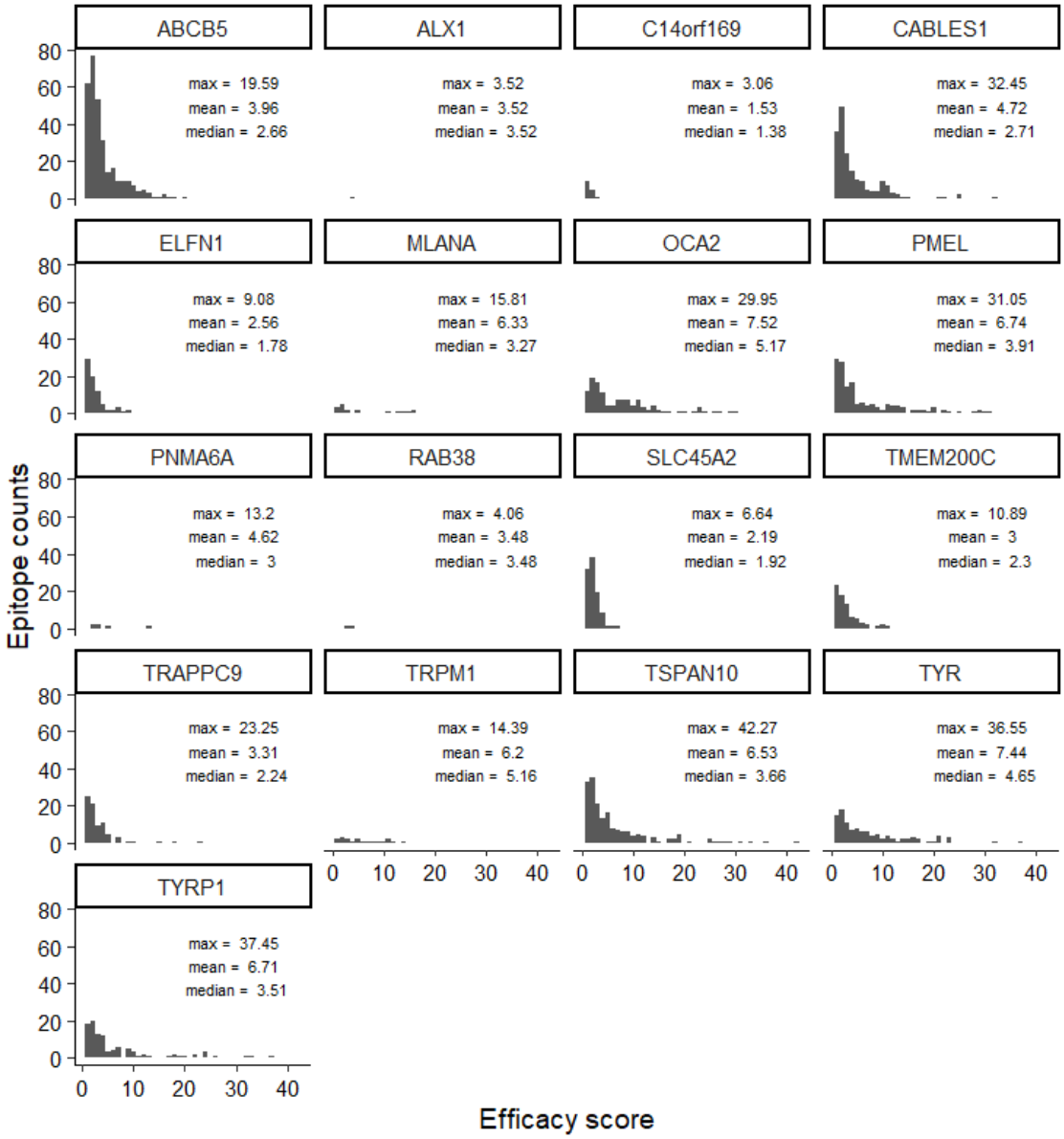


Figure 29: ***Distribution of efficacy scores per gene****. Both the y and x axis have been fixed to the same intervals row-wise and column-wise for easier comparison.*

While the overall distributions showed many options for targeting UM, we were interested in the best-predicted epitopes. Hence, we extracted the top 1% of epitopes for a closer investigation. The cutoff for the top 1% of epitopes was calculated as 28.81. Sixteen epitopes generated from 6 genes cleared this threshold (**Table 7**).

Table 7: ***List of the top 16 epitopes with their gene of origin, predicted preferred allele, and overall efficacy score.***

Peptide	**Gene name**	**HLA allele**	**Efficacy Score**
GLLALAIGL	*TSPAN10*	HLA-A*02:01	42.26
IAIAVVGAL	*TYRP1*	HLA-C*16:01	37.45
MLLAVLYCL	*TYR*	HLA-A*02:01	36.55
SLLGLLALA	*TSPAN10*	HLA-A*02:01	35.69
CIFFPLLLF	*TYRP1*	HLA-A*29:02	33.32
FPFSLLGLL	*TSPAN10*	HLA-B*3501	32.94
MAAAAAAAT	*CABLES1*	HLA-B*3501	32.45
SLGCIFFPL	*TYRP1*	HLA-A*02:01	32.03
LAVLYCLLW	*TYR*	HLA-B*57:01	31.52
AVIGALLAV	*PMEL*	HLA-A*02:01	31.05
ALGGLVVSA	*TSPAN10*	HLA-A*02:01	30.62
MAVVLASLIY	*PMEL*	HLA-B*35:01	29.98
FPMMVVSCTV	*OCA2*	HLA-A*02:01	29.95
AVVLASLIY	*PMEL*	HLA-A*29:02	29.09
FSLLGLLAL	*TSPAN10*	HLA-C*16:01	29.08
LMYALAFGA	*OCA2*	HLA-A*02:01	28.96

Interpreting the ES as a probability, our top candidate, derived from the gene *TSPAN10*, had an overall probability of 42% of being efficacious*. TSPAN10*, from a biomedical perspective, would pose an interesting target since it has been associated with cell migration and metastasis (Seong *et al.*, 2012). A therapy that would exclusively target this in an MCM or UM clinical setting may inhibit cancer mobility and curb the risk of metastasis. *TSPAN10* generated five highly efficacious epitopes, making it a prominent element in our selection for three different alleles; with one being HLA-A*02:01, broad applicability may be feasible for this antigen. The runner-up candidate gene, which produced three high-ranking epitopes, was *TYRP1*. This melanocyte-specific gene is a potential TAA and has been investigated in some clinical trials using monoclonal antibodies with limited success in relapsed or refractory MCM patients (Kobayashi *et al.*, 1999; Khalil *et al.*, 2016). *TYR*, another known MCM gene, ranked third among our top candidates, producing two highly-ranked epitopes for two different alleles. Like *TYRP1*, *TYR* plays a role in pigmentation, is considered melanocyte-specific and has been under translational investigation in the context of MCM (Bentley, Eisen and Goding, 1994; Wolchok *et al.*, 2007; Parlar *et al.*, 2019). *CABLES1* presented with only one highly ranked candidate and is generally not associated with UM or MCM other than being a potential carrier of a driver mutation site in MCM (Dbniak

et al., 2019). Biologically, the gene is involved in cell cycle regulation, having known interactions with several cyclin-dependent kinases (Sakamoto *et al.*, 2008). Curiously, this candidate epitope with the low-complexity peptide sequence `MAAAAAAAT` is restricted to the HLA allele C*16:01, which is relatively rare in Caucasian populations with an overall allelic frequency of 4.1% but prevalent in African populations, e.g., 28.3% in Mali (Middleton *et al.*, 2003; Gonzalez-Galarza *et al.*, 2020). The fifth gene, *PMEL*, another pigmentation gene and MCM antigen, presented itself as a highly-ranked candidate in the context of UM (Zhang *et al.*, 2021). It gave rise to three highly efficacious peptides on three distinct alleles. Finally, the gene *OCA2* produced two efficacious peptides for the allele A*02:01. This gene is attractive as a target for direct epitope intervention and other therapeutic options since it is a transmembrane protein playing a role in pigmentation and melanin synthesis (Sajid *et al.*, 2021). Indeed, this gene is a prognostic and predictive marker for cutaneous MCM and primary UM.

After exploring the biological and biomedical properties of the efficacious epitopes, we checked how the five features in the model contributed to the ranking. Since the model will likely be further expanded or modified, we analyzed how the individual features correlated with poorly ranked (a score of absolute zero) and positively ranked epitopes to inform possible modifications of the models (**Figure 30**).

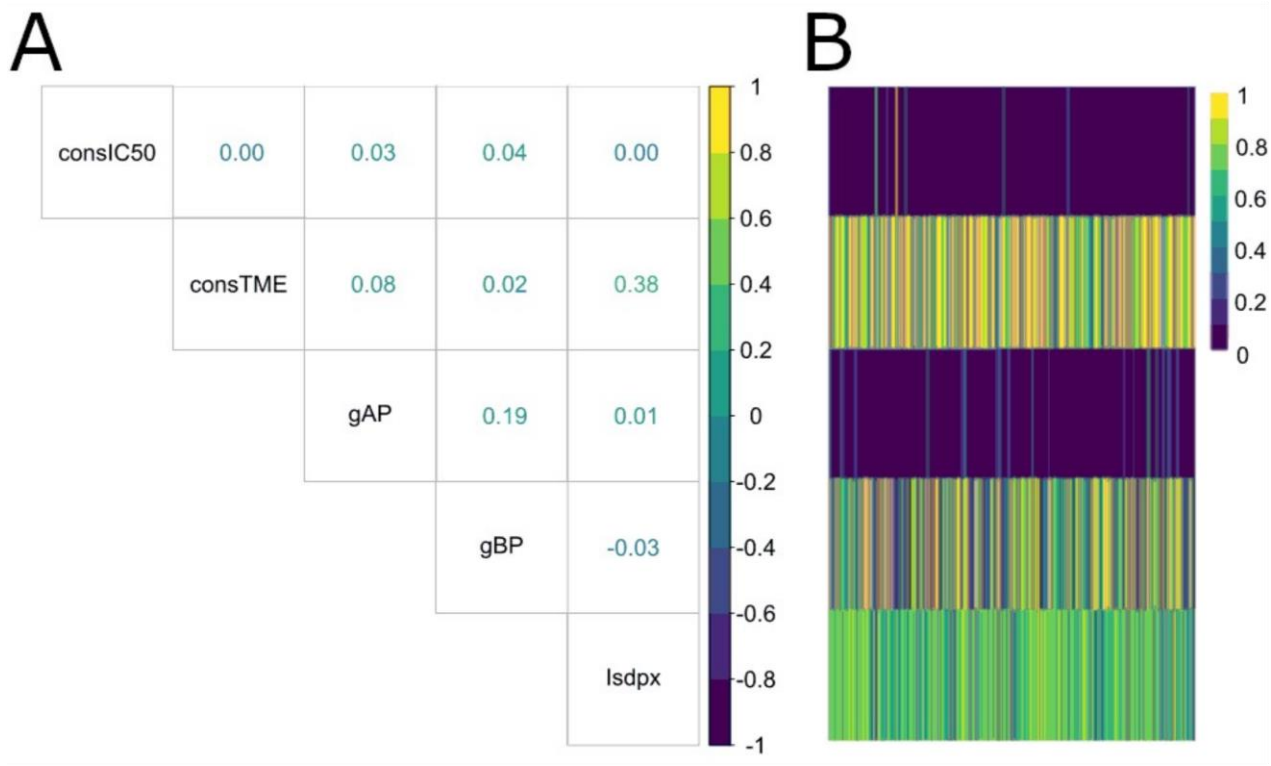


Figure 30: ***(A) Co-correlation matrix of constituents of the ES for all zero-ranked epitopes together with (B) a heatmap illustrating overall distribution. Since the ES is 0 for all selected epitopes, it is excluded from this illustration.***

We found that in contrast to our simpler gPIE model, two factors governed the removal of epitope candidates. Predominantly, the constrained IC50 and the generalized activity probability (gAP) returned all zero values. Of close to 1.43 million zero-ranked peptides, 1.41 million had zero consIC50. The gAP was the second leading cause, with 1.07 million epitopes allotted a zero. The overlap between these two groups was high, with 1.05 million epitopes being zero in consIC50 and gAP. All other factors were negligible, with consTME always being greater than zero by design, the ldspx being zero in 39528 cases, and the generalized binding probability being zero in 129528 cases. Looking at favorably scoring epitopes, the gAP predictor still had a more substantial influence on the overall ES than other factors. Consequently, a modest positive correlation of 0.52 between the ES and the gAP was observed (**Figure 31 A**). In the larger context, no other single feature strongly influenced the ES, with all of them contributing relatively equally (**Figure 31 B**).

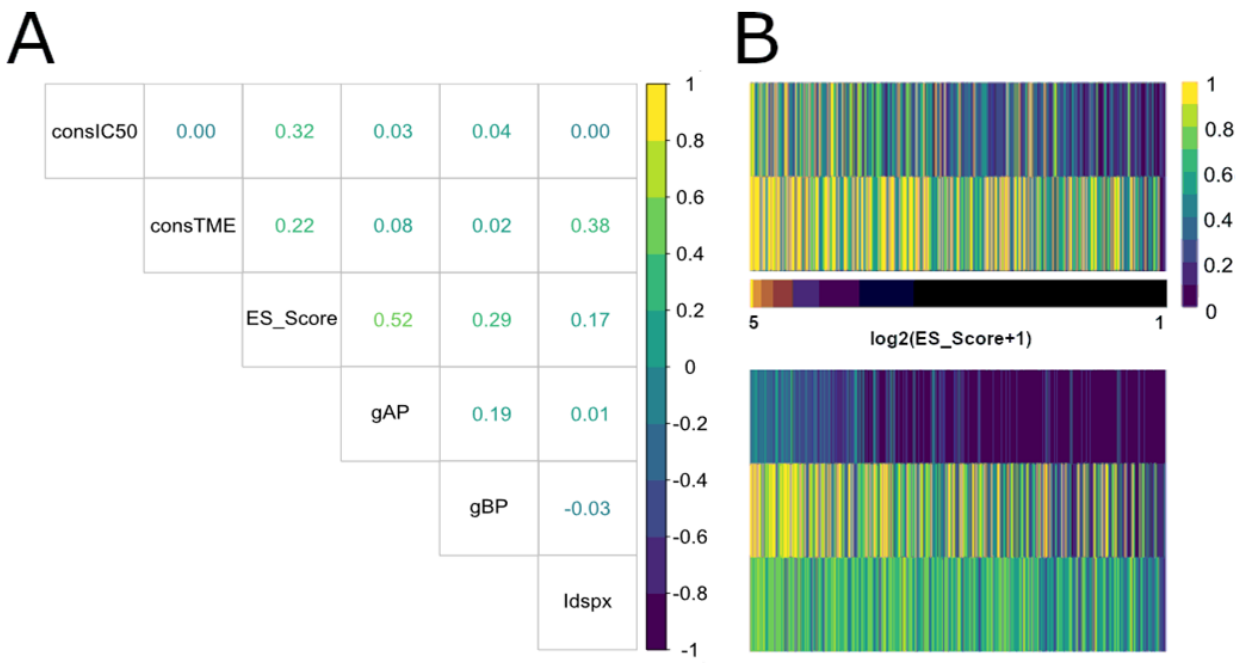


Figure 31: ***(A) Co-correlation matrix of constituents of the ES for all non-zero-ES epitopes together with (B) a heat map illustrating overall distribution and log2-transformed ES.***

4.2.4 Validation of peptide efficacy *in silico* and *in vitro*

Since testing the complete set of 1.4 million ranked MHC-I epitopes is practically impossible, we selected a small subset of candidates for validation. For practical purposes like donor availability, we restricted ourselves to peptides specific to the regionally frequent HLA allele HLA-A*02:01. In accordance with our project partners, we designed experimental procedures that test three different groups of candidates partitioned into different tiers of efficacy. We selected twenty high-efficacy (HE), twenty low-efficacy (LE), and twenty control peptides that were only predicted by an alternative predictor (netMHCpan4.0) (AP) (**Table 8**). To minimize noise from potential cross-presentation on other expressed HLA alleles, we selected the peptides to have minimal efficacy scores for other alleles besides HLA-A*02:01.

When investigating if the contributing variables of the ES were predominantly governed by specific features (e.g., polarity), we observed only weak dependencies (**Figure 30 and Figure 31**). This averts the fallacy that a subset of its contributing variables strongly dominates the ES while others are neglected (**Figure 33**). Since our efficacy tiers were based on a score that abstracted a complex biological process into numerical, sequence-depending features, we wanted to compare our predictions with other *in silico* methods that model structural and spatial aspects of the binding between peptides and the MHC-I. Courtesy of the Gupta group at the SBI at Uni Rostock, we were supplied with molecular docking simulations for our selected epitopes. The simulations were carried out in a blinded manner, i.e., without knowledge by the operator of the tier assignment for each of the 60 peptide sequences supplied. Statistical analysis of the extracted free energy values (in kcal/mol) demonstrated that they significantly differed between the HE and the other two tiers. The HE tier was generally characterized by stronger binding (**Figure 32**).

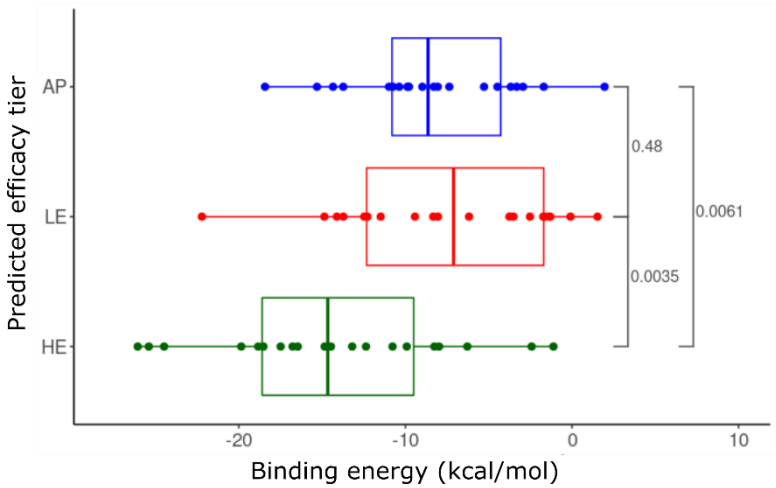


Figure 32: ***Predicted binding energies between MHC (allele A*02:01) and selected peptide candidates grouped by tiers.*** *Docking and molecular dynamics simulations for the 60 peptides were performed blinded for tier assignment. On the right-hand side, uncorrected p-values for pairwise Mann-Whitney U tests are shown, indicating that, on average, the high efficacy (HE) peptide tier formed energetically stabler complexes than the other two tiers. Lower binding energies represent more favorable peptide-MHC-I pairs.*

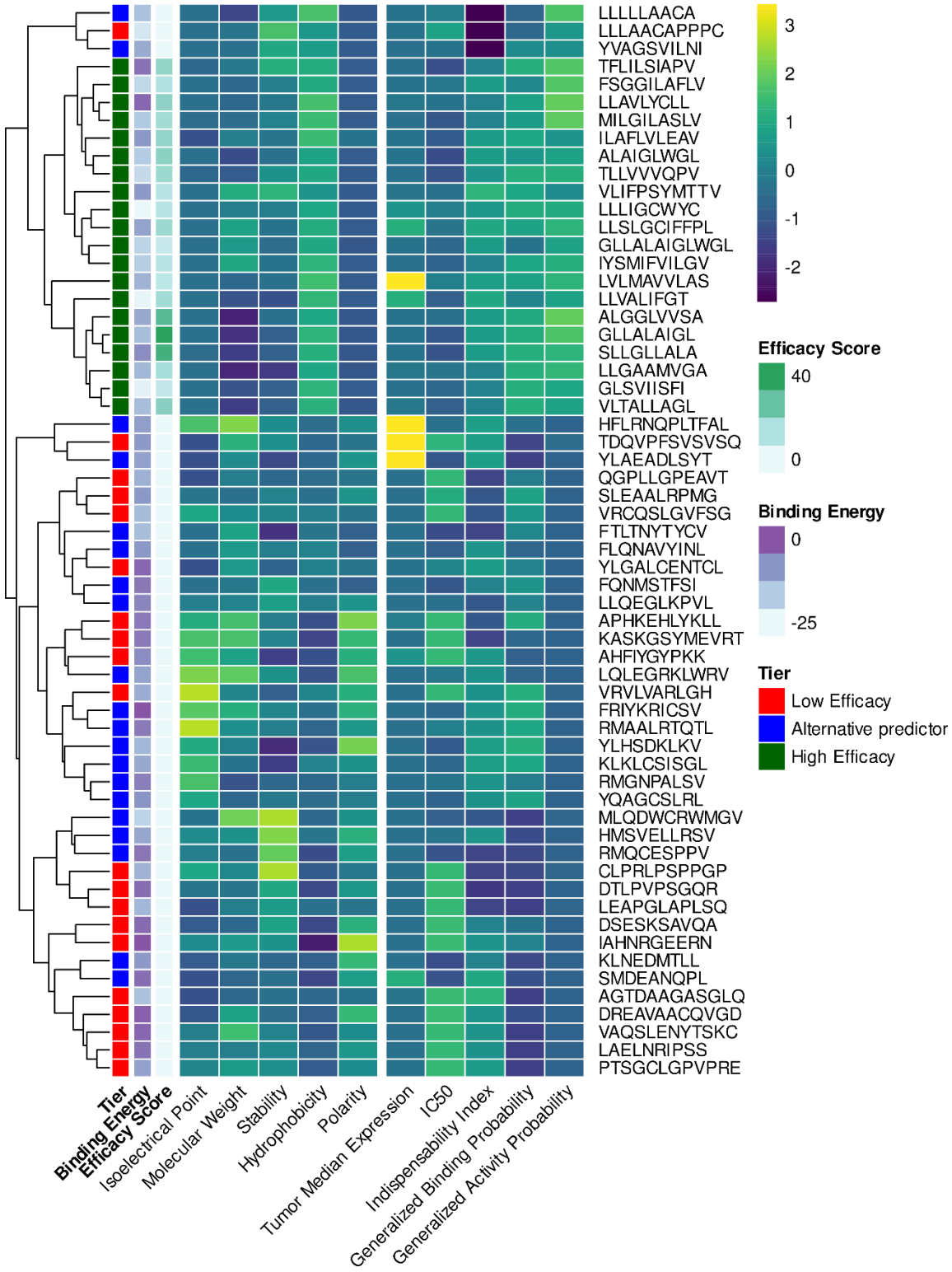


Figure 33: ***Heat map visualizing patterns in the factors contributing to the efficacy scores of the 60 selected peptide candidates****. The columns show the physicochemical peptide features used to train the binding and activity predictors (left) and the factors in the efficacy score (ES) equation (right) after z-score transformation. The IC50 column holds the peptides' netMHCpan-predicted binding affinity to MHC for the HLA-A*02:01 allele. Rows are labeled with the peptide's amino acid sequence and annotated with the ES, the computationally calculated binding energy to MHC (A*02:01), and the allocated ES tier. The high-efficacy group peptides are characterized by high hydrophobicity, an observation that is in line with established knowledge.*

Using a GMP-compliant procedure, our collaborators from the Bruns group at Medizinische Klinik 5, Universitätsklinikum Erlangen, obtained antigen-specific T cells through peptide stimulation of leukapheresis products (Gary *et al.*, 2018). All 60 peptides we had previously allocated to the efficacy tiers were ordered from a commercial service provider. Three peptides failed synthesis at this point (**Table 8**). To streamline experimental demand, we prepared peptide mixtures such that each tier of twenty peptides was split into four pools of up to five, yielding 12 peptide pools. In the HE tiers, we assigned peptides to the pools in order of descending efficacy. The pools of peptides were assigned random labels from 1 to 12 and provided to the experimental team without giving further information about the efficacy tier. This was done to ensure blinded experimental conditions. PBMC preparations from four HLA-A*02:01-positive and CMV-seropositive healthy blood donors were stimulated with each peptide pool for nine days (**Figure 8**). PBMCs were stimulated with CMV-pp65 as a positive control, while unstimulated PBMCs were used as a negative control. To quantify the antigen-specific T-cell activation, we measured interferon-gamma (IFN-γ), a surrogate marker for the activation process, in two different modalities. First, nine days after the initial stimulation, the frequency of IFN-γ-producing T cells within the whole PBMC culture was measured by flow cytometry. Even though our measurements showed substantial inter-donor variability in the responses, we measured an increase in antigen-specific, IFN-γ positive T cells when stimulated with pool HE4 (28%±10) expression compared to the positive control (**Figure 34 A**, **B**).

None of the other tested pools showed statistically significant differences. We corroborated this data by recording IFN-γ concentration in the culture's supernatant by enzyme-linked immunosorbent assays (ELISA).

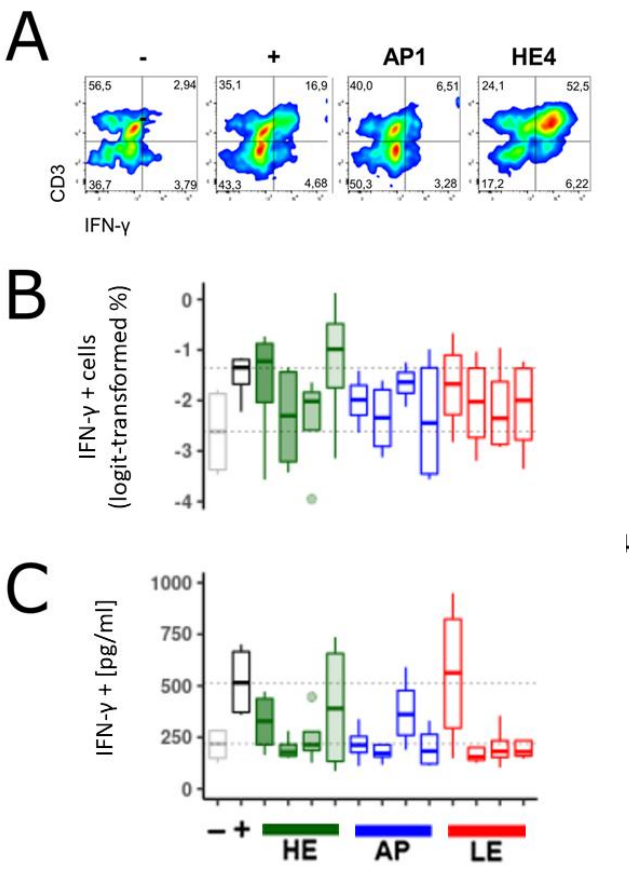


Figure 34: ***FACS and ELISA analysis of stimulated PBMCs. (A)*** *Cells stained for CD3 and IFN-γ on day 9 after stimulation with controls or with two peptide candidate pools from the high-efficacy (HE4) or alternative-predictor (AP1) groups, respectively. Numbers in corners indicate the subpopulation size in the corresponding quadrant expressed as percentage of all plotted cells.* ***(B, C)*** *Box plots of* ***(B)*** *FACS-derived IFN-γ secretion assays and* ***(C)*** *ELISA-derived IFN-γ concentration in culture supernatant (both n=4). Pools of HE peptides are sorted and colored according to decreasing score. The dashed horizontal lines extend the medians of positive and negative controls, respectively, for visual comparability. In (B), percentages of IFN-γ positive cells were logit-transformed before visualization.*

Out of the four HE-stimulated pools, two showed evidence of increased IFN-γ secretion. Only one of the four AP and LE tiers, respectively, showed higher levels of IFN-γ (**Figure 34 C**). Since IFN-γ secretion may be strongly dependent on individual characteristics of the donor and/or cell viability, we also decided to test the direct protective T cell response against a tumor model in a functional cytotoxicity assay. Hence, we co-cultured peptide pool-stimulated and -expanded T cells with the HLA-A*02:01-positive UM cell line 92.1 and followed the extent of cell death with live microscopy (**Figure 35 A**). Again, CMV-pp65 peptide-expanded T cells were used as a positive control due to their known cross-reactivity with 92.1-expressed tyrosinase and CMV's tissue tropism, which includes the choroid, UM's tissue of origin (Sugita *et al.*, 2007; Griewank *et al.*, 2012; Xu *et al.*, 2020). In line with their activated state and ability to secret IFN-γ, expanded T cells from the most reactive HE pool (HE4) showed more cytotoxic activity against the UM cells compared to the negative control and the AP1 tier after 10 hours of culture. A difference in apoptotic cell area between HE4 and LE1 was visible after 24h of co-culture (**Figure 35 B**). To check our hypothesis that these T cells would be self-tolerant, we co-incubated them with autologous macrophages. The T cells did not show any cytotoxic activity; however, due to low cell numbers, we could only perform this experiment for one donor (**Figure 36**). Our data suggests that our HE tier-expanded T cells do not cause measurable off-target cytotoxicity, at least against autologous APC populations. Further, our efficacy score (ES) can select peptides that can activate and stimulate intended T cell populations in an antigen-specific manner under *in vitro* conditions with measurable tumoricidal activity and tolerance for other autologous cell populations.

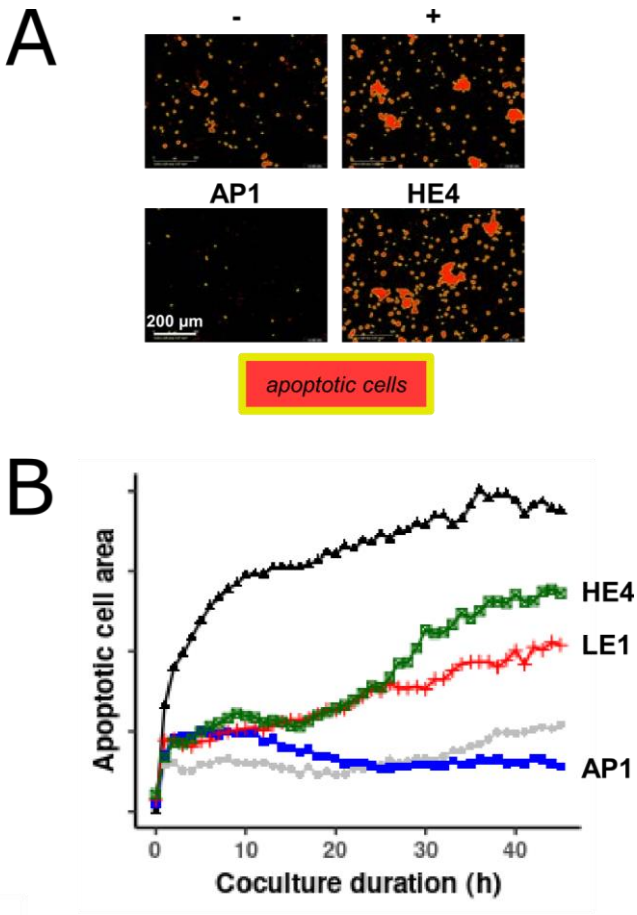


Figure 35: ***Cytotoxicity analysis with representative images taken during live imaging and quantified measurements of apoptotic cell area.*** *(A) Fluorescence images taken during the cytotoxicity assays in which stimulated PBMCs were co-cultured with the UM cell line 92.1. Red regions surrounded by yellow borders were identified as dead cells via image analysis software. (B) Time series analysis and quantification of apoptotic cells in the cytotoxicity assays, as illustrated in panel A. Shown are the averages of three independent experiments with different donor material. HE, high efficacy tier; LE, low efficacy tier; and AP, alternative predictor tier.*

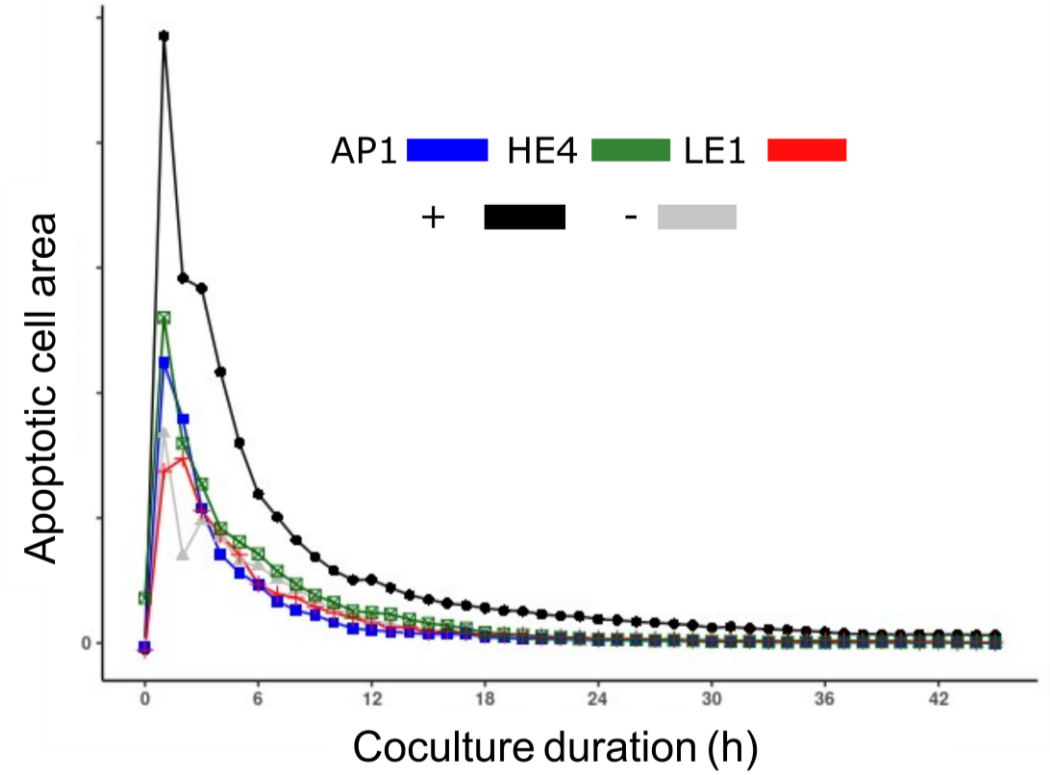


Figure 36: ***Time series quantification of cytotoxicity assays of stimulated T cells co-incubated with autologous macrophages.*** *Pool HE4, which induced T-cell expansion after stimulation and a notable cytotoxic activity against a UM cell line, did not induce measurable cytotoxicity towards autologous macrophages. This may indicate that the stimulation of T cells with our selected peptides indeed produced a self-tolerant response. The initial rise in fluorescence values at the start of the experiment is the consequence of a technical artifact. Due to scarcity of donor material, this assay was only performed once.*

Table 8: ***Peptide candidates selected for experimental validation.*** *Each major column lists peptides of one tier, which are partitioned into pools of size five. High-Efficacy peptides are sorted by efficacy score. ES – efficacy score (unitless). ΔE – a free-energy gain of MHC-peptide complex after computational docking as calculated by RDOCK (kcal mol^{-1}). * – peptide failed synthesis and was absent from the pool during experimental tests.*

	High Efficacy (HE)			Low Efficacy (LE)			Alternative Predictor (AP)		
	Sequence	**ES**	**ΔE**	**Sequence**	**ES**	**ΔE**	**Sequence**	**ES**	**ΔE**
	GLLALAIGL	42.3	-14.9	LEAPGLAPLSQ	0.0	-13.7	YLAEADLSYT	0.0	-9.8
	SLLGLLALA	35.7	-6.3	VAQSLENYTSKC	0.0	-1.3	SMDEANQPL	0.0	-1.7
1	ALGGLVVSA	30.6	-10.8	AHFIYGYPKK	0.0	-6.2	LLQEGLKPVL	0.0	-5.3
	VLTALLAGL	21.2	-14.5	DTLPVPSGQR	0.0	-1.7	MLQDWCRWMGV	0.0	-18.4
	ALAIGLWGL	20.9	-16.8	LLLAACAPPPC	0.0	-22.2	FTLTNYTYCV	0.0	-14.3
	LLAVLYCLL	20.6	-1.1	TDQVPFSVSVSQ	0.0	-8.3	YVAGSVILNI	0.0	-11.0
	TFLILSIAPV	19.6	-2.4	CLPRLPSPPGP	0.0	-12.5	KLNEDMTLL	0.0	-8.3
2	ILAFLVLEAV	19.1	-8.0	SLEAALRPMG	0.0	-8.0	YLHSDKLKV	0.0	-13.7
	TLLVVVQPV	18.0	-19.9	DREAVAACQVGD	0.0	-0.1	RMQCESPPV	0.0	-2.9
	LLSLGCIFFPL	18.0	-9.9	YLGALCENTCL	0.0	-1.4	HMSVELLRSV	0.0	-10.4
	*LLVALIFGT	16.9	-25.4	VRVLVARLGH	0.0	-11.5	FLQNAVYINL	0.0	-8.1
	MILGILASLV	16.7	-16.5	LAELNRIPSS	0.0	-2.5	FQNMSTFSI	0.0	-3.3
3	LLGAAMVGA	15.9	-13.2	APHKEHLYKLL	0.0	-3.5	RMGNPALSV	0.0	-4.5
	FSGGILAFLV	13.6	-18.8	IAHNRGEERN	0.0	1.5	LQLEGRKLWRV	0.0	-9.9
	LLLIGCWYC	12.5	-26.1	DSESKSAVQA	0.0	-1.6	LLLLLAACA	0.0	-15.3
	*IYSMIFVILGV	11.8	-17.5	PTSGCLGPVPRE	0.0	-9.4	YQAGCSLRL	0.0	-7.4
	VLIFPSYMTTV	11.5	-8.3	KASKGSYMEVRT	0.0	-3.8	HFLRNQPLTFAL	0.0	-9.0
4	*LVLMAVVLAS	11.2	-12.4	AGTDAAGASGLQ	0.0	-14.9	RMAALRTQTL	0.0	-3.7
	GLSVIISFI	9.9	-24.5	VRCQSLGVFSG	0.0	-14.1	FRIYKRICSV	0.0	2.0
	GLLALAIGLWGL	9.2	-18.5	QGPLLGPEAVT	0.0	-12.3	KLKLCSISGL	0.0	-10.7

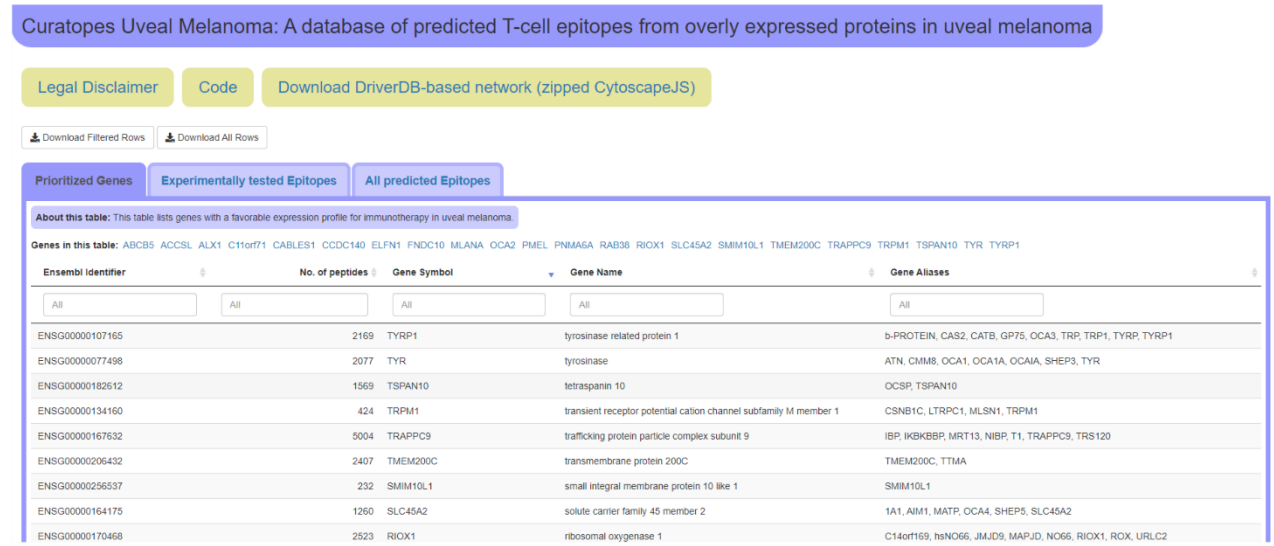
4.2.5 Curatopes 1.5 – Database design and functionality

The Curatopes 1.5 Uveal Melanoma database features the same basic structure and functionality as Curatopes 1.0, designed for MCM. We made some design changes that either resulted from the different structure of the data or would improve user experience (**Figure 37**). The database is accessible under https://curatopes.com/uvealmelanoma and is currently embargoed until the corresponding publication has finished the peer-review process and requires access credential (user: reviewer, password: curuvom2022). When using the database, one has to agree to a legal disclaimer concerning using the peptides in clinical settings. Additionally, the code is made available in a public repository, and finally, one can download the background network based on curated oncogenes (**Figure 37 A**). Since, in the context of molecular tumor boards or biomedical research, gene symbols have recognition value in contrast to machine-readable, version-controlled identifiers commonly used in bioinformatics, we received feedback that putting genes and their symbol as the first information layer would improve the user experience.

The first page lets the user explore different aspects of the genes, such as how many peptides were derived from each gene, its full name, and known aliases. This is especially helpful since different communities or researchers may refer to genes differently; for example, the prominent cancer treatment target Programmed Cell Death 1 gene is referred to as PD1 by the medical community while it also might be known as CD279, even though the official name is *PDCD1* (https://www.genenames.org/data/gene-symbol-report/#!/hgnc_id/8760).

By clicking on a gene symbol, the user is thus taken directly to the GENECARD gateway for further available meta information. The second table of the database consists of the 60 peptides we tested in our *in vitro* setting (**Figure 37 B**). We provide all three tiers – high efficacy (HE), low efficacy (LE), and alternative predictor (AP) – for exploration. Each peptide is shown with its gene of origin, AA sequence, predicted tier, HLA allele, and median tumor gene expression in TPM. Users can add the allele-specific predicted binding affinity, the indispensability index, the generalized binding prediction, the generalized activity prediction, and the peptide length. As with Curatopes 1.0, users can download and query the database or subsets as needed. As a final table, we provide all three epitopes per gene allele combination with the features used in constructing the database.

A



B

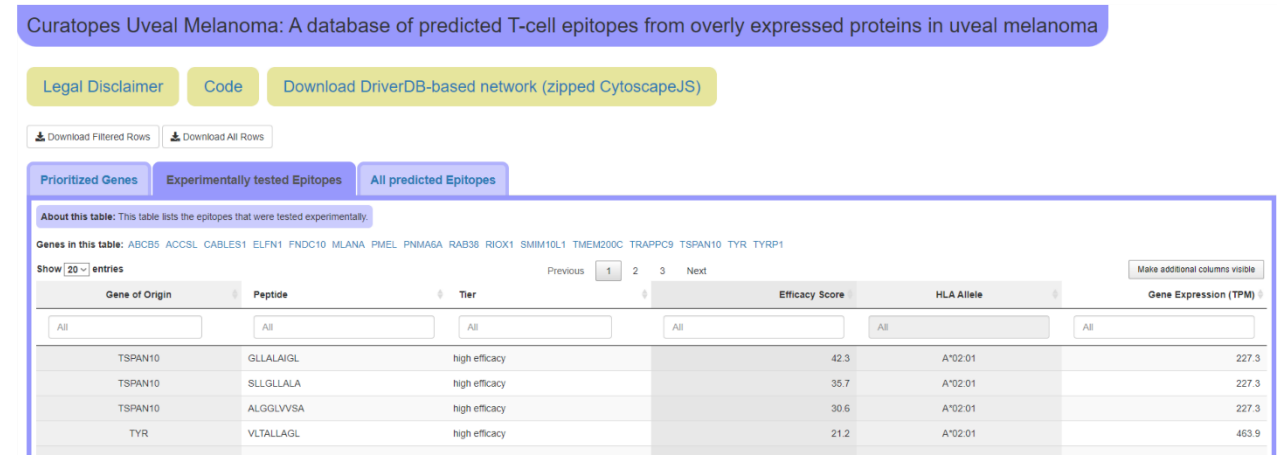


Figure 37: ***Screenshots of the (A) gene overview and (B) tiered-peptide table on the Curatopes 1.5 website for tumor-associated antigens in uveal melanoma.*** *The second table holds the 60 tested peptides and their corresponding tiers. All relevant data, such as gene expression or physiochemical features, may also be shown. Users are informed by a legal disclaimer and can access the code deposited on a public git repository and the background network (DriverDB-based network), which can be downloaded and used for further investigation.*

.

5 Discussion

Tumor-associated antigens in dermato-oncology

Metastatic cutaneous melanoma (MCM) and metastasized primary uveal melanoma (UM) are diseases for which we need efficacious treatment options, be they mono or adjuvant therapies. While there have been improvements in treatment options for MCM patients with the advent of immune checkpoint inhibitors (ICB), complete response rates remain below 20% (Curti and Faries, 2021). Conversely, in UM, these rates are even lower, with overall survival rates for patients teetering in the range of slightly over a year with minimal responses to ICB (Wessely *et al.*, 2020; Elias A.T. Koch *et al.*, 2022). However, while we choose these tumor models due to clinical association, many other cancers share similar prognostic and clinical characteristics. According to the latest Global Burden of Disease Study, cancer is still the second leading cause of death word wide, with fatalities in every age bracket (Abbafati *et al.*, 2020). Thus, methods to leverage the increased availability of targeted immunotherapies (IT) are an important area of research. Especially after the advent of the next generation of vector systems, like mRNA vaccines or antigen-loaded immune cells, the hunt is on for suitable targets (Buonaguro and Tagliamonte, 2020).

Even though there are tremendous amounts of public data gathered today for many different cancers in projects like "The Cancer Genome Atlas" (TCGA, https://www.cancer.gov/tcga), there is a distinct lack of pipelines available which can leverage this data to help discover novel targets for scientific or clinical use (Aran *et al.*, 2017). Hence, in this project, we developed a novel workflow to predict anti-tumor antigen candidates for all tumors for which cohort expression data is available. We then rank order these candidates according to predicted efficacy. We present this algorithm and its derived databases for our two tumor models. Our principal approach is to offer multiple options for efficacious targets for further clinical or *in vitro* investigation, specifically optimized to address the issues of autoreactivity, self-tolerance, and antigen loss. During the first phase of developing the Curatopes pipeline, deemed version 1.0, we designed a system to return candidates for therapy while negotiating the present risk of autoimmunity (Tran *et al.*, 2013; Chodon *et al.*, 2014).

Our analysis found that some epitopes from a clinically and scientifically known MCM TAA were highly ranked in the superior-tolerance set of version 1.0. Regarding the gPIE, the maximum value was generated by an epitope for the allele HLA-A*31:01 and the known MCM tumor-associated antigen (TAA) *MAGEA3.* In contrast, other known antigens were filtered out and generally scored poorly in the gPIE (**Figure 15**). This begs the question of why this occurred. Many known TAAs showed very low tumor restrictiveness in the expression index, suggesting similar expression values in healthy and tumor tissue while also having low predicted affinities (**Figure 16, F1 and F4**). Since many previously known antigens were discovered before high-throughput transcriptomics, a reasonable explanation might be that with increasing sample sizes and sequencing resolution, residual expression of the source genes could be detected by our pipeline, which was not possible at the time of their discovery (**Table 3**, publication dates). Even with the discussed issues, our predicted candidates

should be considered since *MAGEA3* is an important melanoma TAA, and our method's replication of its discovery lends credence to our system **(4.1.3** and **Figure 15)**. While the gene is known and under clinical trial investigation (seven active trials when searching for *MAGEA3* at www.clinicaltrials.gov, last accessed 02.02.2023), our suggested peptide is yet unknown. Thus, we propose that while the net being cast may have large holes, our algorithm may yet produce novel targets. However, with a lack of experimental validation at this stage, we concede that this is speculative.

For antigens predicted by our pipeline without prior knowledge, cohort restrictions are a factor since we could only source 32 melanoma samples with access to raw data (nucleotide sequencing read). Most of our predicted antigens received a gPIE of zero because of no source transcript expression in the tumor (**Figure 16, F3**). Thus, a more extensive data basis could be helpful on the algorithm side. However, access restrictions to patient-derived samples make it hard to collect large cohorts from one source. Public data often suffer from unclear annotation, unknown processing steps, or batch effects caused by their diverse origin (Leek *et al.*, 2010).

The lack of access to primary sequencing data led us to forgo this requirement in the uveal melanoma (UM) database creation phase, and we relied on previously processed expression data. These datasets are made available through consortia like TGCA and allow access to a large set of well-annotated cohorts at the cost of losing control over the primary data processing for patient security purposes. Hence this data cannot always be guaranteed to be comparable with other datasets, e.g., data generated in-house (**Figure 21**). Nevertheless, we feel the trade-off is well worth it, and in-house data may still be used for standalone analysis or validation, an equally important step (**Figure 22, B**).

Few prior works exist on uveal melanoma (UM) TAAs when considering the uveal UM-associated genes we detected. However, many genes that produced highly ranked UM-associated antigens were melanocyte-derived (**Table 7, Figure 29**). The immediate observation can be made that no melanocyte markers were found in our MCM genes, demonstrating the different developmental phases in which cancers were excised. In the case of MCM, samples were gathered from metastases since, commonly, those cases require further intervention, and in some patients, the primary site is unknown (Del Fiore *et al.*, 2021).

In UM, primary tumor samples were collected since, in this tumor, one would like to intervene before metastases occur. Especially after the development of metastases, survival rates plummet with few treatment options (Elias A.T. Koch *et al.*, 2022). Thus, a preemptive treatment option like a cancer vaccine might be a well-suited treatment modality after discovering a UM primary tumor in a patient.

Implemented filtering and scoring systems

Investigating our predicted superior-tolerant MCM epitopes from a probabilistic point of view, a score of 17.3 for the best epitope would translate to roughly 1- in-6 chances of efficaciousness. Similarly, when looking at the enhanced-tolerance epitopes for MCM, we reached a score of 52.65, marginally better than 50%, with some residual expression in healthy tissues. In the ranking of UM epitopes, the best epitope reached a score of 42.26, again translating to a meager 2-in-5 chance. These statements are under the assumption that our scores are modeling the probability of success to some degree.

One reason for relatively low and under strict interpretation close to random scores could be a somewhat expected result of our conservative filtering and scoring methodology. We are potentially filtering out other better candidates in earlier stages through the stringent transcriptomics filter (**Figure 9, Figure 22 A**).

The peptide level filter could later remove highly affine and efficacious epitopes (**Figure 11, Figure 25**). A philosophical argument can be had as to what is the better approach. The first option would be having a more extensive set of potentially highly immunogenic candidates with little consideration of presence in the body outside the tumor. The second option would be a more restrictive set that may combine several low to moderate immunogenicity candidates into one potentially potent approach. Both perspectives are valuable and should even be applied or investigated, especially in the sometimes bleakness of cancer therapy. In our study, we, however, argue for the second approach since the general set of peptides or genes to choose from remains rather large, and choices are plentiful.

Further, there are already ongoing clinical trials where immune checkpoint blockade treatment is administered parallel to an antigen-based treatment (NCT03897881). Hence, when given as an adjuvant therapy, targets with known peripheral presence in healthy tissue may pose a more considerable risk. Especially with the dangers and risks of autoimmune events being well-established in combined therapies (Champiat *et al.*, 2016; Huemer *et al.*, 2020).

Additionally, there is also a point to be made that even well-established MCM antigens that are considered to be highly immunogenic in some settings, when given in trials, like *PMEL* or *MLANA*-derived antigens, did not yield overall success and lead to tumor remission only in a few patients (Steele *et al.*, 2011; Wilgenhof *et al.*, 2011; Chandran *et al.*, 2015). Even when using highly enriched specific T cells in adoptive T cell transfer (ATC), the results were mixed, with a few patients responding but sometimes at the cost of autoimmunity (Chodon *et al.*, 2014).

Expanding on this, considering known MCM antigens in our Curatopes 1.0 analysis, their overall scores were even lower, with the maximum score being 5.49 in this set **(Figure 15, Known-antigen)**.

Indeed, the tumor restrictiveness measured by our expression index was a major cause for assigning a score of zero to the known antigens, supporting our point that historical discovery methods were not sensitive enough to detect the presence in healthy tissues (**Figure 16, F4, Known-antigen**). This was combined with a

considerably lower overall normalized IC50 for this set **(Figure 16 and Figure 17)**. Thus, when applying these antigens in therapy, our highly conservative post-hoc peptide filter may have to supplement traditionally used low throughput measurements of tumor restrictiveness.

Supporting this point is the observation that a peptide-specific highly adverse autoimmunity was observed using an assumed restricted TAA in some patients in a trial targeting *MAGEA3*. In this trial, patients were treated with an HLA-A2-specific engineered T-cell receptor against the peptide `KVAELVHFL` (Morgan *et al.*, 2013). During the study, two patients out of nine treated died due to severe neurotoxicity caused by the therapy. In the following scientific inquiry into the causes of the deaths, it was found that the TCR recognized peptide sequences that were present in other MAGE-family members. These MAGE-family members were expressed in the brainstem of the patients. In our database, which also contains *MAGEA3*-derived peptides, this specific sequence was removed by the peptide level filter due to the exact match with *MAGEA9*. The *MAGE* TAA peptide overlap illustrates that while the current problem is usually viewed from too little immune response when using targeted therapies, it is a fine line to manage (Hirayama and Nishimura, 2016). The issue can quickly shift to mitigating severe autoimmune consequences once success has been achieved in getting T cells to engage. Hence, we feel that our *a priori* conservative approach concerning avoiding the off-tumor presence of an antigen is warranted, and our method is compelling, especially considering our preliminary experimental data produced in version 1.5 in the context of UM (**Figure 36**). Since we used primary donor material for the UM experiment, we could not perform this experiment for all cytotoxicity assays. We did not have enough material available to generate donor APCs in all four assays. However, the testing demonstrates that the filtering procedure, applied in both MCM and UM, can select self-tolerant peptides. Coupled with the ES score, the tested peptides could also elicit IFN-γ secretion in stimulated T cells and induce cytotoxicity against a cell line derived from the targeted tumor entity **(Figure 34, Figure 35)**. In further experiments, it would be interesting to perform tolerance and cytotoxicity experiments in the context of MCM, especially with the *MAGEA3/MAGEA9*-derived peptide to show that these highly autoreactive peptides induce auto-aggression *in vitro*. With our pipeline capable of selecting reactive antigens and threading the balance to tolerance, we believe the cautious, conservative approach is justified.

Comparisons with similar prediction tools and pipelines

A frequent issue in bioinformatics is the "yet-another-tool" problem caused by the re-addressing of an issue believed to be already solved. This begs the question of what makes our approach different or what are the advantages, compared to well-established and known tools like netMHCpan, MHCFlurry2.0, or HLAthena (Jurtz *et al.*, 2017; O'Donnell, Rubinsteyn and Laserson, 2020; Sarkizova *et al.*, 2020). Principally, our pipeline is not a predictor of allele-specific MHC-I binding, nor should it be understood as one. Its focus is to return tumor-specific results aimed explicitly at clinical applications. The input to our pipeline is wholly transcriptomics based and requires no further specification on what alleles or peptides are of interest. It produces a multivariate output and provides in-depth information on gene expression, tolerance, physiochemical properties, and ranking for immunotherapy-based settings. The other tools mentioned highly focus on the binding affinity aspect, and they do not produce an aggregate score but generally a prediction of an empirically measurable value – a chemical affinity. Hence, we see Curatopes as a complementary approach directed at clinical application. The pipeline integrates concepts and values like affinity predictions from tools like netMHCpan directly into its scoring system (Jurtz *et al.*, 2017). Primarily during the first phase of this project, we relied on the affinity predictions provided by netMHCpan4.0 and considered it as one element of the overall gPIE.

Nevertheless, it has recently been shown that many established tools suffer from a significant false-positive rate and allele-specific prediction biases (Prachar *et al.*, 2020). Contrary to intuition, most predictors perform worst for one of the most common alleles, HLA-A*02:01, for which the most training data is available (Kim *et al.*, 2014; Prachar *et al.*, 2020). We show that our two predictors, implemented during version 1.5, which were trained allele-agnostic, would perform slightly better provided our data and testing environment (**Figure 26**). We specifically only used AA sequence-derivable features to train the models that did not include any information on what MHC-I allele the peptide would preferentially bind. We hypothesize that an allele-agnostic training procedure, while probably losing some potential peptide binders, may be more robust. This may especially be the case when predicting self-peptides as we did here, as there might be evolutionary pressure on self-peptides to be well presentable on MHC-I in the context of self-non-self-discrimination of the immune system. However, we cannot remove allele-specific affinity from the model since it remains a significant factor and is generally of high interest to the immunological and medical community that will ultimately decide whether to use an antigen. Consequently, we did not remove allele specificity, but we implicitly down-weighed this element in version 1.5 by only including it as one element of the overall ES score. The ES was constructed to mitigate a general bias towards a single factor and have a constant trade-off between all the variables considered (Tofallis, 2014). We also argue that a chained-probabilities scheme will temper false-positive rates by being overly conservative. The addition of mopeptide or hold it constant but not increase it. Thus, a multi-variate multiplicative score of five independent measurements should generate a more conservative estimate than an additive model.

Experimental validation and its design

Looking at our validation data, however, one could argue that we have overcorrected the issue of high false-positive rates in the model. In one of four low-efficacy pools, high median IFN-γ values were measured, while T cells stimulated by these pools were also quite cytotoxic toward the UM cell line (**Figure 34** and **Figure 35**). To make a case for our score design and the ES score's performance, assume for a moment a hypothetical UM A2 positive patient is supposed to receive peptide treatment in whatever form, say a peptide vaccination with a cocktail of 5 to 10 peptides, as it has been applied recently in other clinical settings (Ikeda *et al.*, 2021; Heitmann *et al.*, 2022).

Further, assume that our score would have been used to select the peptides. In this hypothetical case, the patient would have received peptides from pools and showed the capability of creating IFN-γ+ T cells at higher rates than the positive control in at least two cases (**Figure 34 B**). When comparing worst to best, between a selection based only on the AP and our ES score, it is apparent that our best pool (HE4) performed significantly better than AP1 in all measured categories.

These measurements include the expansion of CD3+ IFN- γ+ cells, the amount of secreted IFN-γ by these cells (**Figure 34**), and cytotoxicity against a UM target and higher than the negative control (**Figure 35**). However, the *in vitro* assays showed substantial inter-donor variability, posing a more significant problem in design and biology. Our pool testing design is one cause of this substantial variability and general noisiness. This setup, intended for short turnaround times in testing our predictions, likely introduced some errors that must be remedied in future setups.

One of these issues is the fact that HLA typing was only performed to confirm that the donors were A2 positive. We thus cannot rule out bystander effects even though we took measures to minimize the possibility of this occurring. However, we observed in our studies that few peptides were predicted to be bound by multiple alleles according to predicted IC50 (**Figure 14**).

Additionally, since our antigens are self-peptides, it cannot be ruled out that in some donors, the T-cell repertoire does not cover the particular peptide, or it might even have an inhibitory effect on the immune response through a regulatory CD4+ T cell clone (Selck and Dominguez-Villar, 2021). This is speculative, however, since our peptides' length is generally considered too short to bind MHC-II well, which would be required for this inhibitory action (Brown *et al.*, 1993). Also, our experiments were performed with PBMCs derived from healthy donors, not patients suffering from UM.

To gather more pre-clinical data, further experiments must be performed with patient-derived autologous T cells. Especially the results from patients under ICB treatment will be quite interesting to see. It is hard to speculate what these results will show, will there be an extreme reaction to using any of our predicted TAAs or none at all? Valuable data to gather novel insights into the interplay of antigen-based treatments and ICB therapy.

Another approach to denoise our validation assays would be to test the peptides individually instead of in pools. This would also provide the ability to characterize antigen-specific cells. With the wisdom of hindsight, pool testing may not be the optimal approach for validation. However, we expected that most of our peptides would not elicit an immune response due to the selective effect of the central tolerance and thus opted for a more time-efficient approach.

Our data showed a different result, though, and thus, an investigation into the central tolerance coverage of tumor-associated antigens would be a fascinating addition to the analysis. A feasible approach would be to investigate the expression profiles of medullary thymic epithelial cells (mTECs) responsible for managing clonal T-cell selection (Takaba and Takayanagi, 2017). One could screen our TAA candidates against these profiles to investigate if these TAAs or known drivers fly under the central tolerance's radar. Overall, T cell repertoire and clonality characteristics are issues we have not addressed in depth in this thesis. Exploring the T cell clonality of the responsive pools through single-cell sequencing would determine if the response was monoclonal or oligoclonal (Pai and Satpathy, 2021). Possible monoclonal responses to given epitopes could be beneficial for direct clinical translation, like a recently established HLA-A*02:01-specific soluble TCR (Nathan *et al.*, 2021). The transcriptomic analysis of the responsive T cells would also provide insights into the activated T cells from a phenotypical perspective helping to understand mechanisms of action.

These deeper analyses must follow the replication of the *in vitro* assay to test our stimulated T cell populations' self-tolerance to fundamentally prove one of our core concepts (**Figure 36**). Additionally, animal experiments may provide a more available, if somewhat more removed, model to test our concepts further. One possible approach could be inducing a tumor or xenograft in a mouse model and vaccinating the mice with our predicted peptides. This would require human HLA transgenic mice, which have been used historically to study T-cell responses to HLA-restricted antigens (Kievits *et al.*, 1987; Belunis, Diseases and Ricerche, 1996; Dipiazza *et al.*, 2017). It also requires fitting the tolerance prediction to the mouse proteome. Also, one may ponder the ethics of potentially inducing severe autoimmunity in animals in pre-clinical experiments if a tolerance prediction is wrong. Many MCM and UM mouse models exist, and xenografts are generally possible. However, a less risky or ethical-debatable option may be the use or organoids or organs on a chip (Kuzu *et al.*, 2015; Balakrishnan *et al.*, 2020; Richards *et al.*, 2020; Leung *et al.*, 2022).

Clinical application and feasibility of our predictions

The success of the soluble bi-specific T cell engager Tebentafusp in a Phase III clinical trial in UM demonstrates that improvements in the standard of care for UM patients are possible (Nathan *et al.*, 2021). The next task is to replicate this success for other HLA alleles. While HLA-A*02:01 has a high allele frequency (AF) in Caucasian populations, ranging from around 25% to up to 50% in some parts of central Europe, it is comparatively rare in Asian and African populations, with AFs as low as 1%, making it a minor allele in these populations (Gonzalez-Galarza *et al.*, 2020). The top-scored peptides from our candidate set also included alleles with broader distribution in terms of populations (**Table 7**). For example, the allele C*16:01 has an AF of 28% in some parts of Africa. Our pipeline can be extended to cover these cases with a larger, more extensive set of alleles. If affinity predictions can be performed for an allele, which generally requires a resolved protein sequence, our database could be extended at the cost of increased computation time (Venkatesh *et al.*, 2020; Reynisson *et al.*, 2021).

A further issue must be kept in mind that affinity predictions can return no binders at all for a given allele if one applies the cut-off 500nm/L **(Figure 12 and Figure 13)**. While it is reasonable to assume that an allele may have truly no binders in a given set, this may also be influenced by available training data for these models. With rarer alleles, studies on their preferred peptides are also scarce. Also, a general cutoff may not be the best option, and an allele-specific cutoff may yield the best results (Bonsack *et al.*, 2019). However, this is unpractical for systematic studies like ours, in which the user may be interested in alleles for which this has not yet been established or titrated (Paul *et al.*, 2013). Ultimately, this is a problem of complexity and where the focus is placed. It seems unrealistic to expect results for appropriately titrated affinities for all known HLA alleles, even in the mid-to-long term. Though even if affinity predictions would be perfect, it would, for the clinical context, still probably not be enough to select immunogenic antigens reliably. Hence our approach tries to optimize several parameters for an efficacious selection.

One reason for this is the concept of immunodominance, e.g., why only a few antigenic peptides out of many viable ones presented on different HLA alleles produce a very strong response while most do not. Despite intensive study, how a specific peptide-MHC-TCR combination produces a response remains mechanistically poorly understood. Current models suggest a combination of germline-encoded and thymically selected TCR-peptide-MHC binding preferences since evolution should converge on TCRs being able to bind MHC, and the thymus selects functional combinations (Rangarajan and Mariuzza, 2014; Szeto *et al.*, 2021). Nevertheless, we still lack a proper understanding of immunogenicity regarding the TCR-epitope interaction, both from the T-cell and the target cell perspective. Because of this and from the clinical and translational perspectives, we would argue more for the shotgun, not the precision approach. We believe supplying different epitopes for different HLA preferences covering different genes is the best option. One could potentially hit an antigen that would provoke a strong response with a beneficial feed-forward loop, like reactivating exhausted

immune cells or engaging other immune populations, reinforcing the anti-tumor response. Since one generally does not have the benefit of a highly restricted, cell population-specific antigen for non-solid tumors like CD19 or CD20, one must opt to administer several different epitopes simultaneously for tumors originating from immunologically less defined tissues. Our databases are designed with specifically this approach in mind. Both databases allow the ranking and selection of candidates of interest by different metrics and allow for the composition of a set of epitopes to match, for example, a patient's HLA profile (**Figure 18, Figure 37, www.curatopes.com**).

With the introduction of the indispensability index in version 1.5, we can now select a set of epitopes that may directly target essential tumor functions on several levels. If one considers our candidate network (**Figure 23**) as a basic circuit for UM, selecting epitopes that disrupt this network on several different network hubs should yield the best results. One might consider not just selecting the highest ES score for application but creating a well-balanced mixture of antigens, thus targeting several different axes in the tumor. For example, using our database, one could target the TAA *TYR*, ES of 21.2, with the peptide VLTALLAGL, covering HLA-A*02:01, while also targeting *PMEL* with MAVVLASLIY, ES of 29.99, covering HLA-B*35:01 and OCA2 with MVVSCTVGM, preferring HLA-C*16:01 with an ES of 16.6. Three peptides, easily synthesized at GMP grade purities and storable for the long term, cover a spectrum of antigens and HLA alleles, making them an off-the-shelf treatment solution. One could imagine the design of a cancer-specific peptide cocktail from our database. The long-term vision would be a ready-made cocktail of immunogenic, self-tolerant peptides, making them a low-risk, cost-effective, and fast turnaround therapy option.

Remaining issues with the prediction of efficacious antigens

It would benefit the design of the proposed antigenic cocktails to better understand the fundamental features of efficacy. We also tried to elucidate these features in our results. We found, however, that other than high hydrophobicity, a well-known attribute of MHC-I binders, we could not show any distinct patterns in the efficacy groups (**Figure 33**)(Altuvia *et al.*, 1994; Huang, Kuhls and Eisenlohr, 2011; Chowell *et al.*, 2015). However, considering we performed pool tests, a significant degree of resolution on individual peptides and their features is expected to be lost. This heterogeneous finding was supported by the observation that there was little cross-correlation between the constituents of the ES for all above zero-scored epitopes (**Figure 31**). At the same time, affinity and generalized activity prediction were the leading causes of zero scores (**Figure 31**). Naturally, we cannot enhance third-party tools used in our pipeline to improve our predictions. Our binding predictor reached an area under the curve (AUC) in its receiver operating curve (ROC) analysis of 0.853, making it a robust model according to current considerations of this metric (Lin *et al.*, 2008; Bonsack *et al.*, 2019). Applying the alternative predictor, netMHCpan4.0, to our testing condition produced an AUC of 0.815 (**Figure 26**). However, we can probably improve the activity predictor, even though it performed slightly better in our condition compared to other published tools. Our model reached an AUC of 0.635, making it what would generally be considered a poor but not random predictor. Comparatively, one widely applied tool reached an AUC of 0.559, making it slightly worse.

While binding prediction generally performed well, immunogenic activity prediction presents as rather difficult. The reasons why this is the case may be manifold. One could hypothesize that there are fewer unknown variables in binding as the process seems easier to measure and potentially model. However, immunogenicity may be a far more complex problem that may also be influenced drastically by milieu effects in which the T cell and its target are located. We hypothesize that the problem in training an activity predictor is the messiness of empirical data. Considering current public datasets, the annotation of immunogenicity is not sufficiently accurate or detailed in many instances. Binding is a rather binary event regarding its empirical measuring and ordinal description; a peptide is either present on MHC-I or is not. It may be expressed continuously as an affinity but abstracted and measured as a binary readout following an arbitrary threshold. However, the accurate measurement and quantification of immunological activity directly or indirectly through surrogate values is a very complex issue.

We explicitly searched for datasets that provided data on immunogenicity but found that this problem is non-trivial. We discovered that while binary labels were assigned to the data, the assays used to determine these labels were highly varied. The MHCBN database we used as training data alone had several different assays in use to assign the labels (information provided was, e.g., “CTL ASSAY”, “CYTOXCITY ASSAY”) (Lata, Bhasin and Raghava, 2009). While assays may measure the same readout in the abstract (T cell activation), it is unclear if they mean the same regarding training data for a model. For example, increased IFN-γ expression

points to antigen-specific activation of a T cell (Haring, Corbin and Harty, 2005; Shen *et al.*, 2017). Does this mean the T cell would be cytotoxic when presented with the target in an *in vivo* context?

Further, using one of the most extensive databases, IEDB, to query for immunogenicity data, one is often presented with mixed labeling schemes over several different assays that, at least from the direct interpretation of the data, do not transport the same information. For example, a peptide is annotated in the category of T cell assays as being "positive-high" for a ^{3}H-thymidine proliferation assay. In contrast, an x-ray crystallography assay is returned with the label positive in the same query for another peptide. How are metrics comparable or on similar scales? It is not obvious how to translate this into solid training data, which can be used in modeling without enormous manual curation. Considering that this query for "Any" Epitope for MHC-I in humans and T cell assays with a positive outcome returned 35,372 entries, this is not an easy task, considering that one probably has to go back to the primary source to find what exactly was measured (SMahajan *et al.*, 2018). Overall, training set availability and cleanliness are significant issues for immunogenicity prediction. The complexity is very high, and adequate biological model data is hard to generate in large quantities (Prachar *et al.*, 2020).

With antigen-based immunotherapies and anti-cancer vaccines likely to rise in importance in the coming years, we should endeavor to improve our efforts to annotate and aggregate existing data better and more precisely. Similarly, while body-map-type sequencing efforts increase and improve our transcriptomic picture of the human body, detecting all cell populations that may present off-site targets is hard or even impossible. Projects like GTEx and HPA keep improving their resolution and granularity, but systematic integration of the two is not widespread. Single-cell technology will help us characterize healthy cell populations by resolving tissue-specific gene expression while also allowing new insights into the functionality of T-cell clones stimulated by an immunogenic peptide-MHC-I combination.

In the meantime, with the results and associated databases presented in this work (Curatopes 1.0 and 1.5), we provide a solution for the rational and quick selection of TAAs and MHC-I-restricted epitopes for application in different antigen-based cancer immunotherapies. We pay special attention to predicting candidates solely based on principle data with being agnostic of prior information while balancing the probability of side effects (Lischer *et al.*, 2019).

6 Conclusion and Outlook

In this thesis, we presented two databases of tumor-associated antigens (TAAs) and their derived MHC-I-restricted epitopes. Our methodology optimizes the selection of TAAs so that, should they be targeted by any mode of antigen-based immunotherapy like adoptive T-cell transfer or anti-cancer vaccination, they are immunogenic and self-tolerant and minimize the risk of damage in healthy tissue. The selection process is designed to work agnostic of prior knowledge on the tumor and is based on a first-principal data source – transcriptomics. As a final step in our *in-silico* pipeline, we validated our antigen selection *in vitro*.

We could show that T cells stimulated by our antigens could kill a target cancer cell line. While we observed substantial inter-donor variability, our predicted high-efficacy MHC-I-restricted peptides elicited a measurable IFN-γ response by T cells. We hope the pipeline can help by finding novel antigens for tumor entities needing mono- or adjuvant therapies, like metastatic cutaneous melanoma, uveal melanoma, or other solid tumors. With the field moving quickly and companies commercializing antigen-based immunotherapies, we must provide rational and empirical target design strategies to ensure reproducibility, robustness, and safety to efficiently translate research into clinical testing.

There are several ways to improve our system in future work. First, we would like to expand experimental testing procedures to continue and extend our work with UM. Due to our pools of predicted peptides showing a highly variable response in healthy donors, we need to repeat the experiments on a single-peptide level to understand whether a particular peptide is highly immunogenic or if the pool composition plays a role. Further, performing T-cell receptor sequencing on the peptide-expanded T cells would show us if we are observing a mono- or oligoclonal response against the antigens and would allow the cloning and further study of this T-cell receptor. In the long term, this could be translated to alternative bi-specific T-cell engagers, like Tebentafusp or an engineered T cell, for treating UM.

On the MCM side, performing similar experimental validation on MCM TAAs would also be of great value since the gPIE model still lacks similar validation as the efficacy score. As approaches for novel projects, we plan to run and perform *in silico* analysis and TAA prediction for all tumor entities in TCGA for which transcriptomics data are available and accessible. We also intend to perform systematic screening studies and hypothesize that there may be shared TAAs between tumor entities, bringing off-the-shelf, cost-effective treatment through antigen-based immunotherapy into the realm of possibility.

These modes of treatment promise quick-turnaround therapy modalities for cancer and other targetable entities. As was seen during the coronavirus pandemic, mRNA or viral vector vaccines offer quick adaption capabilities to immune evasion events. This rapid cycle from bench to therapy has placed science under a new level of scrutiny by the public. The community should hence try to curtail adverse events caused by these novel systems as much as possible. We hope to contribute to this goal and the field with the methodology presented in this thesis.

7 References

Abbafati, C. *et al.* (2020) 'Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019', *The Lancet*, 396(10258), pp. 1204–1222. doi: 10.1016/S0140-6736(20)30925-9.

Abdallah, D. S. A. *et al.* (2011) 'Mouse neutrophils are professional antigen-presenting cells programmed to instruct Th1 and Th17 T-cell differentiation', *International Immunology*, 23(5), pp. 317–326. doi: 10.1093/intimm/dxr007.

Ackerman, A. L. *et al.* (2003) 'Early phagosomes in dendritic cells form a cellular compartment sufficient for cross presentation of exogenous antigens', *Proceedings of the National Academy of Sciences of the United States of America*, 100(22), pp. 12889–12894. doi: 10.1073/pnas.1735556100.

Alfei, F. *et al.* (2019) 'TOX reinforces the phenotype and longevity of exhausted T cells in chronic viral infection', *Nature*, 571(7764), pp. 265–269. doi: 10.1038/s41586-019-1326-9.

Alix, A. J. P. (1999) 'Predictive estimation of protein linear epitopes by using the program PEOPLE', *Vaccine*, 18(3–4), pp. 311–314. doi: 10.1016/S0264-410X(99)00329-1.

Allen, S. L. (1955) 'H-2f, a Tenth Allele at the Histocompatibility-2 Locus in the Mouse as Determined by Tumor Transplantation', *Cancer Research*, 15(5), pp. 315–319.

Almeida, L. G. *et al.* (2009) 'CTdatabase: A knowledge-base of high-throughput and curated data on cancer-testis antigens', *Nucleic Acids Research*, 37(SUPPL. 1). doi: 10.1093/nar/gkn673.

Altuvia, Y. *et al.* (1994) 'Sequence features that correlate with MHC restriction', *Molecular Immunology*, 31(1), pp. 1–19. doi: 10.1016/0161-5890(94)90133-3.

An, N. *et al.* (2015) 'Developmental genes significantly afflicted by aberrant promoter methylation and somatic mutation predict overall survival of late-stage colorectal cancer', *Scientific Reports*, 5. doi: 10.1038/srep18616.

Anderson, G. and Takahama, Y. (2012) 'Thymic epithelial cells: Working class heroes for T cell development and repertoire selection', *Trends in Immunology*, 33(6), pp. 256–263. doi: 10.1016/j.it.2012.03.005.

Andrews, S. *et al.* (2012) 'FastQC'. Babraham, UK.

El Ansary, M. *et al.* (2013) 'Immunotherapy by autologous dendritic cell vaccine in patients with advanced HCC', *Journal of Cancer Research and Clinical Oncology*, 139(1), pp. 39–48. doi: 10.1007/s00432-012-1298-8.

Aran, D. *et al.* (2017) 'Comprehensive analysis of normal adjacent to tumor transcriptomes', *Nature Communications*, 8(1). doi: 10.1038/s41467-017-01027-z.

Balakrishnan, S. *et al.* (2020) 'Organoids: An invaluable tool in pharmacology', *Indian Journal of Pharmacology*, 52(5), p. 422. doi: 10.4103/ijp.ijp_137_19.

Balch, C. M. *et al.* (2009) 'Final version of 2009 AJCC melanoma staging and classification', *Journal of Clinical Oncology*, 27(36), pp. 6199–6206. doi: 10.1200/JCO.2009.23.4799.

Bassani-Sternberg, M. *et al.* (2015) 'Mass spectrometry of human leukocyte antigen class i peptidomes reveals strong effects of protein abundance and turnover on antigen presentation', *Molecular and Cellular Proteomics*, 14(3), pp. 658–673. doi: 10.1074/mcp.M114.042812.

Belunis, C., Diseases, A. and Ricerche, I. M. (1996) 'HLA-DR4-IE Chimeric Class II Transgenic, Murine Class II-Deficient Mice Are Susceptible to Experimental Allergic Encephalomyelitis', 183(June).

Bentley, N. J., Eisen, T. and Goding, C. R. (1994) 'Melanocyte-specific expression of the human tyrosinase promoter: activation by the microphthalmia gene product and role of the initiator', *Molecular and Cellular Biology*, 14(12), pp. 7996–8006. doi: 10.1128/mcb.14.12.7996-8006.1994.

Bhasin, M., Lata, S. and Raghava, G. P. (2007) 'TAPPred prediction of TAP-binding peptides in antigens.', *Methods in molecular biology (Clifton, N.J.)*, 409, pp. 381–386. doi: 10.1007/978-1-60327-118-9_28.

Bhasin, M. and Raghava, G. P. S. (2004) 'Prediction of CTL epitopes using QM, SVM and ANN techniques', *Vaccine*, 22(23–24), pp. 3195–3204. doi: 10.1016/j.vaccine.2004.02.005.

Bierer, B. E. and Burakoff, S. J. (1988) 'T cell adhesion molecules.', *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 2(10), pp. 2584–90. doi: https://doi.org/10.1096/fasebj.2.10.2838364.

Blander, J. M. (2018) 'Regulation of the Cell Biology of Antigen Cross-Presentation', *Annual Review of Immunology*, 36, pp. 717–753. doi: 10.1146/annurev-immunol-041015-055523.

Blank, C. U. *et al.* (2019) 'Defining "T cell exhaustion"', *Nature Reviews Immunology*. Springer US, 19(11), pp. 665–674. doi: 10.1038/s41577-019-0221-9.

Bol, K. F. *et al.* (2016) 'Adjuvant Dendritic Cell Vaccination in High-Risk Uveal Melanoma', *Ophthalmology*, 123(10), pp. 2265–2267. doi: 10.1016/j.ophtha.2016.06.027.

Bonfield, J. K. and Staden, R. (1995) 'The application of numerical estimates of base calling accuracy to DNA sequencing projects', *Nucleic Acids Research*, 23(8), pp. 1406–1410. doi: 10.1093/nar/23.8.1406.

Bonilla, F. A. and Oettgen, H. C. (2010) 'Adaptive immunity', *Journal of Allergy and Clinical Immunology*, 125(2 SUPPL. 2). doi: 10.1016/j.jaci.2009.09.017.

Bonneau, M. *et al.* (2006) ' Migratory monocytes and granulocytes are major lymphatic carriers of Salmonella from tissue to draining lymph node ', *Journal of Leukocyte Biology*, 79(2), pp. 268–276. doi: 10.1189/jlb.0605288.

Bonsack, M. *et al.* (2019) 'Performance evaluation of MHC class-I binding prediction tools based on an experimentally validated MHC–peptide binding data set', *Cancer Immunology Research*, 7(5), pp. 719–736. doi: 10.1158/2326-6066.CIR-18-0584.

Boon, T. and Van der Bruggen, P. (1996) 'Human tumor antigens recognized by T lymphocytes', *Journal of Experimental Medicine*, 183(3), pp. 725–729. doi: 10.1084/jem.183.3.725.

Boulesteix, A. L. *et al.* (2012) 'Overview of random forest methodology and practical guidance with emphasis

on computational biology and bioinformatics', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(6), pp. 493–507. doi: 10.1002/widm.1072.

Bovolenta, L. A., Acencio, M. L. and Lemke, N. (2012) 'HTRIdb: an open-access database for experimentally verified human transcriptional regulation interactions', *BMC Genomics*, 13(1). doi: 10.1186/1471-2164-13-405.

Breedveld, A. *et al.* (2017) 'Granulocytes as modulators of dendritic cell function', *Journal of Leukocyte Biology*, 102(4), pp. 1003–1016. doi: 10.1189/jlb.4mr0217-048rr.

Breiman, L. (2001) 'Random forests', *Machine Learning*, 45(1), pp. 5–32. doi: 10.1023/A:1010933404324.

Brentjens, R. J. *et al.* (2003) 'Eradication of systemic B-cell tumors by genetically targeted human T lymphocytes co-stimulated by CD80 and interleukin-15', *Nature Medicine*, 9(3), pp. 279–286. doi: 10.1038/nm827.

Brentjens, R. J. *et al.* (2013) 'CD19-targeted T cells rapidly induce molecular remissions in adults with chemotherapy-refractory acute lymphoblastic leukemia', *Science Translational Medicine*, 5(177). doi: 10.1126/scitranslmed.3005930.

Brocchieri, L. and Karlin, S. (2005) 'Protein length in eukaryotic and prokaryotic proteomes', *Nucleic Acids Research*, 33(10), pp. 3390–3400. doi: 10.1093/nar/gki615.

Brooks, B. R. *et al.* (2009) 'CHARMM: The biomolecular simulation program', *Journal of Computational Chemistry*, 30(10), pp. 1545–1614. doi: 10.1002/jcc.21287.

Brown, J. H. *et al.* (1993) 'Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1', *Nature*, 364(6432), pp. 33–39. doi: 10.1038/364033a0.

Van Der Bruggen, P. *et al.* (1991) 'A gene encoding an antigen recognized by cytolytic T lymphocytes on a human melanoma', *Science*, 254(5038), pp. 1643–1647. doi: 10.1126/science.1840703.

Buchmann, K. (2014) 'Evolution of innate immunity: Clues from invertebrates via fish to mammals', *Frontiers in Immunology*, 5(SEP). doi: 10.3389/fimmu.2014.00459.

Buhler, S. and Sanchez-Mazas, A. (2011) 'HLA DNA sequence variation among human populations: Molecular signatures of demographic and selective events', *PLoS ONE*, 6(2). doi: 10.1371/journal.pone.0014643.

Bulik-Sullivan, B. *et al.* (2019) 'Deep learning using tumor HLA peptide mass spectrometry datasets improves neoantigen identification', *Nature Biotechnology*, 37(1), pp. 55–71. doi: 10.1038/nbt.4313.

Buonaguro, L. and Tagliamonte, M. (2020) 'Selecting target antigens for cancer vaccine development', *Vaccines*, 8(4), pp. 1–14. doi: 10.3390/vaccines8040615.

Bushnell, B. (2014) 'BBTools'. Available at: https://jgi.doe.gov/data-and-tools/software-tools/bbtools/.

Calis, J. J. A. *et al.* (2013) 'Properties of MHC Class I Presented Peptides That Enhance Immunogenicity', *PLoS Computational Biology*, 9(10). doi: 10.1371/journal.pcbi.1003266.

Carithers, L. J. *et al.* (2015) 'A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx

Project', *Biopreservation and Biobanking*, 13(5), pp. 311–317. doi: 10.1089/bio.2015.0032.

Casan, J. M. L. *et al.* (2018) 'Anti-CD20 monoclonal antibodies: reviewing a revolution', *Human Vaccines and Immunotherapeutics*, 14(12), pp. 2820–2841. doi: 10.1080/21645515.2018.1508624.

Castro, F. *et al.* (2018) 'Interferon-gamma at the crossroads of tumor immune surveillance or evasion', *Frontiers in Immunology*, 9(MAY). doi: 10.3389/fimmu.2018.00847.

Champiat, S. *et al.* (2016) 'Management of immune checkpoint blockade dysimmune toxicities: A collaborative position paper', *Annals of Oncology*, 27(4), pp. 559–574. doi: 10.1093/annonc/mdv623.

Chandran, S. S. *et al.* (2015) 'Persistence of CTL clones targeting melanocyte differentiation antigens was insufficient to mediate significant melanoma regression in humans', *Clinical Cancer Research*, 21(3), pp. 534–543. doi: 10.1158/1078-0432.CCR-14-2208.

Chang, S. C. *et al.* (2005) 'The ER aminopeptidase, ERAP1, trims precursors to lengths of MHC class I peptides by a "molecular ruler" mechanism', *Proceedings of the National Academy of Sciences of the United States of America*, 102(47), pp. 17107–17112. doi: 10.1073/pnas.0500721102.

Chang, W. *et al.* (2023) 'shiny: Web Application Framework for R'. Available at: https://shiny.rstudio.com/.

Chaplin, D. D. (2010) 'Overview of the immune response', *Journal of Allergy and Clinical Immunology*, 125(2 SUPPL. 2), pp. S3–S23. doi: 10.1016/j.jaci.2009.12.980.

Chodon, T. *et al.* (2014) 'Adoptive transfer of MART-1 T-cell receptor transgenic lymphocytes and dendritic cell vaccination in patients with metastatic melanoma', *Clinical Cancer Research*, 20(9), pp. 2457–2465. doi: 10.1158/1078-0432.CCR-13-3017.

Chou, C. H. *et al.* (2016) 'miRTarBase 2016: Updates to the experimentally validated miRNA-target interactions database', *Nucleic Acids Research*, 44(D1), pp. D239–D247. doi: 10.1093/nar/gkv1258.

Chowell, D. *et al.* (2015) 'TCR contact residue hydrophobicity is a hallmark of immunogenic CD8+ T cell epitopes', *Proceedings of the National Academy of Sciences of the United States of America*, 112(14), pp. E1754–E1762. doi: 10.1073/pnas.1500973112.

Clark, D. R. *et al.* (1999) 'T Cell dynamics in HIV-1 infection', *Advances in Immunology*, 73(73), pp. 301–327. doi: 10.1016/s0065-2776(08)60789-0.

Cock, Peter J A *et al.* (2009) 'Biopython: freely available Python tools for computational molecular biology and bioinformatics', *Bioinformatics*. Oxford University Press, 25(11), pp. 1422–1423.

Cock, Peter J.A. *et al.* (2009) 'The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants', *Nucleic Acids Research*, 38(6), pp. 1767–1771. doi: 10.1093/nar/gkp1137.

Conesa, A. *et al.* (2016) 'A survey of best practices for RNA-seq data analysis', *Genome Biology*, 17(1). doi: 10.1186/s13059-016-0881-8.

Crespo, J. *et al.* (2013) 'T cell anergy, exhaustion, senescence, and stemness in the tumor microenvironment', *Current Opinion in Immunology*, 25(2), pp. 214–221. doi: 10.1016/j.coi.2012.12.003.

Cunningham, B. A. (1977) 'The structure and function of histocompatibility antigens.', *Scientific American*, 237(4), pp. 96–107. doi: 10.1038/scientificamerican1077-96.

Curti, B. D. and Faries, M. B. (2021) 'Recent Advances in the Treatment of Melanoma.', *The New England journal of medicine*. United States, 384(23), pp. 2229–2240. doi: 10.1056/NEJMra2034861.

Davis, M. M. *et al.* (1998) 'Ligand recognition by αβ T cell receptors', *Annual Review of Immunology*, 16, pp. 523–544. doi: 10.1146/annurev.immunol.16.1.523.

Dbniak, T. *et al.* (2019) 'Founder mutations for early onset melanoma as revealed by whole exome sequencing suggests that this is not associated with the increasing incidence of melanoma in Poland', *Cancer Research and Treatment*, 51(1), pp. 337–344. doi: 10.4143/crt.2018.157.

Dipiazza, A. *et al.* (2017) 'Avian and human seasonal influenza hemagglutinin proteins elicit CD4 T cell responses that are comparable in epitope abundance and diversity', *Clinical and Vaccine Immunology*, 24(3). doi: 10.1128/CVI.00548-16.

Dobin, A. *et al.* (2013) 'STAR: Ultrafast universal RNA-seq aligner', *Bioinformatics*, 29(1), pp. 15–21. doi: 10.1093/bioinformatics/bts635.

Doherty, P. C. and Zinkernagel, R. M. (1975) 'a Biological Role for the Major Histocompatibility Antigens', *The Lancet*, 305(7922), pp. 1406–1409. doi: 10.1016/S0140-6736(75)92610-0.

Durinck, S. *et al.* (2009) 'Mapping identifiers for the integration of genomic datasets with the R/ Bioconductor package biomaRt', *Nature Protocols*, 4(8), pp. 1184–1191. doi: 10.1038/nprot.2009.97.

Embgenbroich, M. and Burgdorf, S. (2018) 'Current concepts of antigen cross-presentation', *Frontiers in Immunology*, 9(JUL). doi: 10.3389/fimmu.2018.01643.

Fattore, L. *et al.* (2021) 'The Promise of Liquid Biopsy to Predict Response to Immunotherapy in Metastatic Melanoma', *Frontiers in Oncology*, 11. doi: 10.3389/fonc.2021.645069.

Favero, F. *et al.* (2015) 'Sequenza: Allele-specific copy number and mutation profiles from tumor sequencing data', *Annals of Oncology*, 26(1), pp. 64–70. doi: 10.1093/annonc/mdu479.

Feola, S. *et al.* (2020) 'Uncovering the tumor antigen landscape: What to know about the discovery process', *Cancers*, 12(6), pp. 1–28. doi: 10.3390/cancers12061660.

Del Fiore, P. *et al.* (2021) 'Melanoma of Unknown Primary: Evaluation of the Characteristics, Treatment Strategies, Prognostic Factors in a Monocentric Retrospective Study', *Frontiers in Oncology*, 11. doi: 10.3389/fonc.2021.627527.

Frankish, A. *et al.* (2019) 'GENCODE reference annotation for the human and mouse genomes', *Nucleic Acids Research*, 47(D1), pp. D766–D773. doi: 10.1093/nar/gky955.

Gabrielsen, I. S. M. *et al.* (2019) 'Transcriptomes of antigen presenting cells in human thymus', *PLoS ONE*, 14(7). doi: 10.1371/journal.pone.0218858.

Garcia-Diaz, A. *et al.* (2017) 'Interferon Receptor Signaling Pathways Regulating PD-L1 and PD-L2 Expression',

Cell Reports, 19(6), pp. 1189–1201. doi: 10.1016/j.celrep.2017.04.031.

Garstka, M. A. *et al.* (2015) 'The first step of peptide selection in antigen presentation by MHC class I molecules', *Proceedings of the National Academy of Sciences of the United States of America*, 112(5), pp. 1505–1510. doi: 10.1073/pnas.1416543112.

Gary, R. *et al.* (2018) 'Clinical-grade generation of peptide-stimulated CMV/EBV-specific T cells from G-CSF mobilized stem cell grafts', *Journal of Translational Medicine*, 16(1). doi: 10.1186/s12967-018-1498-3.

Gasteiger, G. *et al.* (2017) 'Cellular Innate Immunity: An Old Game with New Players', *Journal of Innate Immunity*, 9(2), pp. 111–125. doi: 10.1159/000453397.

Gjerstorff, M. F., Andersen, M. H. and Ditzel, H. J. (2015) 'Oncogenic cancer/testis antigens: Prime candidates for immunotherapy', *Oncotarget*, 6(18), pp. 15772–15787. doi: 10.18632/oncotarget.4694.

Gonzalez-Galarza, F. F. *et al.* (2020) 'Allele frequency net database (AFND) 2020 update: Gold-standard data classification, open access genotype data and new query tools', *Nucleic Acids Research*, 48(D1), pp. D783–D788. doi: 10.1093/nar/gkz1029.

Gordy, C. and He, Y. W. (2012) 'Endocytosis by target cells: An essential means for perforin-and granzyme-mediated killing', *Cellular and Molecular Immunology*, 9(1), pp. 5–6. doi: 10.1038/cmi.2011.45.

Griewank, K. G. *et al.* (2012) 'Genetic and molecular characterization of uveal melanoma cell lines', *Pigment Cell and Melanoma Research*, 25(2), pp. 182–187. doi: 10.1111/j.1755-148X.2012.00971.x.

Guillaume, P. *et al.* (2018) 'The C-terminal extension landscape of naturally presented HLA-I ligands', *Proceedings of the National Academy of Sciences of the United States of America*, 115(20), pp. 5083–5088. doi: 10.1073/pnas.1717277115.

Guruprasad, K., Reddy, B. V. B. and Pandit, M. W. (1990) 'Correlation between stability of a protein and its dipeptide composition: A novel approach for predicting in vivo stability of a protein from its primary sequence', *Protein Engineering, Design and Selection*, 4(2), pp. 155–161. doi: 10.1093/protein/4.2.155.

Gutiérrez-Martínez, E. *et al.* (2015) 'Cross-presentation of cell-associated antigens by MHC class I in dendritic cell subsets', *Frontiers in Immunology*, 6(JUL). doi: 10.3389/fimmu.2015.00363.

den Haan, J. M. M., Arens, R. and van Zelm, M. C. (2014) 'The activation of the adaptive immune system: Cross-talk between antigen-presenting cells, T cells and B cells', *Immunology Letters*. Elsevier B.V., 162(2), pp. 103–112. doi: 10.1016/j.imlet.2014.10.011.

Hamid, O. *et al.* (2019) 'Five-year survival outcomes for patients with advanced melanoma treated with pembrolizumab in KEYNOTE-001', *Annals of Oncology*, 30(4), pp. 582–588. doi: 10.1093/annonc/mdz011.

Haring, J. S., Corbin, G. A. and Harty, J. T. (2005) 'Dynamic Regulation of IFN-γ Signaling in Antigen-Specific CD8+ T Cells Responding to Infection', *The Journal of Immunology*, 174(11), pp. 6791–6802. doi: 10.4049/jimmunol.174.11.6791.

Harjunpää, H. *et al.* (2019) 'Cell adhesion molecules and their roles and regulation in the immune and tumor

microenvironment', *Frontiers in Immunology*, 10(MAY). doi: 10.3389/fimmu.2019.01078.

Heitmann, J. S. *et al.* (2022) 'A COVID-19 peptide vaccine for the induction of SARS-CoV-2 T cell immunity', *Nature*, 601(7894), pp. 617–622. doi: 10.1038/s41586-021-04232-5.

Henley, S. J. *et al.* (2020) 'Annual report to the nation on the status of cancer, part I: National cancer statistics', *Cancer*, 126(10), pp. 2225–2249. doi: 10.1002/cncr.32802.

Henning, A. N., Klebanoff, C. A. and Restifo, N. P. (2018) 'Silencing stemness in T cell differentiation', *Science*, 359(6372), pp. 163–164. doi: 10.1126/science.aar5541.

Henning, A. N., Roychoudhuri, R. and Restifo, N. P. (2018) 'Epigenetic control of CD8+ T'cell differentiation', *Nature Reviews Immunology*, 18(5), pp. 340–356. doi: 10.1038/nri.2017.146.

Higgins, J. P., Bernstein, M. B. and Hodge, J. W. (2009) 'Enhancing immune responses to tumor-associated antigens', *Cancer Biology and Therapy*, 8(15). doi: 10.4161/cbt.8.15.9133.

Hirayama, M. and Nishimura, Y. (2016) 'The present status and future prospects of peptide-based cancer vaccines', *International Immunology*, 28(7), pp. 319–328. doi: 10.1093/intimm/dxw027.

Howe, K. L. *et al.* (2021) 'Ensembl 2021', *Nucleic Acids Research*, 49(D1), pp. D884–D891. doi: 10.1093/nar/gkaa942.

Huang, L., Kuhls, M. C. and Eisenlohr, L. C. (2011) 'Hydrophobicity as a driver of MHC class i antigen processing', *EMBO Journal*, 30(8), pp. 1634–1644. doi: 10.1038/emboj.2011.62.

Huemer, F. *et al.* (2020) 'Combination strategies for immune-checkpoint blockade and response prediction by artificial intelligence', *International Journal of Molecular Sciences*, 21(8). doi: 10.3390/ijms21082856.

Huppa, J. B. and Davis, M. M. (2003) 'T-cell-antigen recognition and the immunological synapse', *Nature Reviews Immunology*, 3(12), pp. 973–983. doi: 10.1038/nri1245.

Huster, K. M., Stemberger, C. and Busch, D. H. (2006) 'Protective immunity towards intracellular pathogens', *Current Opinion in Immunology*, 18(4), pp. 458–464. doi: 10.1016/j.coi.2006.05.008.

Ikeda, M. *et al.* (2021) 'Phase i studies of peptide vaccine cocktails derived from GPC3, WDRPUH and NEIL3 for advanced hepatocellular carcinoma', *Immunotherapy*, 13(5), pp. 371–385. doi: 10.2217/imt-2020-0278.

Ischenko, I. *et al.* (2021) 'KRAS drives immune evasion in a genetic model of pancreatic cancer', *Nature Communications*, 12(1). doi: 10.1038/s41467-021-21736-w.

Jain, A. and Pasare, C. (2017) 'Innate Control of Adaptive Immunity: Beyond the Three-Signal Paradigm', *The Journal of Immunology*, 198(10), pp. 3791–3800. doi: 10.4049/jimmunol.1602000.

Janeway, C. A. and Medzhitov, R. (2002) 'Innate immune recognition', *Annual Review of Immunology*, 20(2), pp. 197–216. doi: 10.1146/annurev.immunol.20.083001.084359.

Jhunjhunwala, S., Hammer, C. and Delamarre, L. (2021) 'Antigen presentation in cancer: insights into tumour immunogenicity and immune evasion', *Nature Reviews Cancer*, 21(5), pp. 298–312. doi: 10.1038/s41568-021-00339-z.

Joffre, O. P. *et al.* (2012) 'Cross-presentation by dendritic cells', *Nature Reviews Immunology*. Nature Publishing Group, 12(8), pp. 557–569. doi: 10.1038/nri3254.

Jurtz, V. *et al.* (2017) 'NetMHCpan-4.0: Improved Peptide–MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data', *The Journal of Immunology*, 199(9), pp. 3360–3368. doi: 10.4049/jimmunol.1700893.

Kang, K. *et al.* (2020) 'Significance of Tumor Mutation Burden in Immune Infiltration and Prognosis in Cutaneous Melanoma', *Frontiers in Oncology*, 10. doi: 10.3389/fonc.2020.573141.

Kappeler, A. and Mueller, C. (2000) 'The role of activated cytotoxic T cells in inflammatory bowel disease', *Histology and Histopathology*, 15(1), pp. 167–172. doi: 10.14670/HH-15.167.

Kearley, J. *et al.* (2005) 'Resolution of airway inflammation and hyperreactivity after in vivo transfer of CD4+CD25+ regulatory T cells is interleukin 10 dependent', *Journal of Experimental Medicine*, 202(11), pp. 1539–1547. doi: 10.1084/jem.20051166.

Khalil, D. N. *et al.* (2016) 'An open-label, dose-escalation phase I study of anti-TYRP1 monoclonal antibody IMC-20D7S for patients with relapsed or refractory melanoma', *Clinical Cancer Research*, 22(21), pp. 5204–5210. doi: 10.1158/1078-0432.CCR-16-1241.

Kievits, F. *et al.* (1987) 'HLA-restricted recognition of viral antigens in HLA transgenic mice', *Nature*, 329(6138), pp. 447–449. doi: 10.1038/329447a0.

Kim, Y. *et al.* (2014) 'Dataset size and composition impact the reliability of performance benchmarks for peptide-MHC binding predictions', *BMC Bioinformatics*, 15(1). doi: 10.1186/1471-2105-15-241.

Klein, L. *et al.* (2009) 'Antigen presentation in the thymus for positive selection and central tolerance induction', *Nature Reviews Immunology*, 9(12), pp. 833–844. doi: 10.1038/nri2669.

Klein, L. *et al.* (2014) 'Positive and negative selection of the T cell repertoire: What thymocytes see (and don't see)', *Nature Reviews Immunology*, 14(6), pp. 377–391. doi: 10.1038/nri3667.

Kobayashi, T. *et al.* (1999) 'Tyrosinase stabilization by Tyrp1 (the brown locus protein)', *Journal of Biological Chemistry*, 273(48), pp. 31801–31805. doi: 10.1074/jbc.273.48.31801.

Koch, Elias A. T. *et al.* (2022) 'A One-Armed Phase I Dose Escalation Trial Design: Personalized Vaccination with IKKβ-Matured, RNA-Loaded Dendritic Cells for Metastatic Uveal Melanoma', *Frontiers in Immunology*, 13. doi: 10.3389/fimmu.2022.785231.

Koch, Elias A.T. *et al.* (2022) 'Clinical determinants of long-term survival in metastatic uveal melanoma', *Cancer Immunology, Immunotherapy*. Springer Berlin Heidelberg, 71(6), pp. 1467–1477. doi: 10.1007/s00262-021-03090-4.

Koşaloğlu-Yalçın, Z. *et al.* (2021) 'The Cancer Epitope Database and Analysis Resource: A Blueprint for the Establishment of a New Bioinformatics Resource for Use by the Cancer Immunology Community', *Frontiers in Immunology*, 12. doi: 10.3389/fimmu.2021.735609.

Kujala, E., Mäkitie, T. and Kivelä, T. (2003) 'Very Long-Term Prognosis of Patients with Malignant Uveal Melanoma', *Investigative Ophthalmology and Visual Science*, 44(11), pp. 4651–4659. doi: 10.1167/iovs.03-0538.

Kuzu, O. F. *et al.* (2015) 'Current State of Animal (Mouse) Modeling in Melanoma Research', *Cancer Growth and Metastasis*, 8s1, p. CGM.S21214. doi: 10.4137/cgm.s21214.

Kyte, J. and Doolittle, R. F. (1982) 'A simple method for displaying the hydropathic character of a protein', *Journal of Molecular Biology*, 157(1), pp. 105–132. doi: 10.1016/0022-2836(82)90515-0.

Lata, S., Bhasin, M. and Raghava, G. P. (2009) 'MHCBN 4.0: A database of MHC/TAP binding peptides and T-cell epitopes', *BMC Research Notes*, 2. doi: 10.1186/1756-0500-2-61.

Leek, J. T. *et al.* (2010) 'Tackling the widespread and critical impact of batch effects in high-throughput data', *Nature Reviews Genetics*, 11(10), pp. 733–739. doi: 10.1038/nrg2825.

Leung, C. M. *et al.* (2022) 'A guide to the organ-on-a-chip', *Nature Reviews Methods Primers*, 2(1). doi: 10.1038/s43586-022-00118-6.

Levenshtein, V. (1966) 'Binary codes capable of correcting deletions, insertions, and reversals', *Soviet Physics Doklady*, 10(8), pp. 707–710.

Li, H. *et al.* (2009) 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics*, 25(16), pp. 2078–2079. doi: 10.1093/bioinformatics/btp352.

Li, X. C. and Raghavan, M. (2010) 'Structure and function of major histocompatibility complex class i antigens', *Current Opinion in Organ Transplantation*, 15(4), pp. 499–504. doi: 10.1097/MOT.0b013e32833bfb33.

Lin, H. H. *et al.* (2008) 'Evaluation of MHC class I peptide binding prediction servers: Applications for vaccine research', *BMC Immunology*, 9. doi: 10.1186/1471-2172-9-8.

Lischer, C. *et al.* (2019) 'Curatopes melanoma: A Database of Predicted T-cell Epitopes from Overly Expressed Proteins in Metastatic Cutaneous Melanoma', *Cancer Research*, 79(20), pp. 5452–5456. doi: 10.1158/0008-5472.CAN-19-0296.

Liu, S. H. *et al.* (2020) 'DriverDBv3: A multi-omics database for cancer driver gene research', *Nucleic Acids Research*, 48(D1), pp. D863–D870. doi: 10.1093/nar/gkz964.

Liu, Y., Beyer, A. and Aebersold, R. (2016) 'On the Dependency of Cellular Protein Levels on mRNA Abundance', *Cell*, 165(3), pp. 535–550. doi: 10.1016/j.cell.2016.03.014.

Liu, Y., Sun, J. and Zhao, M. (2017) 'ONGene: A literature-based database for human oncogenes', *Journal of Genetics and Genomics*, 44(2), pp. 119–121. doi: 10.1016/j.jgg.2016.12.004.

Lubbers, R. *et al.* (2017) 'Production of complement components by cells of the immune system', *Clinical and Experimental Immunology*, 188(2), pp. 183–194. doi: 10.1111/cei.12952.

Mallet, J. D. *et al.* (2014) 'Implication of ultraviolet light in the etiology of uveal melanoma: A review', *Photochemistry and Photobiology*, 90(1), pp. 15–21. doi: 10.1111/php.12161.

Marshall, J. S. *et al.* (2018) 'An introduction to immunology and immunopathology', *Allergy, Asthma and Clinical Immunology*, 14(Suppl 2). doi: 10.1186/s13223-018-0278-1.

Matys, V. *et al.* (2006) 'TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes.', *Nucleic acids research*, 34(Database issue). doi: 10.1093/nar/gkj143.

Maurer, D. M., Butterfield, L. H. and Vujanovic, L. (2019) 'Melanoma vaccines: Clinical status and immune endpoints', *Melanoma Research*, 29(2), pp. 109–118. doi: 10.1097/CMR.0000000000000535.

McLane, L. M., Abdel-Hakeem, M. S. and Wherry, E. J. (2019) 'CD8 T Cell Exhaustion During Chronic Viral Infection and Cancer', *Annual Review of Immunology*, 37, pp. 457–495. doi: 10.1146/annurev-immunol-041015-055318.

Meyer, D. and Thomson, G. (2001) 'How selection shapes variation of the human major histocompatibility complex: A review', *Annals of Human Genetics*, (1), pp. 1–26. doi: 10.1046/j.1469-1809.2001.6510001.x.

Michalek, M. T. *et al.* (1993) 'A role for the ubiquitin-dependent proteolytic pathway in MHC class l-restricted antigen presentation', *Nature*, 363(6429), pp. 552–554. doi: 10.1038/363552a0.

Middleton, D. *et al.* (2003) 'New allele frequency database: www.allelefrequencies.net', *Tissue Antigens*, 61(5), pp. 403–407. doi: 10.1034/j.1399-0039.2003.00062.x.

Morgan, R. A. *et al.* (2013) 'Cancer regression and neurological toxicity following anti-MAGE-A3 TCR gene therapy', *Journal of Immunotherapy*, 36(2), pp. 133–151. doi: 10.1097/CJI.0b013e3182829903.

Mortazavi, A. *et al.* (2008) 'Mapping and quantifying mammalian transcriptomes by RNA-Seq', *Nature Methods*, 5(7), pp. 621–628. doi: 10.1038/nmeth.1226.

Nair-Gupta, P. *et al.* (2014) 'TLR signals induce phagosomal MHC-I delivery from the endosomal recycling compartment to allow cross-presentation', *Cell*, 158(3), pp. 506–521. doi: 10.1016/j.cell.2014.04.054.

Nathan, P. *et al.* (2021) 'Overall Survival Benefit with Tebentafusp in Metastatic Uveal Melanoma', *New England Journal of Medicine*, 385(13), pp. 1196–1206. doi: 10.1056/nejmoa2103485.

Ness, C. *et al.* (2021) 'Integrated differential DNA methylation and gene expression of formalin-fixed paraffin-embedded uveal melanoma specimens identifies genes associated with early metastasis and poor prognosis', *Experimental Eye Research*, 203. doi: 10.1016/j.exer.2020.108426.

Netea, M. G. *et al.* (2020) 'Defining trained immunity and its role in health and disease', *Nature Reviews Immunology*. Springer US, 20(6), pp. 375–388. doi: 10.1038/s41577-020-0285-6.

Novotny, J. *et al.* (1986) 'Antigenic determinants in proteins coincide with surface regions accessible to large probes (antibody domains)', *Proceedings of the National Academy of Sciences of the United States of America*, 83(2), pp. 226–230. doi: 10.1073/pnas.83.2.226.

O'Donnell, T. J., Rubinsteyn, A. and Laserson, U. (2020) 'MHCflurry 2.0: Improved Pan-Allele Prediction of MHC Class I-Presented Peptides by Incorporating Antigen Processing', *Cell Systems*. Elsevier, 11(1), pp. 42-48.e7. doi: 10.1016/j.cels.2020.06.010.

Oiseth, S. J. and Aziz, M. S. (2017) 'Cancer immunotherapy: a brief review of the history, possibilities, and challenges ahead', *Journal of Cancer Metastasis and Treatment*, 3(10), p. 250. doi: 10.20517/2394-4722.2017.41.

Olsen, L. R. *et al.* (2017) 'TANTIGEN: a comprehensive database of tumor T cell antigens', *Cancer Immunology, Immunotherapy*, 66(6), pp. 731–735. doi: 10.1007/s00262-017-1978-y.

Ooki, A., Shinozaki, E. and Yamaguchi, K. (2021) 'Immunotherapy in Colorectal Cancer: Current and Future Strategies', *Journal of the Anus, Rectum and Colon*, 5(1), pp. 11–24. doi: 10.23922/jarc.2020-064.

Ovchinnikov, D. A. (2008) 'Macrophages in the embryo and beyond: Much more than just giant phagocytes', *Genesis*, 46(9), pp. 447–462. doi: 10.1002/dvg.20417.

Pai, J. A. and Satpathy, A. T. (2021) 'High-throughput and single-cell T cell receptor sequencing technologies', *Nature Methods*, 18(8), pp. 881–892. doi: 10.1038/s41592-021-01201-8.

Palmowski, M. J. *et al.* (2006) 'Role of Immunoproteasomes in Cross-Presentation', *The Journal of Immunology*, 177(2), pp. 983–990. doi: 10.4049/jimmunol.177.2.983.

Parlar, A. *et al.* (2019) 'Engineering antigen-specific NK cell lines against the melanoma-associated antigen tyrosinase via TCR gene transfer', *European Journal of Immunology*, 49(8), pp. 1278–1290. doi: 10.1002/eji.201948140.

Pauken, K. E. *et al.* (2016) 'Epigenetic stability of exhausted T cells limits durability of reinvigoration by PD-1 blockade', *Science*, 354(6316), pp. 1160–1165. doi: 10.1126/science.aaf2807.

Paul, S. *et al.* (2013) 'HLA Class I Alleles Are Associated with Peptide-Binding Repertoires of Different Size, Affinity, and Immunogenicity', *The Journal of Immunology*, 191(12), pp. 5831–5839. doi: 10.4049/jimmunol.1302101.

Pertea, M. *et al.* (2015) 'StringTie enables improved reconstruction of a transcriptome from RNA-seq reads', *Nature Biotechnology*, 33(3), pp. 290–295. doi: 10.1038/nbt.3122.

Peterson, P., Org, T. and Rebane, A. (2008) 'Transcriptional regulation by AIRE: Molecular mechanisms of central tolerance', *Nature Reviews Immunology*, 8(12), pp. 948–957. doi: 10.1038/nri2450.

Prachar, M. *et al.* (2020) 'Identification and validation of 174 COVID-19 vaccine candidate epitopes reveals low performance of common epitope prediction tools', *Scientific Reports*, 10(1). doi: 10.1038/s41598-020-77466-4.

Pruitt, K. D. *et al.* (2009) 'The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes', *Genome Research*, 19(7), pp. 1316–1323. doi: 10.1101/gr.080531.108.

Pugliese, A. (2017) 'Autoreactive T cells in type 1 diabetes', *Journal of Clinical Investigation*, 127(8), pp. 2881–2891. doi: 10.1172/JCI94549.

Rahman, A. *et al.* (2018) 'Importance of Feedback and Feedforward Loops to Adaptive Immune Response

Modeling', *CPT: Pharmacometrics and Systems Pharmacology*, 7(10), pp. 621–628. doi: 10.1002/psp4.12352.

Rähni, A. *et al.* (2022) 'Melanoma-specific antigen-associated antitumor antibody reactivity as an immune-related biomarker for targeted immunotherapies', *Communications Medicine*, 2(1). doi: 10.1038/s43856-022-00114-7.

Rangarajan, S. and Mariuzza, R. A. (2014) 'T cell receptor bias for MHC: Co-evolution or co-receptors?', *Cellular and Molecular Life Sciences*, 71(16), pp. 3059–3068. doi: 10.1007/s00018-014-1600-9.

Raskov, H. *et al.* (2021) 'Cytotoxic CD8+ T cells in cancer and cancer immunotherapy', *British Journal of Cancer*. Springer US, 124(2), pp. 359–367. doi: 10.1038/s41416-020-01048-4.

Reboul, C. F. *et al.* (2012) 'Epitope flexibility and dynamic footprint revealed by molecular dynamics of a pMHC-TCR complex', *PLoS Computational Biology*, 8(3). doi: 10.1371/journal.pcbi.1002404.

Restifo, N. P. and Gattinoni, L. (2013) 'Lineage relationship of effector and memory T cells', *Current Opinion in Immunology*, 25(5), pp. 556–563. doi: 10.1016/j.coi.2013.09.003.

Reynisson, B. *et al.* (2021) 'NetMHCpan-4.1 and NetMHCIIpan-4.0: Improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data', *Nucleic Acids Research*. Oxford University Press, 48(W1), pp. W449–W454. doi: 10.1093/NAR/GKAA379.

Richards, J. R. *et al.* (2020) 'Mouse models of uveal melanoma: Strengths, weaknesses, and future directions', *Pigment Cell and Melanoma Research*, 33(2), pp. 264–278. doi: 10.1111/pcmr.12853.

Robert, C. *et al.* (2015) 'Pembrolizumab versus Ipilimumab in Advanced Melanoma', *New England Journal of Medicine*, 372(26), pp. 2521–2532. doi: 10.1056/nejmoa1503093.

Robertson, A. G. *et al.* (2017) 'Integrative Analysis Identifies Four Molecular and Clinical Subsets in Uveal Melanoma', *Cancer Cell*, 32(2), pp. 204-220.e15. doi: 10.1016/j.ccell.2017.07.003.

Robinson, J. *et al.* (2020) *IPD-IMGT/HLA Database*, *Nucleic Acids Research*. doi: 10.1093/nar/gkz950.

Rosenberg, S. A. *et al.* (1994) 'Treatment of patients with metastatic melanoma with autologous tumor-infiltrating lymphocytes and interleukin 2', *Journal of the National Cancer Institute*, 86(15), pp. 1159–1166. doi: 10.1093/jnci/86.15.1159.

Ross, S. H. and Cantrell, D. A. (2018) 'Signaling and Function of Interleukin-2 in T Lymphocytes', *Annual Review of Immunology*, 36, pp. 411–433. doi: 10.1146/annurev-immunol-042617-053352.

Roth, D. B. (2015) 'V(D)J recombination: Mechanism, errors, and fidelity', *Mobile DNA III*, 2(6), pp. 313–324. doi: 10.1128/9781555819217.ch14.

RStudio (2011) 'RStudio: Integrated development environment for R (Version 0.97.311)', *The Journal of Wildlife Management*. Boston, MA, pp. 1753–1766. Available at: http://doi.wiley.com/10.1002/jwmg.232.

Sadozai, H. *et al.* (2017) 'Recent successes and future directions in immunotherapy of cutaneous melanoma', *Frontiers in Immunology*, 8(DEC). doi: 10.3389/fimmu.2017.01617.

Sajid, Z. *et al.* (2021) 'Genetic causes of oculocutaneous albinism in pakistani population', *Genes*, 12(4). doi:

10.3390/genes12040492.

Sakamoto, H. *et al.* (2008) 'Mechanisms of Cables 1 gene inactivation in human ovarian cancer development', *Cancer Biology and Therapy*, 7(2), pp. 180–188. doi: 10.4161/cbt.7.2.5253.

Sanchez-Mazas, A. *et al.* (2017) 'Common and well-documented HLA alleles over all of Europe and within European sub-regions: A catalogue from the European Federation for Immunogenetics', *Hla*, 89(2), pp. 104–113. doi: 10.1111/tan.12956.

Sarkizova, S. *et al.* (2020) 'A large peptidome dataset improves HLA class I epitope prediction across most of the human population', *Nature Biotechnology*, 38(2), pp. 199–209. doi: 10.1038/s41587-019-0322-9.

Savina, A. and Amigorena, S. (2007) 'Phagocytosis and antigen presentation in dendritic cells', *Immunological Reviews*, 219(1), pp. 143–156. doi: 10.1111/j.1600-065X.2007.00552.x.

Schadendorf, D. *et al.* (2015) 'Pooled analysis of long-term survival data from phase II and phase III trials of ipilimumab in unresectable or metastatic melanoma', *Journal of Clinical Oncology*, 33(17), pp. 1889–1894. doi: 10.1200/JCO.2014.56.2736.

Schneider, V. A. *et al.* (2017) 'Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly', *Genome Research*, 27(5), pp. 849–864. doi: 10.1101/gr.213611.116.

Schoenborn, J. R. and Wilson, C. B. (2007) 'Regulation of Interferon-γ During Innate and Adaptive Immune Responses', *Advances in Immunology*, 96, pp. 41–101. doi: 10.1016/S0065-2776(07)96002-2.

Schuler-Thurner, B. *et al.* (2015) 'Immuntherapie beim Aderhautmelanom: Vakzination gegen Krebs: Multizentrische adjuvante Phase-III-Impfstudie mit Tumor-RNA-beladenen dendritischen Zellen bei neu diagnostizierten, großen Uveamelanomen', *Ophthalmologe*, 112(12), pp. 1017–1021. doi: 10.1007/s00347-015-0162-z.

Seifert, M. and Küppers, R. (2016) 'Human memory B cells', *Leukemia*, 30(12), pp. 2283–2292. doi: 10.1038/leu.2016.226.

Selck, C. and Dominguez-Villar, M. (2021) 'Antigen-Specific Regulatory T Cell Therapy in Autoimmune Diseases and Transplantation', *Frontiers in Immunology*, 12(May), pp. 1–12. doi: 10.3389/fimmu.2021.661875.

Seong, I. *et al.* (2012) 'Sox10 controls migration of B16F10 melanoma cells through multiple regulatory target genes', *PLoS ONE*, 7(2). doi: 10.1371/journal.pone.0031477.

Sewell, A. K. (2012) 'Why must T cells be cross-reactive?', *Nature Reviews Immunology*, 12(9), pp. 669–677. doi: 10.1038/nri3279.

Shen, C. *et al.* (2017) 'Frequency and reactivity of antigen-specific T cells were concurrently measured through the combination of artificial antigen-presenting cell, MACS and ELISPOT', *Scientific Reports*, 7(1). doi: 10.1038/s41598-017-16549-1.

SMahajan, W. *et al.* (2018) 'Epitope specific antibodies and T cell receptors in the immune epitope database',

Frontiers in Immunology, 9(NOV). doi: 10.3389/fimmu.2018.02688.

Smith, M. R. (2003) 'Rituximab (monoclonal anti-CD20 antibody): Mechanisms of action and resistance', *Oncogene*, 22(47), pp. 7359–7368. doi: 10.1038/sj.onc.1206939.

Solberg, O. D. *et al.* (2008) 'Balancing selection and heterogeneity across the classical human leukocyte antigen loci: A meta-analytic review of 497 population studies', *Human Immunology*, 69(7), pp. 443–464. doi: 10.1016/j.humimm.2008.05.001.

Stark, R., Grzelak, M. and Hadfield, J. (2019) 'RNA sequencing: the teenage years', *Nature Reviews Genetics*. doi: 10.1038/s41576-019-0150-2.

Steele, J. C. *et al.* (2011) 'Phase I/II trial of a dendritic cell vaccine transfected with DNA encoding melan A and gp100 for patients with metastatic melanoma', *Gene Therapy*, 18(6), pp. 584–593. doi: 10.1038/gt.2011.1.

Stenger, S. *et al.* (1998) 'An antimicrobial activity of cytolytic T cells mediated by granulysin', *Science*, 282(5386), pp. 121–125. doi: 10.1126/science.282.5386.121.

Stranzl, T. *et al.* (2010) 'NetCTLpan: Pan-specific MHC class I pathway epitope predictions', *Immunogenetics*, 62(6), pp. 357–368. doi: 10.1007/s00251-010-0441-4.

Sugita, S. *et al.* (2007) 'Cross-reaction between tyrosinase peptides and cytomegalovirus antigen by T cells from patients with Vogt-Koyanagi-Harada disease', *International Ophthalmology*, 27(2–3), pp. 87–95. doi: 10.1007/s10792-006-9020-y.

Syn, N. L. *et al.* (2017) 'De-novo and acquired resistance to immune checkpoint targeting', *The Lancet Oncology*, 18(12), pp. e731–e741. doi: 10.1016/S1470-2045(17)30607-1.

Szeto, C. *et al.* (2021) 'TCR recognition of peptide–MHC-I: Rule makers and breakers', *International Journal of Molecular Sciences*, 22(1), pp. 1–26. doi: 10.3390/ijms22010068.

Tabana, Y. *et al.* (2021) 'Reversing T-cell exhaustion in immunotherapy: a review on current approaches and limitations', *Expert Opinion on Therapeutic Targets*, 25(5), pp. 347–363. doi: 10.1080/14728222.2021.1937123.

Takaba, H. and Takayanagi, H. (2017) 'The Mechanisms of T Cell Selection in the Thymus', *Trends in Immunology*, 38(11), pp. 805–816. doi: 10.1016/j.it.2017.07.010.

Tamoutounour, S. *et al.* (2013) 'Origins and functional specialization of macrophages and of conventional and monocyte-derived dendritic cells in mouse skin', *Immunity*, 39(5), pp. 925–938. doi: 10.1016/j.immuni.2013.10.004.

Taylor, J. J., Jenkins, M. K. and Pape, K. A. (2012) 'Heterogeneity in the differentiation and function of memory B cells', *Trends in Immunology*, 33(12), pp. 590–597. doi: 10.1016/j.it.2012.07.005.

Ting, J. P. Y. and Trowsdale, J. (2002) 'Genetic control of MHC class II expression', *Cell*, 109(2 SUPPL. 1). doi: 10.1016/S0092-8674(02)00696-7.

Tofallis, C. (2014) 'Add or Multiply? A Tutorial on Ranking and Choosing with Multiple Criteria', *INFORMS Transactions on Education*, 14(3), pp. 109–119. doi: 10.1287/ited.2013.0124.

Tomiyama, H., Matsuda, T. and Takiguchi, M. (2002) ' Differentiation of Human CD8 + T Cells from a Memory to Memory/Effector Phenotype ', *The Journal of Immunology*, 168(11), pp. 5538–5550. doi: 10.4049/jimmunol.168.11.5538.

Tran, E. *et al.* (2013) 'Immune targeting of fibroblast activation protein triggers recognition of multipotent bone marrow stromal cells and cachexia', *Journal of Experimental Medicine*, 210(6), pp. 1065–1068. doi: 10.1084/jem.20130110.

Trapani, J. A. (1995) 'Target cell apoptosis induced by cytotoxic T cells and natural killer cells involves synergy between the pore-forming protein, perforin, and the serine protease, granzyme B', *Australian and New Zealand Journal of Medicine*, 25(6), pp. 793–799. doi: 10.1111/j.1445-5994.1995.tb02883.x.

Trapnell, C. *et al.* (2012) 'Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks', *Nature Protocols*, 7(3), pp. 562–578. doi: 10.1038/nprot.2012.016.

Trolle, T. *et al.* (2016) 'The Length Distribution of Class I–Restricted T Cell Epitopes Is Determined by Both Peptide Supply and MHC Allele–Specific Binding Preference', *The Journal of Immunology*, 196(4), pp. 1480–1487. doi: 10.4049/jimmunol.1501721.

Uhlén, M. *et al.* (2015) 'Tissue-based map of the human proteome', *Science*, 347(6220). doi: 10.1126/science.1260419.

Venkatesh, G. *et al.* (2020) 'MHCAttnNet: predicting MHC-peptide bindings for MHC alleles classes I and II using an attention-based deep neural model', *Bioinformatics*, 36(Supplement_1), pp. i399–i406. doi: 10.1093/bioinformatics/btaa479.

Vigneron, N. *et al.* (2013) 'Database of T cell-defined human tumor antigens: The 2013 update', *Cancer Immunity*, 13(3), pp. 1–6.

Vono, M. *et al.* (2017) 'Neutrophils acquire the capacity for antigen presentation to memory CD4+ T cells in vitro and ex vivo', *Blood*, 129(14), pp. 1991–2001. doi: 10.1182/blood-2016-10-744441.

De Waard-Siebinga, I. *et al.* (1995) 'Establishment and characterization of an uveal-melanoma cell line', *International Journal of Cancer*, 62(2), pp. 155–161. doi: 10.1002/ijc.2910620208.

Waldman, A. D., Fritz, J. M. and Lenardo, M. J. (2020) 'A guide to cancer immunotherapy: from T cell basic science to clinical practice', *Nature Reviews Immunology*. Springer US, 20(11), pp. 651–668. doi: 10.1038/s41577-020-0306-5.

Wearsch, P. A. and Cresswell, P. (2007) 'Selective loading of high-affinity peptides onto major histocompatibility complex class I molecules by the tapasin-ERp57 heterodimer', *Nature Immunology*, 8(8), pp. 873–881. doi: 10.1038/ni1485.

Weis, E. *et al.* (2016) 'Management of uveal melanoma: A consensus-based provincial clinical practice

guideline', *Current Oncology*, 23(1), pp. e57–e64. doi: 10.3747/co.23.2859.

Wessely, A. *et al.* (2020) 'The role of immune checkpoint blockade in uveal melanoma', *International Journal of Molecular Sciences*, 21(3). doi: 10.3390/ijms21030879.

Wilgenhof, S. *et al.* (2011) 'Therapeutic vaccination with an autologous mRNA electroporated dendritic cell vaccine in patients with advanced melanoma', *Journal of Immunotherapy*, 34(5), pp. 448–456. doi: 10.1097/CJI.0b013e31821dcb31.

Wolchok, J. D. *et al.* (2007) 'Safety and immunogenicity of tyrosinase DNA vaccines in patients with melanoma', *Molecular Therapy*, 15(11), pp. 2044–2050. doi: 10.1038/sj.mt.6300290.

Wolchok, J. D. *et al.* (2017) 'Overall Survival with Combined Nivolumab and Ipilimumab in Advanced Melanoma', *New England Journal of Medicine*, 377(14), pp. 1345–1356. doi: 10.1056/nejmoa1709684.

Wooldridge, L. *et al.* (2012) 'A single autoimmune T cell receptor recognizes more than a million different peptides', *Journal of Biological Chemistry*, 287(2), pp. 1168–1177. doi: 10.1074/jbc.M111.289488.

Wynn, T. A., Chawla, A. and Pollard, J. W. (2013) 'Macrophage biology in development, homeostasis and disease', *Nature*, 496(7446), pp. 445–455. doi: 10.1038/nature12034.

Xiao, F. *et al.* (2009) 'miRecords: An integrated resource for microRNA-target interactions', *Nucleic Acids Research*, 37(SUPPL. 1). doi: 10.1093/nar/gkn851.

Xie, J., Tato, C. M. and Davis, M. M. (2013) 'How the immune system talks to itself: The varied role of synapses', *Immunological Reviews*, 251(1), pp. 65–79. doi: 10.1111/imr.12017.

Xing, Y. and Hogquist, K. A. (2012) 'T-Cell tolerance: Central and peripheral', *Cold Spring Harbor Perspectives in Biology*, 4(6), pp. 1–15. doi: 10.1101/cshperspect.a006957.

Xu, J. *et al.* (2020) 'Ocular cytomegalovirus latency exacerbates the development of choroidal neovascularization', *Journal of Pathology*, 251(2), pp. 200–212. doi: 10.1002/path.5447.

Yao, C. *et al.* (2019) 'Single-cell RNA-seq reveals TOX as a key regulator of CD8+ T cell persistence in chronic infection', *Nature Immunology*, 20(7), pp. 890–901. doi: 10.1038/s41590-019-0403-4.

Yewdell, J. W., Reits, E. and Neefjes, J. (2003) 'Making sense of mass destruction: Quantitating MHC class I antigen presentation', *Nature Reviews Immunology*, 3(12), pp. 952–961. doi: 10.1038/nri1250.

Youngblood, B., Hale, J. S. and Ahmed, R. (2013) 'T-cell memory differentiation: Insights from transcriptional signatures and epigenetics', *Immunology*, 139(3), pp. 277–284. doi: 10.1111/imm.12074.

Yuan, S. *et al.* (2014) 'Comparative immune systems in animals', *Annual Review of Animal Biosciences*, 2(November), pp. 235–258. doi: 10.1146/annurev-animal-031412-103634.

Yuseff, M. I. *et al.* (2013) 'How B cells capture, process and present antigens: A crucial role for cell polarity', *Nature Reviews Immunology*, 13(7), pp. 475–486. doi: 10.1038/nri3469.

Zacharakis, N. *et al.* (2018) 'Immune recognition of somatic mutations leading to complete durable regression in metastatic breast cancer', *Nature Medicine*, 24(6), pp. 724–730. doi: 10.1038/s41591-018-0040-8.

Zajonc, D. M. (2020) ‘Unconventional peptide presentation by classical mhc class i and implications for t and nk cell activation’, *International Journal of Molecular Sciences*, 21(20), pp. 1–13. doi: 10.3390/ijms21207561.

Zhang, S. *et al.* (2021) ‘PMEL as a Prognostic Biomarker and Negatively Associated with Immune Infiltration in Skin Cutaneous Melanoma (SKCM)’, *Journal of Immunotherapy*, 44(6), pp. 214–223. doi: 10.1097/CJI.0000000000000374.

Zhao, W. and Sher, X. (2018) ‘Systematically benchmarking peptide-MHC binding predictors: From synthetic to naturally processed epitopes’, *PLoS Computational Biology*, 14(11). doi: 10.1371/journal.pcbi.1006457.

Zimmerman, J. M., Eliezer, N. and Simha, R. (1968) ‘The characterization of amino acid sequences in proteins by statistical methods’, *Journal of Theoretical Biology*, 21(2), pp. 170–201. doi: 10.1016/0022-5193(68)90069-6.

Zumerle, S., Molon, B. and Viola, A. (2017) ‘Membrane rafts in T cell activation: A spotlight on CD28 costimulation’, *Frontiers in Immunology*, 8(NOV). doi: 10.3389/fimmu.2017.01467.

8 Indices

8.1 List of Figures

8.2 List of Tables

8.3 Glossary & Abbreviations

(m)RNA	(Messenger) Ribonucleic acid
AA	Amino acid
AB	Antibody
AP	Alternative predictor
APC	Antigen-presenting cell
BAM	Binary alignment map
CAR	Chimeric antigen receptor
CD3/4/8/19/20	Cluster of differentiation 3/4/8/19/20 (protein labels)
cDNA	Complementary DNA
CMV	Cytomegalovirus
CTA	Cancer/testis antigen
CTL	Cytotoxic T Lymphocyte
DC	Dendritic cell
DNA	Deoxyribonucleic Acid
Epitope	The combination between peptide and bound HLA allele
ES	Efficacy score
FASTA	Text-based format for representing sequences (nucleotide or amino acid)
FASTQ	Text-based form for representing and storing sequence and Q quality scores
FPKM	Fragments per kilobase per million (shorthand)
GI	Gene importance
gPIE	generalized epitope Predicted Immuno-Efficacy
GTEx	Genotype-Tissue Expression Portal
HE	High efficacy
HLA	Human leukocyte antigen
HPA	Human proteome atlas
Idspx	Indispensability index
IFN	Interferon
IL	Interleukin

LE	Low efficacy
LN	Lymph node
MCM	Metastatic cutaneous melanoma
MHC	Major histocompatibility complex
МФ	Macrophage
ND	Node degree
NGS	Next-generation sequencing
NI	Neighborhood importance
PBMC	Peripheral blood mononuclear cell
PRR	Pattern recognition receptor
Q	Phred score
TAA	Tumor-associated antigen
TCR	T cell receptor
TGCA	The Cancer Genome Atlas
TPM	Transcripts per million
UM	Uveal melanoma

8.4 Publications

1. **Lischer, C*.,** Eberhardt, M.*, Jaitly, T.*, Schinzel, C., Schaft, N., Dorrie, J., Schuler, G., Vera, J., 2019.Curatopes melanoma: A Database of Predicted T-cell Epitopes from Overly Expressed Proteins in Metastatic Cutaneous Melanoma. **Cancer Res**. 79, 5452–5456. *First authors contributed equally
2. Sauerer, T.*, **Lischer, C**.*, Weich, A., Berking, C., Vera, J., Dörrie, J., 2021. Single-Molecule RNA Sequencing Reveals IFNγ-Induced Differential Expression of Immune Escape Genes in Merkel Cell Polyomavirus–Positive MCC Cell Lines. **Front. Microbiol.** 12. *First authors contributed equally
3. **Lischer, C**., Vera-González, J., 2021. The Road to Effective Cancer Immunotherapy—A Computational Perspective on Tumor Epitopes in Anti-Cancer Immunotherapy. Syst. Med. 593
4. Vera, J., **Lischer, C.,** Nenov, M., Nikolov, S., Lai, X., Eberhardt, M., 2021. Mathematical modelling in biomedicine: A primer for the curious and the skeptic. Int. J. Mol. Sci. 22, 1–16.
5. Lai, X., Dreyer, F.S., Cantone, M., Eberhardt, M., Gerer, K.F., Jaitly, T., Uebe, S., **Lischer, C**., Ekici, A., Wittmann, J., Jäck, H.M., Schaft, N., Dörrie, J., Vera, J., 2021. Network- And systems-based re-engineering of dendritic cells with non-coding RNAs for cancer immunotherapy. **Theranostics** 11, 1412–1428.
6. Mougiakakos, D., Bach, C., Böttcher, M., Beier, F., Röhner, L., Stoll, A., Rehli, M., Gebhard, C., **Lischer, C.**, Eberhardt, M., Vera, J., Buttner-Heröld, M., Bitterer, K., Balzer, H., Leffler, M., Jitschin, S., Hundemer, M., Awwad, M.H.S., Busch, M., Stenger, S., Völkl, S., Schütz, C., Krönke, J., Mackensen, A., Bruns, H., 2021. The IKZF1–IRF4/IRF5 axis controls polarization of myeloma-associated macrophages. **Cancer Immunol.** Res. 9, 265–278.
7. Vera J., Lai X., Baur A., Erdmann M., Gupta S., Guttà C., Heinzerling L., Heppt M.V., Kazmierczak P.M., Kunz M., **Lischer C.**, Pützer B.M., Rehm M., Ostalecki C., Retzlaff J., Witt S., Wolkenhauer O., Berking C. Melanoma 2.0. Skin cancer as a paradigm for emerging diagnostic technologies, computational modelling and artificial intelligence**. Brief Bioinform.** 2022 Nov 19;23(6)
8. Karl F., Liang C., Böttcher-Loschinski R., Stoll A., Flamann C., Richter S., **Lischer C.**, Völkl S., Jacobs B., Böttcher M., Jitschin R., Bruns H., Fischer T., Holler E., Rösler W., Dandekar T., Mackensen A., Mougiakakos D. Oxidative DNA damage in reconstituting T cells is associated with relapse and inferior survival after allo-SCT. **Blood**. 2023 Mar 30;141(13):1626-1639.