APPLICATION OF MACHINE LEARNING APPROACHES TO EMPOWER DRUG DEVELOPMENT

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Yue Shen

May 2023

Abstract

Human health, one of the major topics in Life Science, is facing intensified challenges, including cancer, pandemic outbreaks, and antimicrobial resistance. Thus, new medicines with unique advantages, including peptide-based vaccines and permeable small molecule antimicrobials, are in urgent need. However, the drug development process is long, complex, and risky with no guarantee of success. Also, the improvements in techniques applied in genomics, proteomics, computational biology, and clinical trials significantly increase the data complexity and volume, which imposes higher requirements on the drug development pipeline. In recent years, machine learning (ML) methods were employed to support drug development in various aspects and were shown to be highly effective. Here, we explored the application of advanced ML approaches to empower the development of peptide-based vaccines and permeable antimicrobials. First, the peptide-based vaccines targeting pancreatic cancer and COVID-19 were predicted and screened via multiple approaches. Next, novel structure-based methods to improve the performance of peptide: MHC binding affinity prediction were developed, including an HLA modeling pipeline that provides structures for docking-based peptide binder validation, and hierarchical clustering of HLA I into supertypes and subtypes that have similar peptide binding specificity. Finally, the physicochemical properties governing the permeability of small molecules into multidrug-resistant *Pseudomonas aeruginosa* cells were selected using a random forest model. In conclusion, the use of machine learning methods could accelerate the drug development process at a lower cost and promote data-based decision-making if used properly.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction and General Information

1.1 Introduction

Improving human health is one of the major contributions of Life Science innovations. However, scientists are facing new threats, including, but not limited to, cancer, pandemic outbreak, and antimicrobial resistance. Cancer surveillance studies have shown that there were 23.6 million new global cancer cases and 10.0 million cancer deaths in 2019, the rate of diagnosis and death have raised 26.3% and 20.9%, respectively, compared to 2010 [1, 2]. The recent coronavirus disease-2019 (COVID-19) pandemic, caused by Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), outbroke in late 2019 and propagated globally, led to over 611 million confirmed cases, 6.5 million deaths, and huge impact on economy [3]. Antimicrobial resistance has caused at least 1.27 million deaths worldwide and associated with nearly 5 million deaths in 2019, constituting an urgent threat to public health. As existing medicines start to show weakness in fighting against the rapidly increasing threat, new therapies are needed to address these challenges. In order to achieve such goal, efforts should be made in two aspects, discovery of new medicines as well as upgrade in drug development pipeline.

Two new medicines, peptide-based vaccines and permeable antimicrobials were demonstrated to have potential in resolving these challenges. The peptide-based vaccines elicit adaptive immune response using short peptides that bind to major histocompatibility complex (MHC) molecules. Compared to other vaccine strategies, the peptide-based vaccine prevails in safety, development swiftness, and production easiness [4], hence is suitable for personalized cancer vaccines and rapid vaccines against pandemic [5, 6]. As for the multidrug resistance, most antimicrobials were hindered by the outer membrane (OM) and active efflux pumps of Gram-negative bacteria [7, 8]. Thus, efforts have been made to select or modify drug molecules that could penetrate the OM and evade efflux activity as a straightforward solution [9].

Conventional drug development strategy was successfully employed in the past and produced numerous drugs. However, it requires massive time and effort, on average takes more than 12 years and 1.8 billion US dollars to develop one new drug. During the preclinical and clinical trials, 96% of drug candidates failed by showing low efficacy [10]. As a result, in response to the urgent need of new drugs, the drug discovery projects have to be carried out in even larger scale, resulting in higher demand in time and effort. The advancements of in silico approaches provide viable alternatives to conventional experiments, which significantly accelerating the development of new

drugs and therapies [11]. For example, molecular dynamics (MD) and docking methods were applied to virtually screen large chemical libraries for candidates that bind to target molecules. The increased drug screening scale and wide application of in silico methods in combination produce huge amount of data available for detailed analysis but also boost the demand to data mining strategies, which facilitates the rapid development of machine learning (ML) methods [12]. In modern drug development processes, ML methods are vigorously applied to improve the accuracy and throughput of the pipeline, which has been demonstrated to be highly effective. For example, the properties of drug candidates, including, but not limited to, toxicity, efficacy, and permeability, could be predicted using QSAR methods to refine the drug screening result. Also, even if the experimentally determined structure of target molecule is not available, the advanced structural modeling methods could accurately predict it for virtual drug screening. It is foreseeable that ML methods will take more important roles in drug development.

1.2 Major Histocompatibility Complex

Histocompatibility Complex (MHC) is a vital component in adaptive immune system by binding and presenting peptide antigens towards T lymphocytes [13-15]. Studies show that the malfunction of MHC molecules are related to infectious diseases [16], cancer [17], and autoimmunity [18]. A better understanding of the structure and function of the MHC molecules could further facilitate the development of therapies targeting MHC.

1.2.1 MHC genes and structures

There are two major classes of classical MHC, class I and II. Class I molecules are composed of two imbalanced chains: the α chain is larger in size and functionally more important, while the β chain (β macroglobulin, B2M), encoded by non-MHC gene, is almost invariant. Class II molecules also contain α and β chains, while they are balanced in size and functional importance [19].

Molecules of both classes have similar folding (Fig. 1-1A, Fig. 1-2A). From outreached to intracellular, MHC molecules could be divided into several domains, including peptide binding domain, surface proximal domain, transmembrane domain (TM), and cytoplasmic tails. In MHC I molecule, the peptide binding domain is formed by α1 and α2 subunits, the membrane proximal domain is formed by α3 subunit and β chain, while only α chain processes TM domain and cytoplasmic tail. The MHC II molecules are roughly symmetric, the α1 and β1 subunits form the

peptide binding domain, the α2 and β2 subunits form the membrane proximal domain, and both chains have TM domains as well as cytoplasmic tails [20].

The human version of MHC is also known as Human Leukocyte Antigen (HLA). HLA genes are located on the short arm of chromosome 6, spanning approximately 3.6 mega base pairs. Classical class I regions consist of A, B, and C loci that encode the α chains of HLA molecules, while classical class II regions consist of DP, DQ, and DR genes, each containing A and B loci encoding α and β chains, respectively [13].

1.2.2 Function of MHC in adaptive immune system

The MHC is first known as transplantation antigen [21]. Its major biological function is regulation of immune response, including, but not limit to, antigen presentation and T cell activation. MHC class I molecules appears on the surface of almost all nucleated cells and present endogenous peptides towards CD8+ T cells (cytotoxic T cells) (Fig. 1-3A). Class II molecules are mainly expressed by antigen presenting cells (APCs), such as dendritic cells, macrophages, Langerhans cells and B cells, and present exogenous peptides towards CD4+ Tells (helper T cells) (Fig 1-3B) [13].

The peptide-MHC complex (pMHC) elicits activation of T cells carrying certain T cell receptors (TCR) that can recognize such pMHC. The phenomenon that T cells can only respond to antigens that presented by self MHC molecules is referred to as MHC-restricted antigen recognition, or MHC restriction [19].

1.2.3 Peptide MHC interaction

Peptide antigen fit into the binding groove on top of the peptide binding domain, which is formed by two roughly parallel α helices above a floor of β plate, one helix and half of the plate come from each of the two subunits, α1 and α2 for class I, and α1 and β1 for class II [19].

Peptides that bind to MHC class I usually have 8-10 residues, since both ends of the binding groove are closed. Peptide residues occupy 6 binding pockets along the binding groove, namely A to F, among which two deep pockets B and F correspond to anchor residues on peptide, usually P2 and PΩ, and contribute most to peptide MHC interaction (Fig. 1-1B). In peptide-MHC I complex (pMHC I), the position of both ends of peptides are conserved, while the middle section show flexibility, as it often bulges from the groove [19].

Class II molecules have open binding groove thus can hold longer peptides, usually 13-25 residues long, with both ends of peptides extruding from the groove. There are 9 pockets (P1-P9) assigned along the binding groove, accommodating 9-mer core region of consecutive residues on the peptide. Four deep pockets, 1, 4, 6, and 9, correspond to anchor residues of peptide (Fig. 1-2B). Compared to pMHC I, peptides that bind to MHC II adopt more linear conformation, and residues flanking the core region also contribute to the interaction with MHC II [19].

1.2.4 HLA polymorphism

HLA is highly polymorphic in order to interact with diverse antigens. As the most polymorphic region in human genome, there are huge number of HLA alleles recorded in IPD-IMGT/HLA Database, including 24,075 classical HLA I and 9,515 classical HLA II genes, encoding 13,970 and 5,907 unique proteins, respectively [22]. It should be noted that class II molecules are composed of two chains, thus the number of unique HLA II molecules is as large as 613,013. In addition, human beings are diploid, making the HLA haplotype extremely complex.

It is discovered that although peptide binding specificity of each HLA allele is unique, some alleles have largely overlapped specificities. To reduce the complexity in discriminating huge number of alleles, supertypes were defined to include alleles with similar functions, hence alleles in a supertype could be represented by a few representatives.

1.2.5 HLA nomenclature

To maintain data integrity, HLA alleles are given unique names according to the nomenclature adopted by the WHO Nomenclature Committee for Factors of the HLA System. The allele name starts with the HLA prefix and gene name, followed by up to four sets of digits that identify the allotype group, specific protein, synonymous DNA variations within the coding region, and DNA variations in non-coding regions, respectively. For example, alleles HLA-A*01:01:01:01 and HLA-A*01:01:01:02 belong to the same gene locus and allotype group, encode the same proteins, and have the same DNA sequences in the coding regions, but differ in the non-coding regions. For convenience, all alleles are referred to as unique proteins by the gene name and first two sets of digits, and the full names are used only when necessary.

1.3 Peptide-based vaccine targeting MHC class I molecules

Vaccines have been applied for several decades and are effective to reduce morbidity and mortality of infectious disease and cancer. Peptide-based vaccines use short peptide to elicit adaptive immune response (Fig. 1-4). According to the mechanism, peptide-based vaccines could be concluded into three types. The MHC class I epitopes, usually composed of 9-11 residues, activate cytotoxic (CD8+) T cells. The MHC class II epitopes, usually 15-mers, activate helper (CD4+) T cells. And antibody epitopes that stimulate B cells by multiple mechanisms.

1.3.1 Advantage of peptide-based vaccines

Conventional vaccines that rely on attenuated or inactivated forms of pathogens or components that contain mixture of multiple antigens, which stimulate polyclonal and long-lasting immune response. However, the conventional vaccines also have disadvantages: the complex component may induce allergic or autoimmune reactions, the attenuated or inactivated pathogen may return to its virulent state, and it is difficult to culture certain pathogens under laboratory conditions. As one of the fast-developing alternative techniques, compared to whole pathogens and proteins, peptide vaccines could elicit epitope-specific immune response and minimize the risk of allergic or autoimmune reactions, since peptide vaccines are designed to be epitope specific, and the purification of peptides is easy to perform. Also, because of the easiness of chemical synthesis and stability in storage condition, time and resource consumption in both development and production phases of peptides are much lower.

1.3.2 Development of peptide-based vaccines containing MHC class I antigens

Peptide-based vaccines, both preventive and therapeutic, could elicit immune response via the activation of T cells or B cells. Here, we focus on the short peptides, usually 8-10 residues, that bind to MHC class I and activate CD8+ T cells. In the context of infectious diseases, cytotoxic immune response is elicited by the pathogen specific peptides that generated by degraded pathogen proteins inside infected cells. While in immune oncology, cancer cells carry somatic mutations producing tumor specific peptides, which are also known as neoantigens, that leads to T cell antitumor activity.

Among all factors that contribute to effective vaccines, the selection of peptide antigen is the most important one. Such a peptide must be specific to pathogen or tumor, presented by MHC I in

abundance, and recognized by T cells. Correspondingly, experimental and *in silico* methods were developed for peptide selection. Based on whole exome sequencing (WES), the pathogen or tumor specific peptides could be identified. Affinity between peptide and MHC is the most widely accepted measurement of MHC presentation, thus the assays and predicting tools are actively developed. Still, the effective identification of immunogenic and naturally presented peptide remains a challenge.

1.4 Peptide MHC binding affinity prediction

One MHC allele can only bind a small portion of all possible peptides, which is described as peptide binding specificity. The selection of peptides is characterized by different affinity of peptides toward a certain MHC allele, and peptides with high affinity are more likely to be presented. Thus, determining peptide MHC affinity is extremely important for understanding the function of MHCs.

1.4.1 Experimental methods of measuring peptide MHC interactions

The competition assays are the most widely used method for determining peptide MHC affinity, which employs the competition of binding to MHC molecule between the test peptide and radiolabeled indicator peptide that strongly binding to such MHC allele. Purified MHC molecules, indicator peptides, and test peptides at different concentrations are first incubated together for at least two days, then the level of inhibitory of indicator peptides was measured. According to the dose-response curve (level of inhibitory versus test peptide concentration), the concentration of test peptide required to inhibit 50% indicator peptide binding (IC_{50}) is calculated, which is used to represent the affinity of test peptide. A low IC_{50} indicates strong binding, and $IC_{50} < 500$ nM are usually adopted as threshold of strong binders.

Scintillation proximity assay is used to measure the stability of pMHC I complex that is demonstrated to have higher correlate with immunogenicity. By mixing MHC I α chain, radiolabeled β chain, and test peptide, pMHC I complexes are formed, and its dissociation was monitored by consecutive measurement of the scintillation stimulated by radiation. The scintillation frequency and duration of experiment are fitted to one-phase dissociation model, where the half-life of complex is calculated and used to represent the stability. A half-life > 1h should be used as a threshold for immunogenic peptide.

The concept of MHC associated peptides (MAP) describes peptides that are presented by MHC molecules, which undergone protein splicing, selective transportation, MHC binding, and editing, thus are highly immunogenic. To identify MAPs, presented peptides are eluted from pMHC on cell surface, then sequenced using LC-MS.

1.4.2 Peptide-MHC binding affinity prediction methods

Measuring peptide MHC affinity via experiments is highly time and resource consuming. The huge number of both MHC molecules and peptides determines that it is impossible to measure the affinity between all possible combinations of peptides and MHCs via experiments. Therefore, *in silico* methods are developed, which could be roughly divided into approaches that based on scoring matrix, deep learning models, and structures.

Currently, deep learning methods are most widely used, as such methods have better performance than scoring matrix-based methods, while the computation is fast enough. Structure based methods calculate the binding affinity employing docking and molecular dynamics, which leads to high accuracy but consumes high amount of computation, thus is not suitable for high throughput prediction.

Among the deep learning-based methods, two training strategy were applied, allele-specific and pan-specific, the resulting predictors address different pros and cons. The allele-specific methods treat each MHC allele separately and corresponds to one neural network trained on its own peptide binding affinity data. Such methods achieve high accuracy, however, only cover a small number of alleles, as peptide binding assays are focused on populated alleles. The pan-specific methods train a single model that apply for all alleles, by combining binding data of different alleles, sometimes from different experimental methods. Such methods could be applied to theoretically all MHC alleles, but the accuracy is limited. In addition, both methods perform poorly on understudied alleles. With the accumulation of experimental data and development of deep learning algorithms, this issue is likely to be resolved in the future.

1.5 Multidrug resistance in bacteria

As one of the most revolutionary discoveries in modern medicine, antimicrobials have become the most important treatment against pathogenic bacteria. However, the evident increase in

antimicrobial resistance threatens the effective prevention and treatment of infectious diseases. To face this challenge, WHO adopted a global action plan on antimicrobial resistance [23].

1.5.1 Mechanisms of antimicrobial resistance

Since most antimicrobial substances are natural product, the resistance mechanisms have existed since ancient time as a result of evolution. It is hypothesized that the trade-offs between antimicrobial resistance and fitness in the absence of antimicrobials, including competitiveness, growth rate, and virulence, has limited the spread of resistance genes [24-27]. However, the extensive production and application of antimicrobials have broken the balance, as a result, common pathogens that was susceptible acquired resistance due to gene mutations or horizontal gene transfer [28].

The mechanisms of antimicrobial resistance could be concluded into four categories (Fig. 1-5): (1) permeability barrier. Intrinsically, the cell wall of bacteria is a natural barrier to antimicrobials, especially hydrophilic molecules such as β-lactams, tetracyclines and some fluoroquinolones since they rely on porins to cytoplasm and periplasm [29]. (2) modified drug target. A change in the drug target, for example a mutation in the key residue on an enzyme, may decrease even inhibit drug binding [30, 31]. (3) inactivating drug molecules. Antimicrobials are inactivated by degradation or modification [32-34], a most typical example of which is the β-lactamases [35, 36]. (4) efflux pumps. The efflux pumps actively transport a large variety of toxic molecules out of bacterial cell, thus enables the ability of resistance [37-39]. According to clinical observations, about 80% of the severe bacterial infections are caused by Gram-negative bacteria with multidrug resistance (MDR), which is mainly because of their two-membrane barrier and active efflux pumps.

1.5.2 Permeability barrier of Gram-negative bacteria

The Gram-negative bacteria are protected by two-membrane cell envelope outside the inner membrane, which is composed of the cell wall, a thin layer of peptidoglycan, and the outer membrane (OM). The OM is asymmetric bilayer, containing lipopolysaccharides (LPS) in the outer leaflet and phospholipids in the inner leaflet, with size-exclusion porins and selective transporters embedded. By combining the highly hydrophobic bilayer and porins that filter by size, the OM functions as a selective barrier that protect the bacteria without interrupting the exchange

of required material. Therefore, the permeability properties of this barrier have a strong impact on the susceptibility to drug molecules whose targets are located in the cytoplasm and the periplasm.

1.5.3 Efflux pumps

The efflux pump is the predominant mechanism of MDR. Efflux pumps are active transporters located on bacteria cell membrane that recognize and extrude noxious molecules from cytoplasm and periplasm. Currently, six families of efflux pumps have been identified in bacteria, including the ATP-binding cassette (ABC) family, the major facilitator superfamily (MFS), the multidrug and toxin extrusion (MATE) family, the small multidrug resistance (SMR) family, and the resistance-nodulation-cell division (RND) superfamily. The ABC family members are primary active transporters that directly consume ATP, and all other five families are secondary active transporters that rely on electrochemical energy of transmembrane concentration gradients.

1.6 Objectives of the thesis

The purpose of this thesis is to apply ML methods to empower the development of peptide-based vaccines and permeable small molecule antimicrobials. The peptide vaccine development mainly relies on prediction of peptides that strongly bind to certain MHC alleles carried by the population. However, the accuracies of widely used predictors are not satisfying among less studied alleles. In order to improve the performance of peptide-based vaccine development pipeline, we applied varies methods, knowledge-based and structure-based, to filter the prediction results. In chapter 2 and 3, we demonstrated the design of peptide-based vaccines targeting cancer and infectious disease employing affinity predictions, and several methods to refine the crude prediction result. In chapter 4 and 5, we attempted to improve the performance of affinity predictions using structure-based methods, as well as investigated the relationship of peptide binding specificities of HLA alleles. To aid the development of antimicrobials against multidrug resistant *Pseudomonas aeruginosa* strains, in chapter 6 we predicted the permeability of small molecules by using machine learning models based on their physicochemical features calculated from structures, which could facilitate the drug design and screening.

1.7 Conclusion

In conclusion, we developed and applied ML methods to aid drug development based on sequence and structure information, which showed improvements in accuracy and throughput of the pipeline.

Also, we demonstrated that the performance of ML methods directly relies on the availability of high-quality training data. Thus, the experiments, simulations, and ML methods are complimentary and should be performed with the same emphasis. With the accumulation of experimental data and development of ML algorithms, the ML methods are likely to improve further, benefiting human health.

Appendix

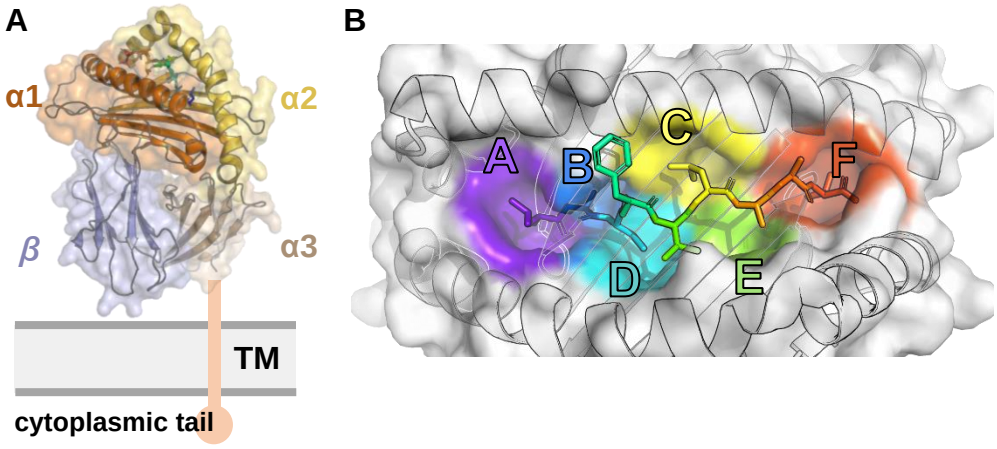


Figure 1-1. Structure of peptide-MHC I complex (PDB ID: 3to2). (A) The whole complex demonstrating the domains and peptide binding groove. The transmembrane domain (TM), cytoplasmic tail, and lipid bilayer are shown schematically. (B) The top-down view showing six characteristic binding pockets (labeled A-F) along the binding groove. The peptide is shown in rainbow color with the N-terminus in purple and the C-terminus in orange. Pockets are colored the same with closest peptide residue.

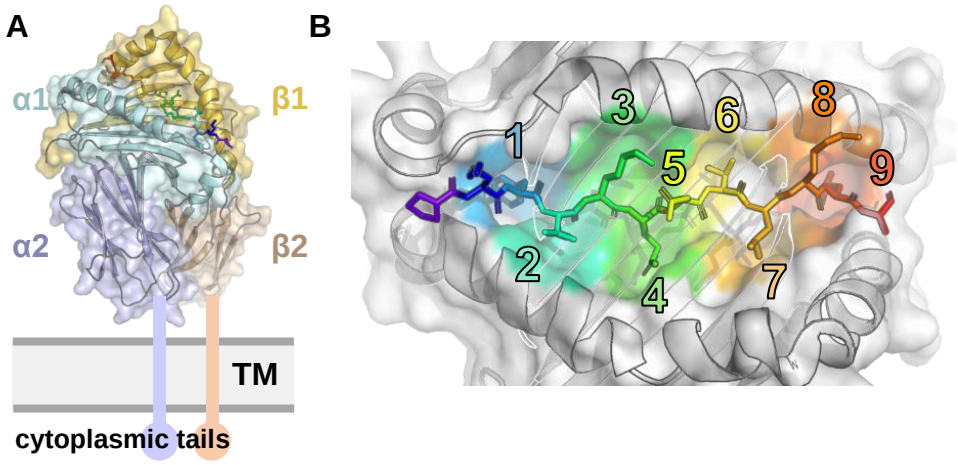


Figure 1-2. Structure of peptide-MHC II complex (PDB ID: 1dlh). (A) The whole complex demonstrating the domains and peptide binding groove. (B) The top-down view showing nine characteristic binding pockets (labeled 1-9) along the binding groove. The notifications are the same as Fig 1-1.

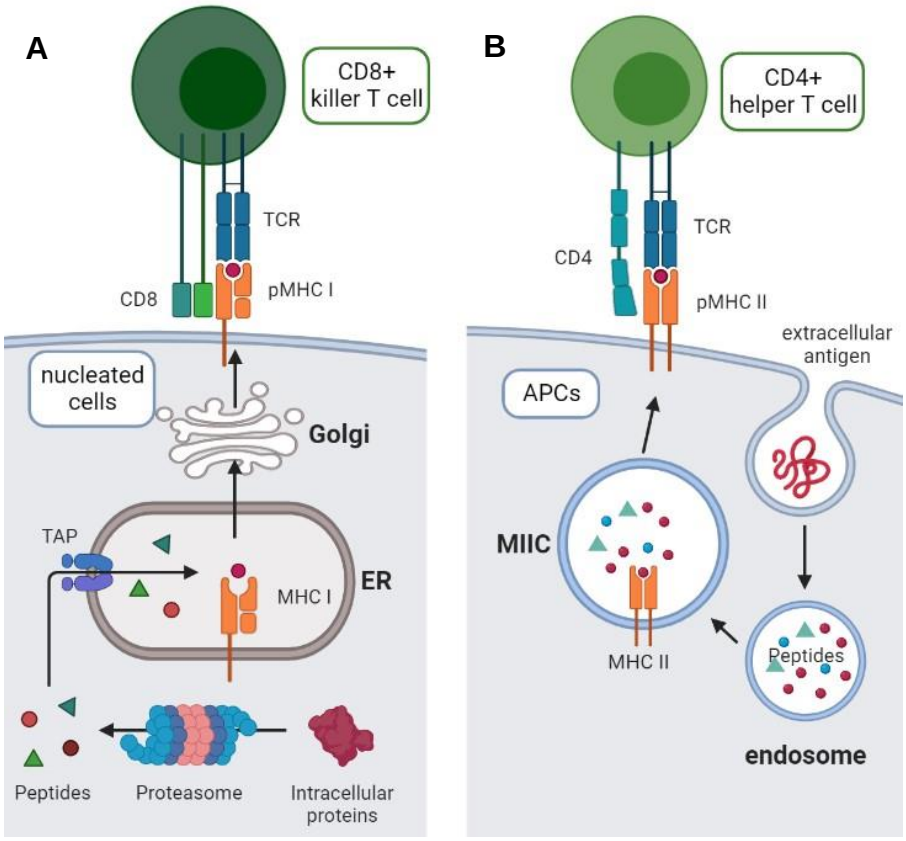


Figure 1-3. Peptide presentation pathway of (A) MHC I and (B) MHC II.

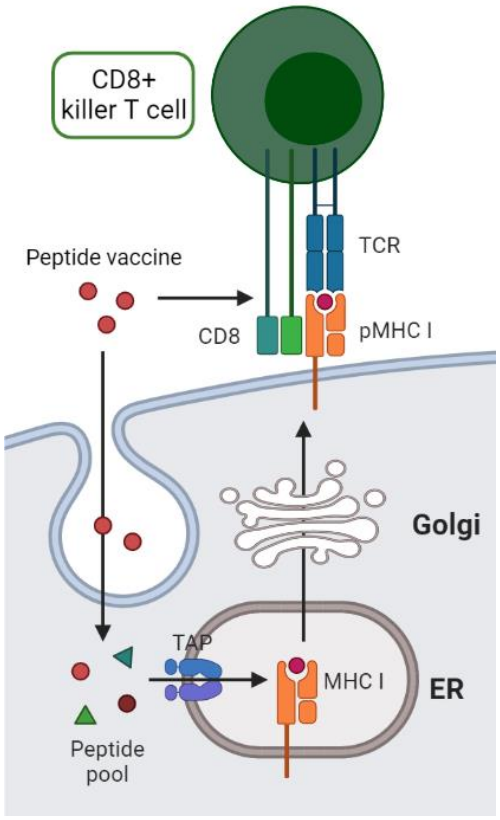


Figure 1-4. The mechanism of MHC I epitope peptide-based vaccines.

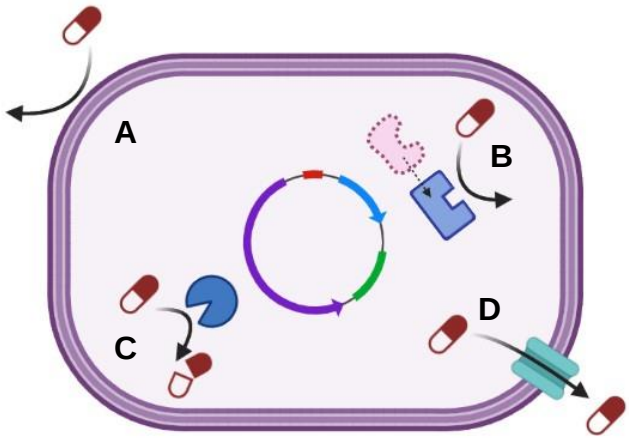


Figure 1-5. Mechanisms of antimicrobial resistance. (A) decreased permeability. (B) modified drug target. (C) inactivating drug molecules. (D) efflux pumps.

Chapter 2

Neoantigen Prediction Targeting Pancreatic Cancer

A version of this chapter was originally published as:

Reham Ajina, Zoe X. Malchiodi, Allison A. Fitzgerald, Annie Zuo, Shangzi Wang, Maha Moussa, Connor J. Cooper, Yue Shen, Quentin R. Johnson, Jerry M. Parks, Jeremy C. Smith, Marta Catalfamo, Elana J. Fertig, Sandra A. Jablonski, Louis M. Weiner; Antitumor T-cell Immunity Contributes to Pancreatic Cancer Immune Resistance. *Cancer Immunol Res* 1 April 2021; 9 (4): 386–400. https://doi.org/10.1158/2326-6066.CIR-20-0272

My contribution to this project is to build a pipeline line of predicting MHC class I restricted neoantigens from tumor and healthy tissue genomes. The original publication was trimmed to the parts that directly relevant to tumor growth and neoantigen prediction.

Author contributions:

R. Ajina: Conceptualization, formal analysis, validation, investigation, visualization, methodology, writing-original draft. Z.X. Malchiodi: Validation, investigation. A.A. Fitzgerald: Investigation. A. Zuo: Investigation. S. Wang: Investigation. M. Moussa: Investigation, visualization. C.J. Cooper: Data curation and formal analysis. Y. Shen: Data curation and formal analysis. Q.R. Johnson: Data curation, formal analysis. J.M. Parks: Data curation, formal analysis, validation. J.C. Smith: Data curation, formal analysis, validation. M. Catalfamo: Validation, visualization. E.J. Fertig: Data curation, formal analysis, validation. S.A. Jablonski: Supervision, validation, project administration. L.M. Weiner: Conceptualization, resources, supervision, funding acquisition, validation, writing-original draft.

Abstract

Pancreatic ductal adenocarcinoma (PDAC) is the third leading cause of cancer death in the United States. Pancreatic tumors are minimally infiltrated by T cells and are largely refractory to immunotherapy. Understanding how pancreatic tumors respond to immune attack may facilitate the development of more effective therapeutic strategies. In this study, we illustrated that T cell induced immune response partially regulates tumor growth. The predicted neoantigen epitopes were demonstrated to be highly tumor specific, showing the instability of pancreatic tumor genome and complexity of vaccine design. These findings provide future directions for treatments that specifically disable this mechanism of resistance in PDAC.

2.1 Introduction

Worldwide, nearly half a million people are diagnosed with pancreatic cancer every year [40], and the mortality of these patients within five years is more than 90% [41]. Therefore, there is an urgent need to develop more effective therapeutic strategies.

The PDAC is thought to arise from ductal epithelium that undergoes neoplasia [42]. In most individuals, this process is entirely somatic, while up to 10% of patients with PDAC have a germline predisposition to malignancy [43]. The PDAC is considered as an immunologically cold neoplasm, as the T cell infiltration is minimal. Even if some T cells successfully infiltrate pancreatic tumors, they are likely to skewed toward protumor function [44], got suppressed [45, 46], or physically trapped within the stroma [47]. As a consequence, vast majority of immune therapies targeting pancreatic cancer failed in clinical trials [48].

Albeit showing a myelosuppressive microenvironment, previous studies suggested that T cell antitumor response is efficacy against pancreatic cancer. Potent T cell cytotoxic immune response has been identified in long-term survivors [49]. Also, pancreatic cancer patients that exhibit high microsatellite instability are demonstrated to benefit from immune-checkpoint blockade therapy, which promotes T cell activity [50]. Thus, peptide-based vaccines that stimulate cytotoxic T cell activity are considered as novel potential treatment, and a better understanding of the mechanism of immune evasion will surely promote the development of immune therapies.

In this research, the potential peptide epitopes that are specific to mT3-2D tumor cells were predicted. The exomes of mT3-2D mice pancreatic cancer cells in immunocompetent wild-type

(WT) and severe combined immunodeficiency (SCID) C57BL/6J mice were compared, then the affinity of peptides that containing tumor specific point mutations to MHCs was predicted, finally strong binders were further screened using varies approaches. The differences in predicted neoantigen T cell epitopes between tumor under these two difference environments demonstrated that T cells could incompletely regulate tumor growth.

2.2 Materials and Methods

2.2.1 Data used in study

Whole-exome sequencing (WES) were performed for five WT mT3-2D subcutaneous tumors, five SCID mT3-2D subcutaneous tumors, and one mT3-2D cell sample. The experiment was carried out by Reham Ajina of Georgetown University Medical Center.

2.2.2 Variant identification

WES reads were aligned to the mm10 reference genome using the Burrows-Wheeler Alignment tool version 0.7.17 [51]. Genome Analysis Toolkit (GATK) best practices were used for sorting, duplicate marking, and variant calling with the HaplotypeCaller [52]. Variants were annotated using Ensemble-VEP version 90.9 [53].

2.2.3 Neoantigen prediction

There are three classical MHC class I loci in mice, equivalent to human HLA class I, including H-2K, H-2D, and H-2L. Laboratory mice are inbred so that each strain is homozygous and has a unique MHC haplotype, which is designated by a small letter. All C57BL/6J mice carry two copies of H-2Kb and H-2Db genes, and no H-2L genes [54].

All possible 9-mer peptides that included a missense variant were identified using pVACtools version 1.0.8 [55]. MHCI binding affinities (H-2-Db and H-2-Kb) were then predicted for each WT and corresponding mutant peptide with NetMHC 4.0 [56]. The following filtering steps were per- formed. Mutant peptides with predicted binding affinity (IC50) >500 nM were removed. Peptides from olfactory genes were excluded because they are prone to somatic mutations, are not highly expressed, and yield false positives as cancer-associated genes [57]. Mutations that corresponded to known polymorphic genes were also excluded. Peptides with low predicted agretopic indices [58], i.e., ratio of mutant to WT binding affinity <4, were removed (Fig. 2-2).

2.2.4 Peptide vaccine efficacy assay

The immunogenicity of predicted peptides was tested using IFN- γ ELISPOT assay, then their protection against pancreatic cancer in mice was examined. The experiments were carried out by Chris Priestly-Milyanta, Hong Zuo, and Sandra Jablonski of Georgetown University Medical Center.

2.3 Results

Because missense mutations may generate T-cell neoepitopes that can drive antitumor T-cell responses, WES and variant calling was performed to identify such mutations (Table 2-1). To determine the influence of in vivo tumor growth and antitumor immune responses on the generation of tumor neoepitopes, we analyzed the mT3-2D cell line, five WT tumors, and five SCID tumors. In total, we identified 1,248 peptides with predicted IC50 <500 nM. After filtering, 125 of these mutations were predicted to be putative cancer neoantigens (Fig. 2-2). Among the 14 potential neoepitopes that were shared among the three groups (Table 2-2), 13 belong to unique genes and seven were validated by Sanger sequencing. We also observed that many mutations were generated during in vivo tumor growth compared with the mT3-2D cell line. However, the majority of these mutations were not shared between the tumor samples. These observations suggest that antitumor T-cell responses in WT mice were likely tumor neoantigen-specific. In the follow-up ELISpot, the neoepitope QNYDYAVYL derived from Mrps35 gene was able to elicit activation of tumor infiltrating T cells (Fig. 2-3A), and vaccinated mice was demonstrated to have significantly lower tumor growth rate (Fig. 2-3B), demonstrating the accuracy of this neoantigen prediction pipeline.

2.4 Discussion

In this project, we have established a pipeline to predict neoantigens that have high affinity with provided MHC alleles in mice, which represent possible T-cell epitopes. In this mouse model, we identified 125 putative cancer neoantigens within 11 samples, with only 14 potential neoepitopes predicted to be shared between the three groups, half of which were validated by Sanger sequencing. The slow tumor growth rate in immunocompetent mice suggested that at least some of these mutations stimulated T-cell responses. However, it is likely that T-cell responses were also directed against tumor-associated antigens [59]. Although pancreatic cancers generally have

lower mutational load compared with other cancers, such as melanoma, several comprehensive genomic analyses have shown that human pancreatic tumors exhibit a considerable number of mutations. For example, Waddell and colleagues report that the average number of mutations/Mb in 100 human PDAC samples is 2.64, ranging from 0.65 to 28.2 mutations/Mb. More importantly, pancreatic cancers are also likely to express immunogenic mutations [60]. Balachandran and colleagues reported that the median number of putative neoantigens detected per pancreatic cancer patient is 38 in the MSKCC cohort and 32 in the ICGC cohort [61]. Bailey and colleagues demonstrate that the number of neoantigen-related mutations in human PDAC ranges from 4 to 4,000 neoantigens per sample [62, 63]. As a result, the detection of 125 putative cancer neoantigens within 11 samples suggested that the mT3-2D mouse model was reminiscent of human disease, supporting the notion that pancreatic tumors may use mechanisms to evade antitumor T-cell immunity even when PDAC tumors exhibit sufficient immunogenicity.

One major issue we encountered when performing neoantigen prediction is that the genome background of each individual mouse is unique but unknown, as the WES was not performed on every test subject. Instead, the normal genome used in variant calling was reference genome mm10, which leads to both false-positive and false-negative mutations. As a consequence, a number of individual-specific neoantigens were hidden, and extra validation step using Sanger sequencing was carried out. Also, to minimize the impact of possible inaccurate mutation information, we only collect neoantigens that derived from oncogenes. Neoantigens generated from mutation in other genes, also known as passenger mutations [64], were neglected. Such epitopes are also important targets for treatment and should be taken into account in future studies [65].

The mT3-2D cell line may reflect the immunogenicity of human PDAC and is considered a model for testing the concept neoantigen vaccines targeting pancreatic cancer. Compared to mice, human genome and MHC system are way more complexed. First, the human genome is bigger with more unidentified regions. Second, the HLA system possesses much more alleles, a large portion of which have no accurate peptide affinity prediction models. In addition, the oncogenesis in real patients is much more complicated than well-established animal model. These factors determined that the development of neoantigen vaccine still need further efforts. By illustrating the neoantigen

prediction pipeline, we demonstrated the general technique and rule of thumb of neoantigen vaccine development, which could facilitate future studies.

Appendix

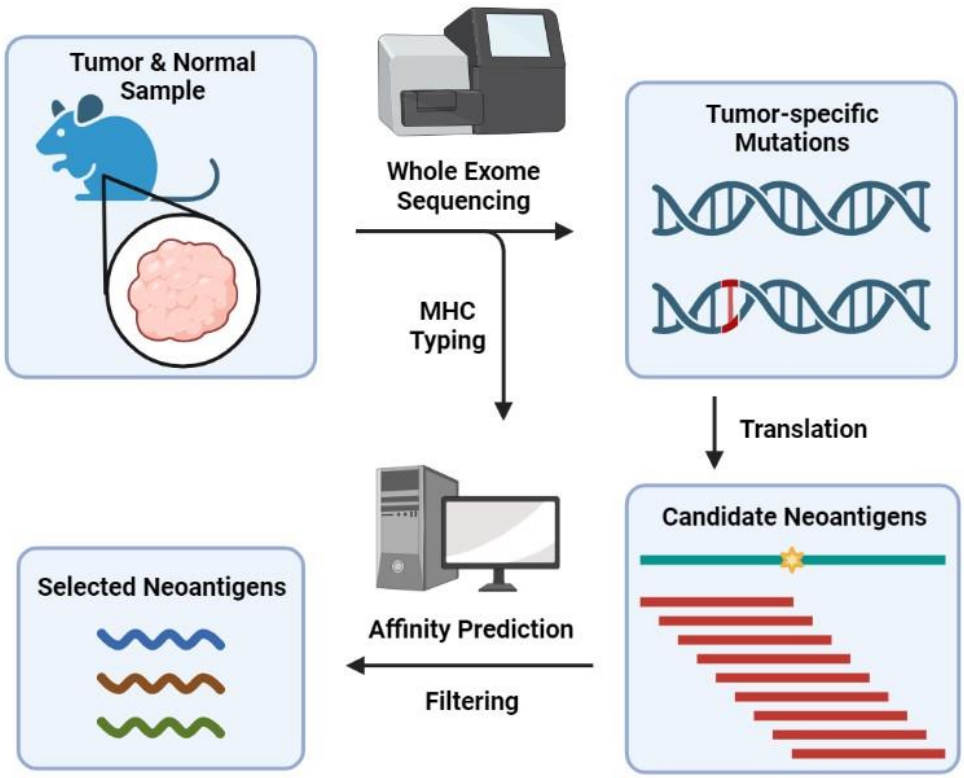


Figure 2-1. Pipeline for predicting antitumor neoantigen vaccines.

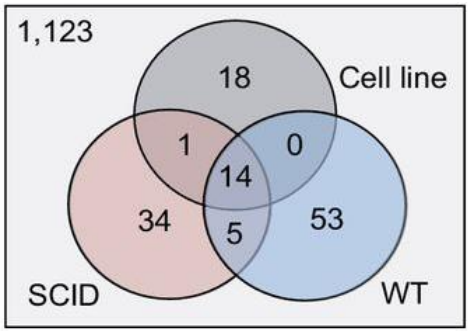


Figure 2-2. Number of predicted neoepitopes of mT3-2D cell line (n=1), tumor in WT (n=5) and SCID (n=5) mice.

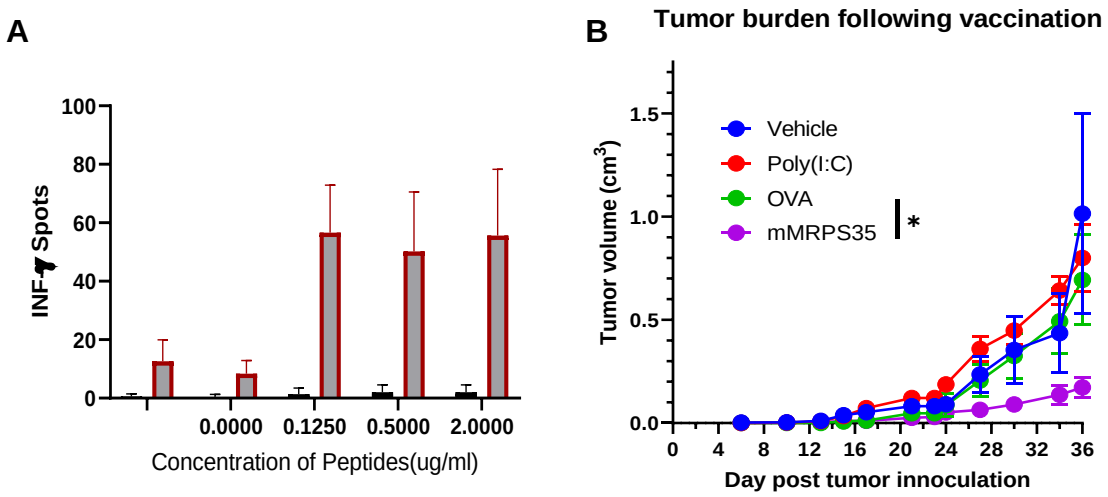


Figure 2-3. Predicted peptide vaccine (QNYDYAVYL) efficacy. (A) IFN-γ ELISPOT Assay on Mouse Splenocytes Stimulated with Predicted Vaccine Peptide. (B) Tumor burden following vaccination. Figure and data are generated by Chris Priestly-Milyanta, Hong Zuo, and Sandra Jablonski of Georgetown University Medical Center.

Table 2-1. Total number of missense mutations occurred in tumors in WT and SCID mice.

SCID	number of missense mutations	average
AGATCGC-35175-L1	10318	
ATCCTGT-35175-N	14593	
CCGTGAG-35175-R1	8140	11237.6
GACTAGT-35175-R2	10289	
GATAGAC-35175-L2	12848	
WT		
AGCCATG-31252-R1	14097	
AGTGGTC-31253-L2	14972	
CATCAAG-31253-R2	13691	14086.2
CTAAGGT-31253-L1	15296	
TGGCTTC-31253-R1	12375	

Table 2-2. Neoantigens predicted and filtered by the pipeline. The analysis was carried out by Connor J. Cooper of University of Tennessee, Knoxville.

MT_Sequence	Mutation	Position	MT_Score (nM)	WT_Score (nM)	WT/MT	Gene	Validated by sequencing
YAEKNRPPL	E/A	2	53	6166	116	AC168977.1	No
STFKKKRYL	E/K	5	140	680	5	AW822073	Yes
YNISLSTVI	M/V	8	437	1985	5	Ces1c	No
STFKKKRYL	E/K	5	140	680	5	Duxf3	No
VIPWIFARA	V/A	7	214	1524	7	Glp2r	Yes
VGLRFTSAT	N/T	9	328	6609	20	Mansc4	Yes
QNYDYAVYL	C/Y	3	23	792	34	Mrps35	Yes
TSYTQSTSV	S/Y	3	308	8950	29	Muc4	No
ASHTCHTFL	M/T	7	295	1404	5	Nlrc5	Yes
TSWQQFARL	R/S	2	8	168	22	Rrs1	No
KQRVYAQNL	R/Q	2	296	1689	6	Shank3	Yes
FTGEHFPRI	S/F	6	471	3247	7	Sirpb1a	No
SNLDFSIRI	N/S	1	66	578	9	Sirpb1b	No
VFSTLFALL	V/F	6	184	810	4	Tm7sf3	Yes

Chapter 3

Prediction of potential peptide vaccines from SARS-COV2 early-stage proteins

My contribution to this project was performing sequence analysis and predicting peptide antigens.

The epitope validation using flexible docking methods was performed by Michelle Aranha of Center for Molecular Biophysics.

The Sequence Analysis and Mutational Entropy Calculations section was included in:

Acharya, A., R. Agarwal, M.B. Baker, J. Baudry, D. Bhowmik, S. Boehm, K.G. Byler, S.Y. Chen, L. Coates, C.J. Cooper, O. Demerdash, I. Daidone, J.D. Eblen, S. Ellingson, S. Forli, J. Glaser, J.C. Gumbart, J. Gunnels, O. Hernandez, S. Irle, D.W. Kneller, A. Kovalevsky, J. Larkin, T.J. Lawrence, S. LeGrand, S.H. Liu, J.C. Mitchell, G. Park, J.M. Parks, A. Pavlova, L. Petridis, D. Poole, L. Pouchard, A. Ramanathan, D.M. Rogers, D. Santos-Martins, A. Scheinberg, A. Sedova, Y. Shen, J.C. Smith, M.D. Smith, C. Soto, A. Tsaris, M. Thavappiragasam, A.F. Tillack, J.V. Vermaas, V.Q. Vuong, J. Yin, S. Yoo, M. Zahran, and L. Zanetti-Polzi, *Supercomputer-Based Ensemble Docking Drug Discovery Pipeline with Application to Covid-19.* Journal of Chemical Information and Modeling, 2020. **60**(12): p. 5832-5852.

Author contributions in this article:

All authors assisted in the drafting and revising the manuscript. A.A. assisted in MPro protonation state selection. R.A. assisted in AD-GPU for Summit development and testing, he also assisted in clustering and performed ensemble docking calculations. J.B. performed clustering, ensemble docking, and docking analysis and curated repurposing ligand database. M.B.B. created and deployed ligand pre-processing pipeline at scale on Summit. S.B. assisted in the development of workflow management solutions for deploying giga-docking calculations on Summit. D.B. assisted in development of preliminary AI-based clustering analysis tools. K.G.B. performed ensemble docking calculations and assisted in docking analysis. S.Y.C. assisted in the development and deployment of NLP tools for rapid literature review. L.C. directed production of room temperature crystal structures for MPro and assisted in MPro protonation decisions. C.J.C. performed ensemble docking calculations and assisted in target assessment. O.D. assisted in preliminary rescoring of docked ligands. I.D. assisted in MPro protonation state selection. J.D.E. performed ensemble docking calculations. S.E. performed ensemble docking calculations. S.F. assisted in development of AD-GPU for Summit for Summit. J.G. assisted in the development of giga-docking workflow and giga-docking analysis techniques. J.C.G. assisted in MPro protonation

state selection and spike target discussion. J.G. assisted in drafting the manuscript and suggested architecture-specific acceleration plans. O.H. led and facilitated the development of AD-GPU for Summit. S.I. co-lead QM Analysis of S-Protein Docking Results. D.W.K. produced the room temperature protein structure of the Main Protease from SARS-CoV-2 (6WQF). A.K. produced the room temperature protein structure of the Main Protease from SARS-CoV-2 (6WQF). J.L. assisted in development of AD-GPU for Summit for Summit. T.J.L. led evolutionary entropy analysis. S.L. assisted in development of AD-GPU for Summit. S.H.L. performed ensemble docking and hot-spot analysis. J.C.M. assisted in hot-spot analysis and provided feedback on protein-protein interface targeting. G.P. assisted in development and deployment of NLP tools for rapid literature review. J.M.P. assisted in target selection and MPro protonation state selection. A.P. assisted in MPro protonation state selection. L.P. performed ensemble docking calculations and assisted in clustering. D.P. assisted in development of ADGPU for Summit. L.P. assisted in development and deployment of NLP tools for rapid literature review. A.R. led preliminary AIbased clustering efforts. D.M.R assisted in molecular dynamics analysis and development of AD-GPU for Summit workflow. D.S.-M. assisted in development of AD-GPU for Summit. A.S. assisted in AD-GPU for Summit. A.S. led literature review, assisted in AD-GPU for Summit testing and development, and assisted in mutational entropy calculations. Y.S. assisted in ligand library management and entropy calculations. J.C.S.* conceived project idea and coordinated collaborative efforts. M.D.S. initiated preliminary study, led T-REMD simulation efforts, assisted in clustering, co-designed molecular dynamics analysis, and prepared systems for simulation. C.S. assisted in development and deployment of NLP tools. A.T. assisted in development of AI-based clustering tools. M.T. assisted in testing and development of AD-GPU for Summit. A.F.T. assisted in development of AD-GPU for Summit. J.V.V. assisted in the development of AD-GPU for Summit, performed three TREMD simulations, generated figures assisted in analysis of MD trajectories, and benchmarked AD-GPU performance. V.Q.V. performed QM refinement and rescoring of ligand-spike docking results. J.Y. assisted in development and deployment of preliminary AI-based clustering. S.Y. led NLP tool development and deployment. M.Z. performed ensemble docking calculations and assisted in clustering analysis. L.Z.-P. assisted in MPro model assessment and protonation state selection.

Abstract

The outbreak of COVID-19 has raised alert to widespread infectious disease, therapies, including vaccines, are in urgent need. Due to the high infectivity, virulence, and mutation rates of SARS-CoV-2, the development of conventional vaccines could no longer satisfy the requirements. Peptide-based vaccines are demonstrated to have unique advantages over other vaccination strategies, thus are gradually valued and studied. In this research, we developed a method for predicting peptide-based vaccines targeting infectious diseases, especially viruses. This *in silico* methods could rapidly design vaccines composed of HLA class I and II epitopes using rival genome sequence and deep neural network predictors and will surely facilitates further research and development.

3.1 Introduction

Started from late 2019 and still ongoing, the pandemic of COVID-19 caused by SARS-CoV-2 has already caused over 611 million confirmed infection cases and 6.5 million deaths [3]. Also, the direct economic losses and health care burden brought by potential long-term effect led to huge impact on global development. In response to the pandemic, treatments are under active development, more than 7 thousand clinical trials were launched, yet the therapeutic options are still limited.

As a viable strategy against infectious disease, vaccines can effectively protect a large fraction of healthy individuals from infection and slow the spreading. Currently, multiple vaccination strategies are applied in the development of COVID-19 vaccines, including inactivated or attenuated whole virus, viral nucleic acids (DNA or RNA), recombinant proteins, synthetic peptides, and viral vectors [66]. Compared to other strategies, the peptide-based vaccines show advantage in safety and productivity, which is owning to the specific immune response, easiness of chemical synthesis and purification, stability in storage conditions as well as instability in human body [67].

SARS-CoV-2 is an enveloped, positive-sense single-stranded RNA virus that belongs to the genus of Betacoronavirus [68], which also includes severe acute respiratory syndrome coronavirus (SARS-CoV), human coronavirus OC43 (HCoV-OC43), HCoV-HKU1, and Middle East respiratory syndrome coronavirus (MERS-CoV). The entry of SARS-CoV-2 into human cells is

mediated by angiotensin-converting enzyme 2 (ACE2) [69-72], the same receptor as SARS-CoV [73] and alphacoronavirus HCoV-NL63 [74].

The SARS-CoV-2 virion is formed by structural proteins including nucleocapsid (N), membrane (M), envelope (E) and spike (S) proteins [75]. The lifecycle of viruses starts by entering the host cell mediated by the binding between S protein and ACE2 receptor. Next, viral RNAs were translated into non-structural proteins (NSPs) in capable of suppressing host mRNAs and producing viral RNAs and structural proteins. Finally, the newly synthesized viral molecules assembled into new viral particles and were released for further infection (Fig.3-1) [75, 76].[75, 76].

Multiple studies attempted to develop peptide-based vaccines against SARS-CoV-2 by predicting T cell and B cell epitopes, and most in-depth studies focus on structural proteins, especially the S protein, for their relatively high expression level and vital role in infection. However, the structural proteins are not a good target for vaccine for the possible low efficiency in clearing the viruses. The structural proteins are expressed in the late phase of viral life cycle [77], and are much more stable in human cells than NSPs [78]. On the contrary, the virion release starts shortly (8 hours) after infection [79], meaning that T cells that recognize epitopes derived from structural proteins can only decrease the virion production rate rather than prevent symptoms. Also, in response to immune pressure, the mutations of S protein appears frequent [80], and leads to significant functional and epidemiological change, including the D614G mutation associated with higher SARS-CoV-2 loads [81, 82], N439R with increased transmissibility [83], N501Y with enhanced binding to ACE2 [84], and E484K with antibody escape [85, 86]. The above concerns suggest that the NSPs expressed in the early stages of SARS-Cov-2 life cycle might be a better vaccine target.

In this study, we demonstrated *in silico* development method of peptide vaccines targeting SARS-CoV-2 early-stage proteins. The peptides were extracted from annotated viral genome, then the affinity to populated HLA I alleles was predicted. By applying multiple filtering methods, we suggested 21 peptides that are potential COVID-19 vaccines. The demonstrated method could be applied to the development of peptide-based vaccines targeting other pathogens.

3.2 Methods and Results

3.2.1 Structure of SARS-Cov-2 genome

The SARS-Cov-2 genome is 30 kb positive-sense, single-stranded linear RNA, encoding 14 open reading frames (ORFs) (Fig. 3-2). The ORF1a and ORF1b are translated into polypeptide 1a (pp1a, 440-500 kDa) and 1ab (pp1ab, 740-810 kDa), respectively, which are then cleaved into 16 non-structural proteins. To note, due to a -1 ribosome frameshift located at the upstream of the ORF1a stop codon, the translation of ORF1a would continue till the end of ORF1b, resulting in pp1ab. On the 3' side of the genome, there are 13 ORFs encoding four structural proteins: spike (S), envelope (E), membrane (M) and nucleocapsid (N) [87, 88], and nine accessory proteins: ORF3a, ORF3b, ORF3c, ORF3d, ORF6, ORF7a, ORF7b, ORF8, ORF9b, ORF9c and ORF10 [89].

3.2.2 Viral protein extraction

Genes were annotated according to the reference genome NC_045512.2. First, the ORFs on each unannotated genome, including both forward and backward directions, were captured using Python script. Next, the captured ORFs were compared to annotated reference genes using blastn command in BLAST+ package, the ORF with highest bitscore was designated as corresponding gene. The protein sequences were derived from translated gene sequences.

3.2.3 Sequence analysis and mutational entropy calculations

We performed an analysis of available sequences of the SAR-CoV-2 virus to look for numbers of mutations and map these locations on the proteins we were using as drug discovery targets. SARS-CoV-2 genomes were downloaded and aligned using NC_045512.2 (EPI_ISL_402125) as the reference genome. Shannon entropy [90] was calculated by Travis J. Lawrence of ORNL for every column of each protein alignment using a custom script. For visualization of the mutation entropy per residue of the proteins studied in this paper, entropy values were color-coded in protein PDB structures (Fig. 3-3). Known SARS-CoV and SARS-CoV-2 structures were downloaded from the Protein Data Bank, their sequences were aligned with the SARS-CoV-2 reference genome (NC_045512.2), and the calculated entropy of the sequences was embedded in the PDB file in the place of the B factor column using a custom Python script.

3.2.4 Prediction of candidate peptides

Peptides of 8, 9, and 10 amino acids were extracted from early-stage proteins, including NSP 1-16, using sliding window algorithm. Then the binding affinity of these peptides to 12 populated HLA alleles suggested by IEDB [91] were predicted using multiple predictors in pVACtools (version 1.5) [55]. The applied predictors included MHCflurry [92], MHCnuggets [93], NetMHC [56], NetMHCpan [94], PickPocket [95], SMM [96], and SMMPMBEC [97]. The final affinity was defined as the lowest score among all predictors. A high affinity threshold was set to < 500 nM for all alleles.

3.2.5 Candidate peptides filtering

Candidate peptides that were predicted to have high binding affinity with populated HLA alleles were further filtered using multiple methods, in order to select antigens of high immunogenicity. First, according to the mutational entropy, peptides that were located at the highly unstable regions were left out, as vaccines targeting region can only cover a subset of viral strains and will gradually lose efficacy with the accumulation of mutations. Second, the candidate peptides were searched in IEDB for T cell assays. The peptides that have positive experiment results were preferred. Also, because the TCR recognize HLA I restricted epitopes by the central region [15], we extracted position 4 to -2 (the second from C terminus) as the core region sequences and searched against reference human proteome to filter out epitopes that mimics auto antigens. Finally, the stability of pHLA complex composed of 9-mer candidate peptide and HLA molecules was validated by Michelle Aranha of Center for Molecular Biophysics, ORNL using structure-based method to remove false positives of high affinity binders [98]. As a result, 21 candidates that covers 9 HLA alleles were selected (Table 3-1).

3.3 Discussion

In this research, we demonstrated the method of *in silico* development of peptide-based vaccine targeting viruses, which is required to 1) bind to populated HLA alleles and 2) elicit strong and specific immune response. Besides the most widely used peptide-HLA affinity predictors, we also included other tests and filter methods to remove false positive candidates and increase true positive rate. This is valuable for further experimental assays as a high true positive rate of prediction results could significantly lower the time and effort consumption.

Apart from peptide vaccines that come into effect by binding to HLA molecules and activating T cells, the strategy of blocking pathogen infection or replication process by peptides was applied to fight against infectious diseases. One way is to block the function of viral proteins, for example the S protein, using competitive binding peptides, so that the viral entrance, replication, or immune suppression would be interfered. Also, peptides could be used to target B cell epitopes by binding to B cell receptors and elicit antibody response. These strategies could be combined to generate stronger immune response and broaden the coverage of therapies.

Some studies proposed strategies based on HLA class II epitopes that activate helper T cells. However, the prediction methods for peptide HLA II affinity are not well developed, thus the prediction accuracy is not satisfying. As a result, the development of HLA II epitope-based peptide vaccines still heavily relies on developer's experience and experimental assay, so it is not included in our research.

In real application of peptide-based vaccines, the flanking residues one both ends will be added to ensure the synthetic peptides last longer in human body. Apart from turning the identified epitopes into peptide-based vaccines directly, the result could improve vaccines employing other strategies, for example refining DNA/RNA based vaccines by introducing specific mutations to confront new virus strains or removing certain regions to avoid autoimmunity. Beside SARS-CoV-2, the demonstrated method could also be applied to other pathogens.

Appendix

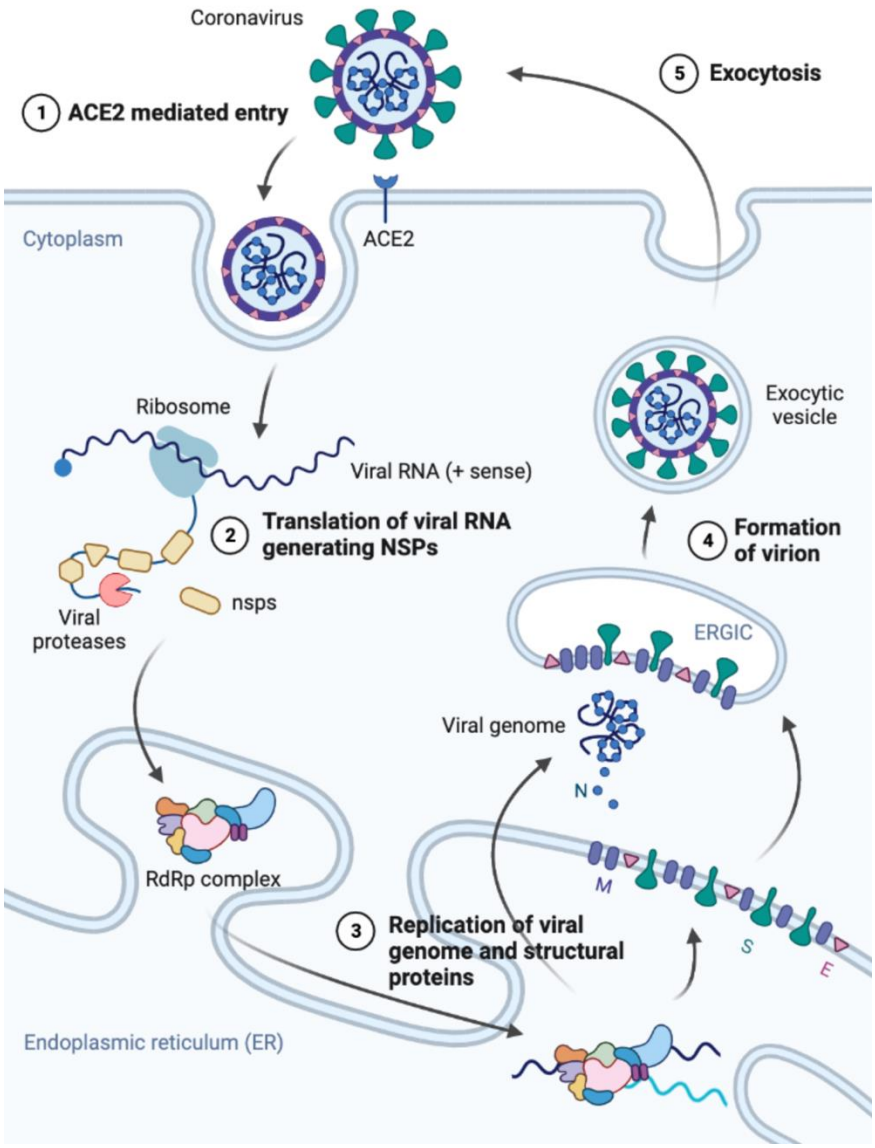


Figure 3-1 Life cycle of SARS-Cov-2.

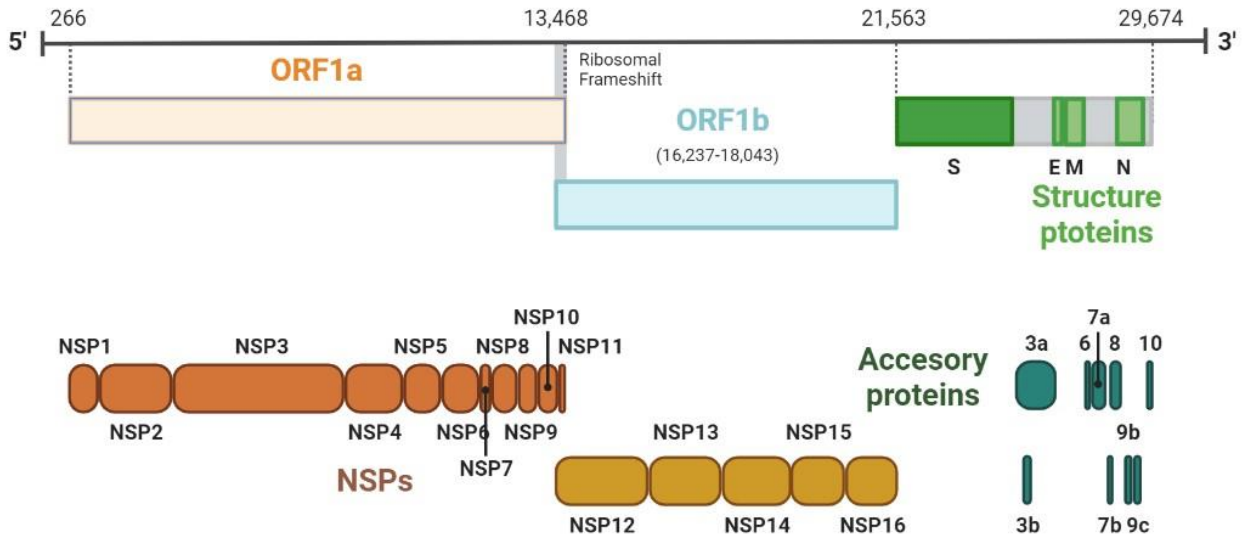


Figure 3-2. Genome structure of SARS-Cov-2. Genes and products were annotated according to NC_045512.2 reference genome.

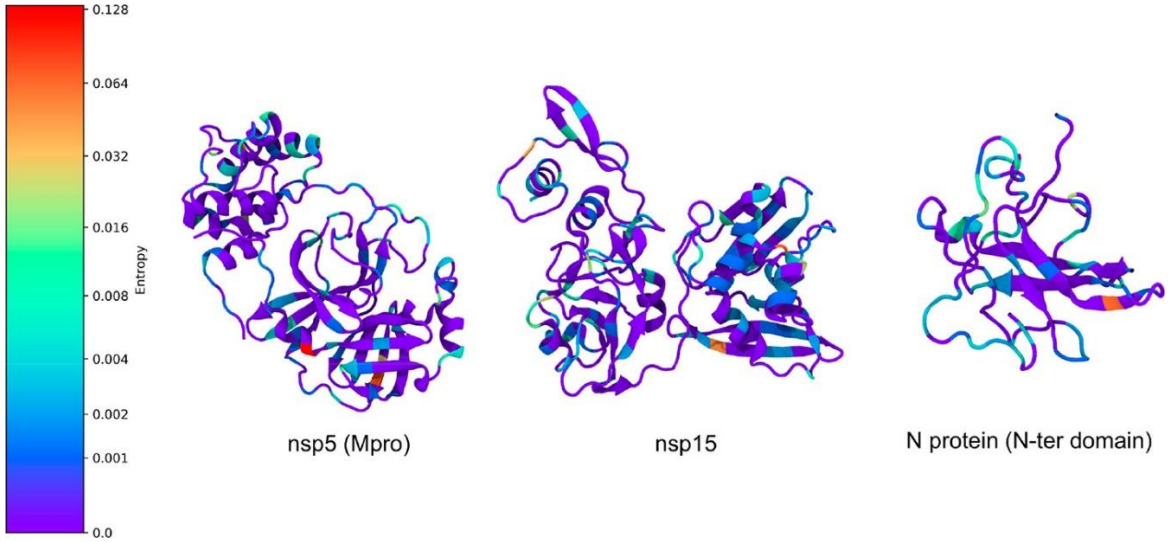


Figure 3-3. Example mutational entropy analysis. Residues are colored by entropy, with redder colors corresponding to greater entropy. Figure was generated by Ada A. Sedova of Oak Ridge National Laboratory.

Table 3-1. Predicted and filtered peptide antigens. Preferred features were marked in green.

HLA allele	gene	Epitope Seq	conserved (1=in nCoV, 2=in all Cov)	IEDB T cell assay (positive/all)	Structure-based prediction (1=bind, NA=not tested)
A0201	nsp14	YLDAYNMMI	2	2/2	1
A0301	nsp3	RMYIFFASFY	2	1/1	NA
	nsp13	VVYRGTTTYK	2	1/1	NA
	nsp12	RLYYDSMSY	2	3/3	1
	nsp4	VLAAECTIFK	2	1/1	NA
	nsp10	TVCTVCGMWK	2	1/1	NA
A1101	nsp16	SSYSLFDMSK	2	1/1	NA
	nsp10	TVCTVCGMWK	2	1/1	NA
B0702	nsp6	MPASWVMRIM	2	1/1	NA
	nsp10	HPNPKGFCDL	2	1/1	NA
B0801	nsp13	FLGTCRRCPA	2	0/0	NA
	nsp3	FLGRYMSAL	2	0/0	NA
	nsp2	MGRIRSVYPV	2	0/0	NA
	nsp5	TPKYKFVRI	2	0/0	NA
B3501	nsp2	IAFGGCVFSY	1	0/0	NA
B4001	nsp12	SEMVMCGGSL	2	1/1	NA
B4402	nsp14	EEAIRHVRAW	2	1/1	NA
	nsp12	QEYADVFHLY	2	1/1	NA
	nsp9	TELEPPCRF	2	1/1	1
B4403	nsp12	QEYADVFHLY	2	1/1	NA
	nsp9	TELEPPCRF	2	1/1	1

Chapter 4

High Throughput HLA Homology Modeling Pipeline

Abstract

High quality HLA structures are necessary for the application of structure-based validation method for peptide-HLA binding. Homology modeling is a strategy suitable for building HLA structures since all HLA alleles are homologous and there are a number of HLA crystal structures available. However, currently available homology modeling pipelines and applications that runs on local machine require specifying templates and parameters task by task, while online services usually don't capacity large enough. In addition, modeling of HLA class II molecules requires multichain modeling ability, which is rarely supported in available methods. Here, we established a new homology modeling pipeline based on RossettaCM protocol that automated the modeling of HLA class I and II molecules, with easy-to-use command line user interface. By using parallel computing, polished algorithms, and optimized parameters, this pipeline is able to swiftly build models with high accuracy.

4.1 Introduction

Structure-based methods are demonstrated to be useful in studying protein-protein interactions. In the prediction of peptide HLA affinity, structure-based methods that employ docking and molecular dynamics are demonstrated to show high accuracy [98]. However, such methods are limited by the availability of high-quality HLA structures.

Due to the extreme polymorphism, not all HLA alleles have crystal structures available. To minimize such a gap and provide insights in peptide HLA interactions, *in silico* structure prediction methods provide viable alternatives. Compared to *ab initio* approaches, homology modeling requires much less computational cost, thus is the most widely accepted strategy for modeling proteins that have homologous structures available [99].

The homology modeling is based on the observations that the protein structure is primarily determined by its amino acid sequence, and the structure is more stable than sequence during evolution. The higher the sequence similarity, the more confident that the two proteins have similar folding. Referring to the HLA molecules, despite the huge number of alleles and multiple loci (Table 4-1), they are highly similar in protein sequences (Table 4-2). In addition, most mutations among HLA alleles are point mutations, while the insertions and deletions are rare, making the protein folding consistent [22]. Thus, homology modeling is a suitable approach for modeling

HLA molecules. We choose to build the homology modeling pipeline based on RosettaCM protocol [100], which was demonstrated to outperform other methods when the sequence similarity between target protein and template is high.

The RosettaCM could be summarized in the following steps: template selection, target-template sequence alignment, model building, optimization, and validation. The model building and optimization steps are the most computationally intensive, as Monte Carlo sampling method is applied, and multiple candidate models are built to choose from. As the development of high-performance computing (HPC), large scale parallel computing could perfectly solve this issue because of the parallel nature of Monte Carlo methods.

In conclusion, we developed high throughput HLA homology modeling pipeline. Steps including template selection, job management on HPC cluster, and model picking, are automated via python scripts. The resulting models were validated to have high accuracy and could be applied in structure-based HLA studies.

4.2 Methods

4.2.1 Collection of HLA sequences and crystal structures

HLA protein sequences used as modeling targets were downloaded from the IPD-IMGT/HLA database [22], which were extracted from locus-wise MSA files rather than plain sequence files to acquire the residue numbering information. The target sequences were trimmed to peptide binding domain(s) and membrane proximal domain, since the TM domain and cytoplasmic tail is missing in all currently available HLA crystal structures thus cannot be accurately modeled. In total, there are 13,970 class I and 5,907 class II alleles collected.

Crystal structures used as templates were downloaded from the IMGT/3Dstructure-DB [101]. The allele name (nomenclature) of each crystal structure was validated by extracting the protein sequences and comparing to the record in the IPD-IMGT/HLA database. For alleles that have multiple crystal structures available, the one with best resolution was chosen. There were 64 class I and 16 class II crystal structures collected. Structure cleaning was performed with PyMOL (version 2.5.2) [102], during which water molecules, ions, binding peptides, and the beta chain specifically for class I, were removed.

4.2.2 General information of the modeling pipeline

The homology modeling pipeline was built to automatically model HLA class I and II structures, based on RosettaCM and HPC cluster. The RosettaCM protocol suggests practicing modeling in the following steps: *partial_thread*, *fragment_picker*, *hybridize_mover*, and *FastRelax* (Fig. 4-1), which is not optimized for multichain proteins and large number of proteins. To automate the pipeline and accommodate parallel computing on HPC cluster, modifications were made: 1) We added a setup step at the beginning of pipeline for template selection and performing pairwise sequence alignment between target and template; 2) The *fragment_picker* was modified for multichain modeling; 3) We re-wrote *partial_thread* in Python for improved speed and easy-to-use; 4) The unbiased relaxation step using FastRelax was replaced by constraint relaxation implemented in the model building step using hybridize_mover. In addition, files and job information were maintained using Python scripts, including *database_access.py*, *job_center.py*, *job_launcher.py*, and *job_pick.py*, providing command line user interface for users to view, launch, edit, and collect results from modeling. The details of each step are illustrated in the five following sections.

4.2.3 Setup: clean and perform sequence alignment

The pipeline starts from MSA files downloaded from IMGT/HLA database, which is not included in the published package. In the setup step, user should get such sequence files and execute *IMGT_input.py* to generate sequence file containing all non-redundant alleles. Then, execute *database_access.py* initial function to generate job record file and specify the root directory for the project. Alleles that have sequence length shorter than a threshold (the default is 100) are inactivated. Finally, the *job_center.py* should be run to create hierarchical tree of directories, each allele was assigned to one directory. For class II, since the α chains and β chains are in separated files, this step also combines both chains into one file in cross product style, which means that each α chain will be combined with each β chain. Template selection is based on the bitscore calculated by BLOSUM90 substitution matrix using *blastp* function from the BLAST+ package. For class I, simply the crystal structure that have highest bitscore was selected as template. While for class II, the crystal structure that have the highest sum of bitscore of both chains was selected.

4.2.4 Threading: thread target sequence onto template backbone

The py_thread.py script was used for threading, in the place of ROSETTA partial thread module that has significant shortcomings. The py_thread function was designed to address the following issues of ROSETTA partial thread: 1) Only accept sequence alignment input in unique Grishin format. Unlike other alignment file format such as FASTA, CLUSTAL, and PHYLIP, the Grishin format is unique, thus there's no existing parser for this format. Also, the Grishin format has strict restriction on name space. The pdb file name corresponds to Grishin file header, strictly 5 letters long. Thus, it's not portable and robust for large number of jobs. 2) Lack multi-chain support. Inherited from the Grishin format, the ROSETTA partial thread module doesn't distinguish between multiple chains, which means it's impossible to operate on a specific chain. The threading operation is executed from the start position specified in Grishin file to the end of sequence file. Threading on multi-chain template is possible but need to give full sequence of all chains. Also, the partial thread seems to have problem when overwrite cysteine that forms disulfide bond. 3) It is slow. The partial thread is built with standard ROSETTA API which aims at providing uniform flags and options for all modules in ROSETTA. So, several setting files specifying parameters are needed. It adds more unnecessary operations. The module itself is slow as well, which needs ~ 30 seconds to finish. It is also inconvenient for submitting jobs on HPC cluster. If the threading is performed interactively on login node, it takes substantial time since the number of HLA alleles is huge. If the threading jobs are carried out in compute nodes, the queue management system will be flooded by large number of small jobs, causing unnecessary waiting time in queue.

In comparison, our rebuilt py_thread function shows advantages. It takes sequence alignment input in FASTA format, which better supports multichain threading, with significantly less runtime. The script was written in Python with simple logic (Fig. 4-2), using functions implemented in BioPython.

4.2.5 Fragment picking: select fragments for sampling folding

The Rosetta structure prediction algorithm starts from Monte Carlo searches in folding space, represented fragments extracted from a database of known structures. The use of fragments decreased the size of folding space that needs to search against comparing to whole molecules. The ROSETTA *fragment_picker* function picks the fragments from predefined database based on the sequence profile and secondary structure similarity. The sequence profile is generated via PSI-

BLAST [103], based on which the secondary structure was predicted by PsiPred [104]. However, the PSI-BLAST is slow since searching against the general database is redundant for HLA. To speed up this process, we bypassed the searching step in PSI-BLAST and generate sequence profile via MSA provided by IMGT/HLA database. Also, we designed new front-end scripts for PsiPred replacing the original one for better file management. Then the sequence profile and predicted secondary structure is provided to ROSETTA fragment_picker, which is executed in multithread mode on HPC cluster, generating 3-mers and 9-mers fragment file.

4.2.6 HybridizeMover: build models

As the major step, the models were built via *HybridizeMover* implemented in *rosetta_scripts* function, following the RosettaCM protocol. This process could be concluded in three steps. First, the coarse-grained model was built based on the threaded model and fragments. Second, the model was upsampled to all-atom model, and undergo further optimization. Finally, the model was relaxed with constraint on backbone to minimize potential clashes.

4.2.7 Model selection: pick final model based on energy score

The standalone relaxation step was omitted, because the HLA structures were modeled in apo form and unbiased relaxation worsen the model with collapsed binding groove. For each allele, multiple replicas were built to choose from. Based on total score calculated by *ref_2015* energy function [105], the replica with lowest score was selected as final model.

To optimize the number of replicas that achieves the balance that minimizes computational cost and also ensures the near-native folding is sampled, homology models of allele HLA-A*02:01, one of the most studied alleles, were built with different number of replicas. The distribution of total scores for each number of replicas was shown in histogram, visualized using matplotlib.

4.2.8 Model validation

To access the model quality, models of alleles that have crystal structures available was predicted using second best template (the best template is of course itself) and were compared to corresponding crystal structures. The use of second-best template reproduced the situation that most target allele and available template have differences. The similarity between models and crystal structures was measured by Root Mean Square Deviation (RMSD) using PyMol.

4.3 Results

Here we introduced the pipeline that automates homology modeling based on Rosetta CM, which is built on Python scripts, with command line user interface that handles HPC jobs and directory structures. Scripts and instructions are available at https://github.com/yshen25/rosetta_CM.

From the distribution of total score calculated by ref_2015 energy function, the optimized number of replicas was selected as 100, as it has similar distribution with that of more replicas (Fig. 4-3).

The quality of models measure by RMSD between homology models and crystal structures of the same allele shows that they are highly similar, though the models were built using second best template (Table 4-3). Usually, an RMSD less than 1 Å is considered as very high model quality, while this pipeline achieves an average RMSD of 0.74 Å among 28 tested alleles, demonstrating high confidence that the models are correct.

4.4 Discussion

Structures are closely related to function, and the use of structure-based method should provide more insights into studies of protein functions. However, the availability of high-quality structures is limited, as determining protein structure via experimental methods requires significant time and effort. Quite a number of HLA crystal structures have been released, however, compared to the huge number of alleles, the availability of structures is still low. The homology modeling perfectly fits this situation to build structures based on closely homologous structures. With the development of HPC, the large-scale parallel computing for homology modeling is possible, which enables to predict structures for huge number of HLA alleles to decrease the knowledge gap and support structure-based HLA studies and tools.

Rosetta package provides multiple functions for molecular modeling and analysis, including, but not limited to, ligand docking, protein design, and structure prediction. To optimize the large-scale homology modeling on HPC cluster, we developed this automated pipeline using Python scripts implementing publicly available libraries. Different from other homology modeling solutions, for example the Robetta server that is also based on RosettaCM protocol, the current pipeline focuses on HLA structures using customized sequence and structure databases, as well as modified algorithms to achieve better performance.

Appendix

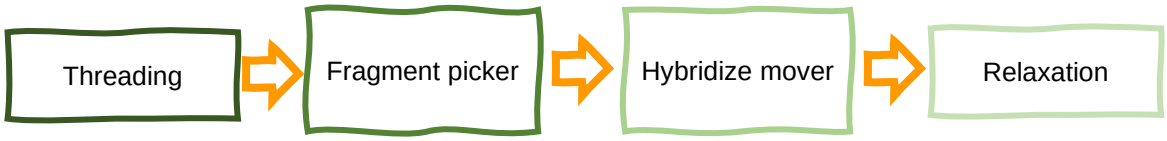


Figure 4-1. Homology modeling pipeline suggested by RosettaCM protocol.

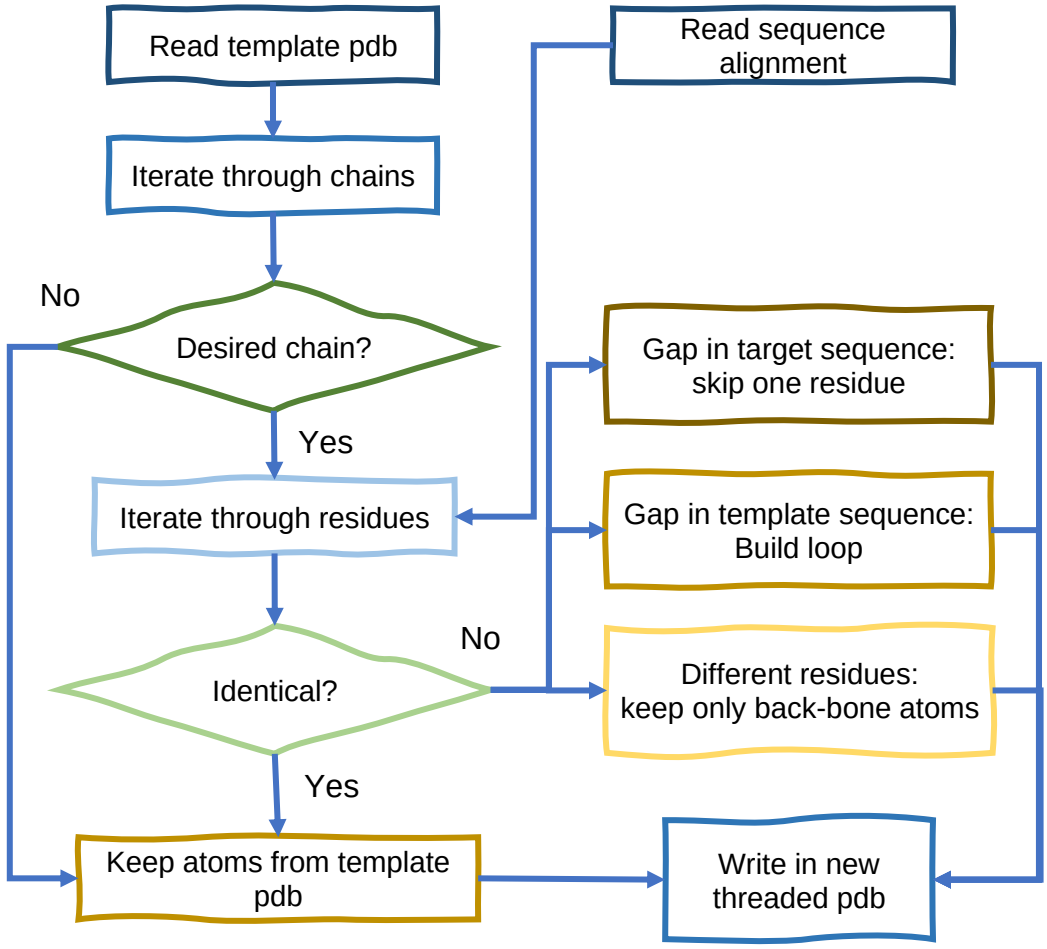


Figure 4-2. The flow chart describing the algorithm of py_thread.py script.

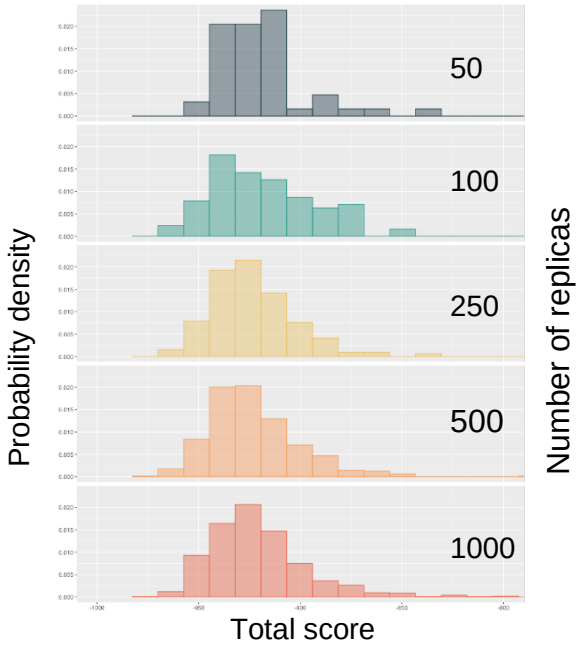


Figure 4-3. Histogram of total score of HLA-A*02:01 models built with different number of replicas.

Table 4-1. The total number of HLA alleles and unique proteins. Data collected from IPD-IMGT/HLA database in November 2022.

Gene	HLA I			HLA II									
	A	B	C	DRA	DRB	DQA1	DQA2	DQB1	DQB2	DPA1	DPA2	DPB1	DPB2
Alleles	7644	9097	7609	43	4258	508	40	2330	18	491	5	2221	6
Proteins	4450	5471	4218	5	2823	244	11	1455	8	223	0	1321	0

Table 4-2. The total number of HLA alleles with crystal structures. Data collected from IMGT/3Dstructure-DB in November 2022.

Gene	HLA I			HLA II					
	A	B	C	DRA	DRB	DQA	DQB	DPA	DPB
Crystal structures	413	264	40		102		33		6
Alleles with crystal	20	40	11	1	11	6	3	1	2

Table 4-3. Model quality measured be RMSD between models and corresponding crystal structures of the same allele. The homology models were build using second best template.

Target allele	Template	RMSD vs crystal(Å)
A*01:01	A*11:01	0.66
A*02:01	A*92:35	0.77
A*02:03	A*92:35	0.80
A*02:06	A*92:35	0.74
A*02:07	A*92:35	0.70
A*02:24	A*92:35	0.65
A*03:01	A*11:01	0.89
A*11:01	A*03:01	0.79
A*23:01	A*24:02	0.95
A*24:02	A*23:01	0.98
A*68:01	A*02:06	0.67
DPA1*01:03_DPB1*02:01	DPA1*01:03_DPB1*123:01	0.69
DPA1*01:03_DPB1*123:01	DPA1*01:03_DPB1*32:01	0.78
DPA1*01:03_DPB1*32:01	DPA1*01:03_DPB1*123:01	0.53
DPA1*01:03_DPB1*51:01	DPA1*01:03_DPB1*123:01	0.78
DPA1*02:02_DPB1*22:01	DPA1*01:03_DPB1*123:01	0.62
DQA1*01:02_DQB1*06:02	DQA1*03:01_DQB1*03:02	0.89
DQA1*03:01_DQB1*03:02	DQA1*05:01_DQB1*02:01	0.68
DQA1*05:01_DQB1*02:01	DQA1*03:01_DQB1*03:02	0.95
DRA*01:01_DRB1*01:01	DRA*01:01_DRB1*15:01	0.78
DRA*01:01_DRB1*03:01	DRA*01:01_DRB1*14:02	0.68
DRA*01:01_DRB1*04:01	DRA_01:01_DRB1*04:05	0.64
DRA*01:01_DRB1*04:05	DRA_01:01_DRB1*04:01	0.61
DRA*01:01_DRB1*11:01	DRA_01:01_DRB1*14:02	0.62
DRA*01:01_DRB1*14:02	DRA_01:01_DRB1*11:01	0.58
DRA*01:01_DRB1*15:01	DRA_01:01_DRB1*04:01	0.69
DRA*01:01_DRB3*01:01	DRA_01:01_DRB1*14:02	0.97
DRA*01:01_DRB5*01:01	DRA_01:01_DRB1*04:01	0.57

Chapter 5

HLA Class I Supertype Classification Based on Structural Similarity

Abstract

Human leukocyte antigen (HLA) class I proteins, a critical component in adaptive immunity, bind and present intracellular antigens to CD8+ T cells. The extreme polymorphism of HLA genes and associated peptide binding specificities leads to challenges in various endeavors, including neoantigen vaccine development, disease association studies and HLA-typing. Supertype classification, defined by clustering functionally similar HLA alleles, has proven helpful in reducing the complexity of distinguishing alleles. However, determining supertypes via experiment is impractical while current *in silico* classification methods exhibit limitations in stability and functional relevance. Here, we have developed a classification approach employing structural modeling, a newly defined structure distance metric, and hierarchical clustering that improves stability, resolution flexibility, and cluster cohesion relative to previous classifications.

5.1 Introduction

The complexity arising from HLA polymorphism, in particular the largely unknown yet dissimilar functions, i.e., peptide binding specificities, of alleles, is a challenging problem for researchers. For example, in peptide vaccine development, a key step is to find antigenic peptides that bind tightly to the specific HLA alleles carried by the patient, but only 150 alleles have experimentally characterized binding epitopes [106] and only 112 alleles have accurate allele-specific prediction models [92, 107]. Furthermore, the large number of alleles makes it is difficult to identify associations between individual HLA phenotypes and disease susceptibility [108, 109].

Although HLA alleles do differ in peptide binding specificities, they are not always functionally distinct. Since the 1990s, studies have shown that some HLA alleles have largely overlapping peptide binding specificity [110-116]. Accordingly, most HLA alleles could be clustered into supertypes and thus represented by a few typical alleles [117, 118], which greatly reduces the difficulty of discriminating between the huge number of HLA alleles.

Determining supertypes via experiments is very time and effort consuming, and thus is impractical for classifying large number of alleles [119-121]. As viable alternatives, several in silico classification methods have been proposed. Among these sequence-based approaches cluster alleles based on global (whole sequence) [122, 123] or local (binding groove or contact residues) [124-126] sequence similarity using sequence alignment. Prediction-based approaches calculate

peptide binding specificity from predicted peptide-HLA affinities, instead of experimental data [127, 128]. Structure-based methods use 3D information derived from HLA structures, such as spatial similarity and molecular interaction fields [129], the number of hydrogen bonds, interface area, and the gap volume between HLA and peptide [130].

Current HLA I supertype classifications have proven to be helpful and are widely used but have limitations. Some methods cluster alleles based on sequence or structural similarities between the molecules. However, the correlation of sequence or structure with functional similarity is not well established, and thus the resulting supertypes are not guaranteed to include functionally similar alleles. Prediction-based methods are limited by the coverage and accuracy of prediction methods, as allele-specific predictors have good performance but cover a very limited number of alleles, while the pan-specific predictors are less accurate though have better coverage, and both types of predictors perform poorly on rare or uncharacterized alleles [93, 131]. Structure-based methods are potentially more informative than sequence-based methods but have been limited by the availability of high-quality HLA structures and the overall complexity of structure analysis. In addition, due to the existence of inter-locus interactions and coevolution of HLA genes, association analysis on the haplotype level has advantages over single-locus genotype approaches [132, 133], but most widely used supertype classifications are locus specific, and thus are not capable of describing functional relationships between alleles at different loci. These limitations indicate that advanced approaches to supertype classification are needed to further facilitate HLA studies.

In this work, we explore the use of 3D structural similarity to measure peptide binding specificity and cluster HLA alleles. By using the revolutionary structural modeling tool, ColabFold [134] and an automated analysis pipeline, issues of structure-based methods in structure availability and analysis complexity were addressed, enabling the present method to be straightforwardly extended over the whole of the HLA class I space and not be restricted to alleles with sufficient experimental data. Also, we establish that structural similarity between allele pairs is highly correlated with peptide binding specificity. Finally, based on structural similarity, 449 populated alleles are hierarchically clustered into 12 supertypes and 20 subtypes, giving flexibility in the resolution of describing epitope similarities between alleles. Compared to previous classifications, the present clustering method has better performance (cohesion), meaning that the classification better represents similarity in peptide binding specificity. Also, higher stability infers better confidence

and extensibility, ensuring that users can add more custom alleles without perturbing the existing classification structure.

5.2 Methods

5.2.1 HLA-Clus: HLA class I clustering method based on structure distance metric

HLA-Clus is a pipeline for clustering HLA class I alleles based on structure distance. The HLA I structures were modeled and coarse grained as cloud of labeled points, then the pairwise distances were calculated using newly defined structure distance metric that taking into account physicochemical similarities. Finally, alleles were clustered via hierarchical clustering or nearest neighbor clustering approaches.

Processing structures into labeled point cloud. HLA I structures in PDB format were represented by labeled point cloud in csv format, which include information of coordinates, amino acid name, and weight factor of each residue. First, structures were trimmed to peptide binding domain (residue 2-180). Then, trimmed models were superimposed onto the structure of peptide binding domain of allele HLA-A*02:01 (PDB id: 1i4f) by aligning the residues that form α-helixes and β-strands, as these residues are more conserved and stable across different alleles. In this way, all models were aligned, and no pairwise structure alignment is needed. Next, the structures were coarse grained, each residue was represented by the center of mass of its sidechain, while the backbone atoms were hidden. After that, weight factors were assigned to residues according to the importance of the position in determining peptide binding affinity, the details were described in the following section.

Residue weight factor. As residues in the binding groove have different levels of contribution towards peptide binding specificity, a weight factor was assigned to each residue (Table 5-1). According to a previous study, point mutation in 25 positions leads to altered peptide binding specificity compared to the wild type, a higher alteration means that the residue has more influence. We calculated the weight factors based on residue 7 that has minimal change in peptide binding specificity, 4.5%, which was given weight factor 1. The residue 24 leads to second least change, 5.2%, and was given weight factor $5.2/4.5 \approx 1.2$, and so on. Other residues that do not have significant contact with peptide were given weight factor 0 and ignored in calculating structure distance.

Coarse grained structure distance metric SD. Structural similarity between alleles were measured by newly defined structure distance metric SD, which was adapted from Sup-CK method, a point cloud based all-atom structure distance metric. We modified the metric for coarse grained models, addressing three major concerns: 1) to obtain improved calculation speed, 2) to incorporate weight factors for residues, 3) to implement physicochemical similarity that was usually studied at residue level. First, A kernel function is defined:

$$F(X) = \frac{1}{\cosh^k(\sigma \cdot X)} \qquad [1]$$

Such a kernel function turns distance between two residues X ($X \in R | x \geq 0$) into value between 0 and 1. Parameters k and σ fine tune the shape of the kernel function that determines the sensitivity to displacements between the two corresponding residues (Table 5-7). For large σ and k values, distant residues will have small kernel function output, while small σ and k are more tolerant thus distant residues will have higher values.

Based on this kernel function, the similarity between two alleles P1 and P2 was then defined as:

$$K(P1, P2) = \sum_{i \in P1} \sum_{j \in P2} \sqrt{w_i w_j} \cdot S_{ij} \cdot \frac{1}{\cosh^k(\sigma \cdot \|x_i - x_j\|)} \qquad [2]$$

In this equation, x_i represents the Cartesian coordinates of coarse-grained residue i, and $\|x_i - x_j\|$ is Euclidean distance between residues i and j. The residue similarity S_{ij} measures the physicochemical similarity between i and j, which will be described in detail in the next section. The weight factor w_i controls the importance of residue i, as described in previous section.

Finally, the structure distance *SD* is defined as follows, giving a metric that fulfills the axioms of distance metric including minimality, symmetry, and triangle inequality:

$$SD(P1, P2) = \sqrt{K(P1, P1) + K(P2, P2) - 2K(P1, P2)} \qquad [3]$$

Residue similarity matrix S. A similarity matrix was used to determine the physicochemical similarity between two residues. Three similarity matrices were implemented for users to choose from: Grantham, PMBEC, and SM_THREAD_NORM. All three similarity matrices were normalized to [0, 1]. The Grantham similarity matrix was calculated as $1 - G/215$, where G is Grantham Distance which was based on the physicochemical property similarities between

residues (Table 5-8). The PMBEC, peptide-MHC binding covariance matrix, was derived from peptide MHC class I binding data. The SM_THREAD_NORM was based on energy potential calculated by THREADER force field. We suggest using the Grantham similarity matrix as it achieved best clustering cohesion compared to other two matrices, if the shape parameters were tuned properly (Table 5-7).

Calculating pairwise SD matrix. Given alleles that waiting to be clustered, the matrix of pairwise *SD* is first calculated, based on which the hierarchical clustering is then performed. Instead of calculating *SD* for each pair of alleles directly, the calculation procedure was optimized by the following method. First, the matrix of structure similarity K between each pair of alleles, including similarity to itself, is calculated in a single round-robin manner, to avoid repeated computation of $K(P1, P2)$ and $K(P2, P1)$. Next, when calculating the $K(P1, P2)$, calculations are performed between every residue in P1 and P2, resulting in 21×21 calculations, if weight factors are assigned as described in section 1.3. We split the calculation into three parts, including the kernel function, the similarity score S, and the geometric average of weight factors. All three parts are calculated, resulting in three 21×21 NumPy arrays, the product of which are then calculated, taking advantage of vectorized computation feature. Finally, the matrix of pairwise structure distance SD is calculated by eq.3, the similarity values are looked up in the similarity matrix.

Hierarchical clustering. Based on pairwise structure distance *SD* matrix, hierarchical clustering was performed using scikit-learn *AgglomerativeClustering* function with complete linkage method, in preference to a coherent and compact clustering result.

Nearest-neighbor clustering. To simplify calculation and maintain when clustering custom HLA alleles, an alternative nearest-neighbor clustering approach is proposed. Anchor alleles for each supertype and sub-supertype were used as cluster centers (Table 5-10), which are selected based on the standard that such allele has crystal structures or is supported by NetMHC, one of the best performing allele specific peptide-HLA affinity predictor, since those alleles are better studied with abundant experimental data. Two sub-supertypes, B50 and B57, don't include such alleles, thus their centroid alleles when performing hierarchical clustering are selected as anchors. Then for each allele needs to be clustered, the distances to these anchor alleles are calculated, and the tested alleles are classified to the supertype represented by the nearest anchor allele. Such an

alternative method has 96% consensus in supertype level and 94% consensus in sub-supertype level comparing to the hierarchical clustering method, while significantly simplifies calculation.

5.2.2 HLA allele frequency analysis

Due to the large number of HLA alleles, protein structural modeling and supertype classification were limited to populated alleles. Allele frequency data were derived from the Allele Frequency Net Database (AFND) [37]. Alleles with frequency > 0.01 in any population with more than 50 samples were determined as populated alleles and selected for subsequent structural modeling and clustering. There were 128 HLA-A, 235 HLA-B, and 86 HLA-C alleles that meet the above criterion (Table 5-4).

5.2.3 Collecting HLA sequences and crystal structures

Protein sequences of populated HLA alleles were downloaded from the IPD-IMGT/HLA database [38]. Crystal structures were downloaded from the IMGT/3Dstructure-DB [39]. The allele name of each crystal structure was validated by extracting the HLA α chain protein sequence and comparing to the record in the IPD-IMGT/HLA database. There were 397 crystal structures of 40 alleles collected (Table 5-5). Structure cleaning was performed with PyMOL (version 2.5.2) [40], during which water molecules, ions, binding peptides, and the beta chain were removed, leaving the alpha chain only.

5.2.4 Dataset split for validation purpose

Alleles were split into several datasets for different purposes. The model quality evaluation set was used to assess how closely the structural models reproduce experimentally determined crystal structures for the 40 alleles that have crystal structures available. The reference panel was used to estimate the performance of present and previous methods, including 31 HLA-A, 57 HLA-B, and 22 HLA-C alleles that were classified into supertypes with high confidence in previous studies (Table 5-6). The HLA-A and HLA-B alleles were taken directly from the reference panel used in ref [25], the supertype classification of which were based on experimentally established peptide binding motif. The HLA-C alleles were picked from the alleles of the consensus of the two supertype classification methods reported in ref [28].

5.2.5 Structural modeling

A total of 451 HLA structures, including 449 populated alleles plus 2 rare alleles that belongs to the reference panel (B*08:02 and B*27:09), were modeled using ColabFold [33], an implementation of AlphaFold2 [41], that accelerates predictions by using fast homolog sequence searching with MMseqs2 [42, 43]. The transmembrane (TM) domains were trimmed and not modeled, as the TM domain is far from the peptide binding groove and is expected to have a negligible influence on the folding of the binding domain. Also, because the models were evaluated by overall quality, a poorly modeled TM domain may dominate scoring of model quality, concealing details in the peptide binding domain. Models were generated with the AlphaFold2_batch.ipynb (version 1.3) Google Colaboratory notebook. For each allele, five models were built with three rounds of recycles. The model with the highest pLDDT score (predicted Local Distance Difference Test-Cα) was then relaxed using the Rosetta FastRelax protocol [44] with fixed backbone to minimize steric clashes and optimize side chain positioning, because the constraint is not available in current ColabFold Amber implementation. We generated 20 relaxed replicas for each model and selected the one with the lowest total score calculated by the ref_2015 Rosetta energy function.

5.2.6 Evaluating model quality

The quality of ColabFold models was measured by the Root Mean Square Deviation (RMSD) between models in the model quality evaluation set and crystal structures of the same allele. Some alleles have multiple crystal structures available. To represent the average of each of such alleles, the centroid structure was selected for comparison: first, the mean structure was generated in the way that the coordinates of any atom in the mean structure are the average coordinates of that atom in all crystal structures; next, the all-atom RMSDs between crystal structures and the mean structure were calculated using PyMOL; and finally, the structure with the lowest RMSD from the mean structure was selected as the centroid structure. The use of a centroid structure, rather than the mean structure itself, avoids comparisons with an unphysical structure obtained by averaging. To estimate natural structural variation, RMSDs between crystal structures and the centroid structure were also calculated.

Three kinds of RMSD were used to describe structural similarity: all-atom, backbone (Cα), and coarse-grained. The all-atom and backbone RMSDs were calculated using PyMOL, and the coarse grained RMSD between two structures, P1 and P2, was calculated with Eq. 1:

$$RMSD(P1, P2) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} {\delta_i}^2} \quad [4]$$

where δ_i is the distance between residue *i* of *P1* and the equivalent residue of *P2*, and *N* is the number of matching residues between *P1* and *P2*.

5.2.7 Peptide binding specificity distance, PD

The function of an allele as defined in the present study is represented by its peptide binding specificity. By this definition, the functional relationships between alleles are measured by peptide binding specificity distances, which were calculated using the method adapted from MHCcluster [127]. First, the binding affinities of each allele to a set of 50,000 random peptides of length 8-14 (Table 5-9) were calculated using the NetMHCpan 4.1 server [94]. The random peptides was generated by Expasy RandSeq tool [135], and the ratio of different lengths was in accordance with natural peptide length preference [136]. Next, the peptide binding specificity distance $PD(P1, P2)$ between two alleles *P1* and *P2* was calculated by the correspondence of the top 10% strongest binders (including $50{,}000 \times 10\% = 5{,}000$ peptides) of each allele, calculated as:

$$PD(P1, P2) = 1 - \frac{n(b1 \cap b2)}{5000} \quad [5]$$

where *b1* and *b2* are the top 10% strongest binding peptides of alleles *P1* and *P2*, respectively, and $n(b1 \cap b2)$ is the number of peptides that are strong binders to both alleles. If both alleles have the same set of strong binders ($n(b1 \cap b2) = 5{,}000$), the distance is 0, whereas for completely different sets of strong binders ($n(b1 \cap b2) = 0$) the distance is 1.

5.2.8 Tuning shape parameters σ and k

To select the optimal shape parameters σ and k, grid search technique was performed using scikit-learn *ParameterGrid* function. We tested 6 different values for σ and 5 values for k, 30 sets of combinations in total. For each set of shape parameters, pairwise *SD* was calculated and compared

to PD. We find that σ=0.3, k=1 achieved highest Pearson correlation coefficient between *SD* and peptide binding specificity distance, thus was chosen for suggested parameters.

5.2.9 Correlation between SD and PD

The locus-wise matrices of structure distance SD and peptide binding specificity distance PD of reference panel alleles were compared. Two pairwise distance matrices (SD and PD) for each of the three loci (HLA-A, HLA-B, and HLA-C) were calculated using Python scripts and were visualized as heatmaps by Matplotlib (version 3.4.3) [64] and Seaborn (version 0.11.2) [65]. The correlation between the two distances was further investigated via linear regression. Both SD and PD matrices of the same locus were normalized to the range in [0,1] using the scikit-learn (version 1.1.1) [66] MinMaxScaler function, then the PD was linearly fitted on SD using the SciPy (version 1.8.0) [67] linregress function. The correlation between PD and SD was derived with the Pearson correlation coefficient, which ranges in [−1, 1], and a larger absolute value of this coefficient indicates stronger correlation.

5.2.10 Assessing clustering performance

The performance of the clustering, cluster cohesion, was measured by the sum of squared errors (SSE) and silhouette coefficient (SC) using Python scripts with the subsequent procedure.

The calculation of the SSE requires identification of the cluster centers, which is usually the mean of cluster members. As a practical alternative, the centroids of clusters were used here instead, given the following equation:

$$SSE = \sum_{C} \sum_{i \in C} distance^2(i, centroid(C)) \qquad [6]$$

In which the distance refers to structure distance metric SD when calculating SD SSE, and peptide binding specificity distance PD when calculating PD SSE. Given the same number of clusters, a smaller SSE value suggests better clustering because members of each cluster are more homogeneous. Usually, as the number of clusters increases, the SSE decreases monotonically. If all clusters are homogeneous, or the number of clusters equals to the number of samples, the SSE reaches its minimal value of 0.

SC compares the intra-cluster variation to the distances to the neighboring clusters. For each allele i that belongs to cluster C, the average distance to all other alleles in the same cluster is:

$$a(i) = \frac{1}{size(C) - 1} \sum_{i,j \in C, i \neq j} distance(i, j) \quad [7]$$

If cluster C contains only 1 allele, then a(i) is set to zero to avoid a divide-by-zero error. Next, the average distance to the closest neighboring cluster for each allele i is defined as:

$$b(i) = \min_{C \neq D} \frac{1}{size(D)} \sum_{k \in D} distance(i, k) \quad [8]$$

where k is the member of neighboring cluster D. The silhouette coefficient for the clustering result regarding all alleles is calculated as:

$$SC = \frac{1}{N} \sum_{i}^{N} \frac{b(i) - a(a)}{max(a(i), b(i))} \quad [9]$$

where N is number of alleles included in clustering. The range of the silhouette coefficient is [-1, 1], and a high silhouette coefficient indicates proper clustering.

5.2.11 Hierarchical clustering of the reference panel

Alleles in the reference panel were clustered locus-wise by the method described in section 3.1, as well as three other methods for comparison. A second method, “direct clustering”, also uses complete linkage hierarchical clustering but was based on the peptide binding specificity distance (PD). Thirdly, we reproduced the classifications from two previous studies: Ref [25] proposed 6 HLA-A and 6 HLA-B supertypes, and ref [28] proposed 3 HLA-A, 3 HLA-B, and 2 HLA-C supertypes. Finally, the “random even split” method provides the baseline performance, randomly clustering alleles into any given number of clusters of equal size. The performance of each method was assessed by the PD SSE of its clustering result. Because the SSE is dependent on the number of clusters (N), methods including the structure clustering, direct clustering, and random even split were performed on a series of N clusters to show the trend.

5.2.12 Hierarchical clustering of 449 populated HLA alleles

All 449 populated HLA class I alleles were hierarchically clustered across all alleles simultaneously for two purposes. First, possible inter-locus functional overlapping is investigated. Also, such an approach ensures that supertypes are clustered with the same degree of functional similarity for all alleles rather than vary locus to locus. The SD matrix of populated alleles was

calculated, based on which clustering was performed as described in section 3.1. The optimal number of clusters was determined by analyzing the elbow plot and silhouette plot, which show the SSE and SC based on SD as a function of the number of clusters (N), respectively. With stepwise increasing of N, clustering was performed, and the SSE and SC were calculated for the resulting clusters. The optimal values of N are indicated by elbow points and SC peaks. The corresponding dendrogram was generated with the SciPy dendrogram function and visualized with the Interactive Tree Of Life (iTOL) web server [68].

5.2.13 Cluster stability estimated by bootstrapping

One major issue with hierarchical clustering is unsatisfying stability against independent resampling, which is a type of robustness measurement [69-71]. The stability of the hierarchical clustering result was calculated here using a bootstrapping strategy. The HLA alleles were randomly sampled 100 times with replacement, then the SD-clustering was performed on the 100 bootstrapped samples, with the same setting as the original sample. The correspondence between bootstrapped cluster A and corresponding original cluster B was calculated using the Jaccard index [72], which is defined as:

$$J(A,B) = \frac{A \cap B}{A \cup B} \qquad [10]$$

The stability of each cluster was then calculated as the mean Jaccard index of that cluster among 100 bootstrapped clustering results.

5.3 Results

5.3.1 HLA class I structural modeling

Models of 451 alleles were built using ColabFold, then relaxed using Rosetta FastRelax. The models were of high confidence, as indicated by their average pLDDT score of 96.2, and a minimum score of 94.0, with values >90 indicating high accuracy [137]. The quality of the models was further tested by comparing to the crystal structures of the same allele. The RMSDs, calculated with respect to the corresponding crystal structures or centroid structures, of the models in the model quality evaluation set were compared to the natural structural variations measured by RMSDs between crystal structures and centroid structures (Fig. 5-1). The all-atom RMSDs of the models (mean = 1.26 Å) are slightly larger than those of the crystal structures (mean = 0.85 Å).

However, only one model (B*08:01, RMSD = 1.97 Å) exceeds the maximum RMSD of crystal structures (PDB id 4QRP, RMSD = 1.74 Å, compared to the centroid structure 4QRS) (Fig. 5-1A). The backbone RMSDs of the two groups are smaller than the all-atom RMSDs with the average 0.62 Å for models and 0.41 Å for the crystal structures (Fig. 5-1B), showing that the highest inaccuracy is in the sidechain positioning. The distribution of coarse grained RMSDs differs from the all-atom RMSDs (Fig. 5-1C), while the average (1.35 Å for models and 0.81 Å for crystal structures) is similar to the all-atom structures, showing that coarse graining has little influence on model quality. We also examined the model of B*08:01 that performed poorly in all three RMSD tests. Here, the major differences between the model and crystal structure occur in the loop regions between the beta strands, which is distant from the binding groove, and thus has only a minor impact on the clustering results, in line with the observation that the loop regions show flexibility in crystal structures as indicated, for example, by high B-factor values [19].

5.3.2 Structure distance metric *SD* compared to peptide binding specificity distance *PD*

The *SD* between alleles in the reference panel were compared to their *PD*. Both distances were calculated locus-wise and then normalized to [0, 1], resulting in three pairs of intra-locus distance matrices (Fig. 5-2). It is evident that the *PD* and *SD* heatmaps display similar patterns, suggesting that the two distances are in general agreement. This finding was further examined by linear regression between *SD* and *PD*. Among the three loci, HLA-B shows the highest coefficient (R=0.79), followed by HLA-C (R=0.75) and then HLA-A (R=0.73), which confirms the strong correlation between the two distances and supports the hypothesis that peptide binding specificity of HLA class I molecule is mainly determined by the structural landscape of the binding groove. We also calculated the correlation coefficients between PD and the pseudo-sequence (contact residues applied in NetMHCpan [94]) distances based on BLOSUM62 substitution matrix [138], the results are 0.69, 0.69, and 0.59 in HLA-A, B, and C, respectively, which is significantly inferior than the SD (Fig. 5-8).

5.3.3 Performance of reference panel clustering

The reference panel alleles were clustered locus-wise by four methods: 1) *SD*-clustering, as described in Methods section 3.1; 2) hierarchical clustering based on pairwise *PD* (“direct clustering”), which should provide the most accurate clustering amongst available binding data; 3) random even split representing the baseline performance; 4) clustering results from two previous

studies [126, 129]. The performance, i.e., the cluster cohesion, was calculated by the *PD* SSE, because the primary aim of supertype classification is to group alleles with similar peptide binding specificities (Fig. 5-3). In all three loci, the *PD* SSEs of the present clustering method are significantly lower than the random even split. Compared to the direct clustering curve, the *SD*-clustering methods have very close SSE values, which indicates that the clustering based on *SD* is a reasonable approximation to clustering based on *PD*. The consistency between the two distances is also demonstrated by the *SD* SSE and the *PD* SSE curves, which show very similar shapes after scaling.

When compared to previous clustering methods, given the same number of clusters the present method outperforms ref [129] in all three loci, and outperforms ref [126] in HLA-A but not in HLA-B. It is important to note that the HLA-A and HLA-B alleles in the reference panel were derived directly from ref [126], with experimentally validated peptide binding motifs. These results indicate that the present classification method achieves accuracy comparable to previous methods, including classifications based on experimental data.

5.3.4 Supertype and subtype classification of populated HLA class I alleles

Because the three classical HLA I loci - A, B and C - are homologous, it is possible that alleles of different loci have overlapping peptide binding specificities. Some classification methods proposed mixed supertypes that include alleles from different loci [107, 125, 128]. Here, we clustered 449 populated alleles (frequency > 0.01) to investigate the possibility of overlapping peptide binding specificities, and to ensure the clustering of all HLA I alleles achieves the same level of functional similarity among all clusters.

The optimal number of clusters was determined by the elbow plot and silhouette plot method (Fig. 5-4). The most significant elbow point and silhouette coefficient peak appears at N = 5. However, dividing all HLA class I alleles into only five supertypes may hide useful details and thus is not selected. Two optimized numbers of clusters, N=12 and N=20, are suggested by both the elbow points and silhouette coefficient peaks. Accordingly, the 449 populated HLA alleles were clustered into 12 supertypes and 20 subtypes, representing two levels of resolution (Fig. 5-5 and Table 5-2). In this way, the supertypes describe the overall functional similarities, while the subtypes provide more details at enhanced resolution, which provides flexibility in application.

Subtypes are named after the most abundant allotype group in the cluster, following the naming convention, while supertypes were named after the included subtypes. The sizes of the supertypes and subtypes are imbalanced: each supertype contains 1 to 3 subtypes and 9 to 72 alleles, while each subtype contains 4 to 44 alleles, suggesting that the distribution of sampled 449 alleles is uneven in peptide binding specificity space.

Though the supertypes and subtypes were clustered based on structural similarity, they generally agree with allotype groups (the first set of digits of allele nomenclature) that are based on sequence similarity, because alleles that belong to the same group are usually clustered together. This agreement indicates that the allotype groups are reasonable approximations to supertypes and subtypes, as has been done previously [139-141]. As for allotype groups that were separated in multiple clusters, single point mutations in the key residues and subsequent change of side chain positioning explains their split. For example, allele B*15:09 was clustered in subtype B14 rather than B15 where the majority of B*15 allele group is situated. The full protein sequence comparison shows that B*15:09 has 8 mismatches with B*15:01, which belongs to B15, and 10 mismatches with B*14:01, which belongs to B14. However, among the 21 key residues defined in the Methods section 2.2, B*15:09 has 7 mismatches with B*15:01 but 4 mismatches with B*14:01. In addition, the major structural difference between the three alleles is located near the F pocket, the orientation of residue 97 in B*15:01 is different from B*15:01 but similar to B*14:01 (Fig. 5-6).

At both the supertype and subtype level, functional overlap between different loci is not significant, as most supertypes and subtypes are locus-specific, the only exception is subtype C02 that includes two HLA-B alleles, B*46:01 and B*56:03 (Table 5-2). The allele B*46:01 is validated to have close functional relationship with HLA-C [142], which provides circumstantial evidence that supports the present classification.

5.3.5 Stability of present supertype and subtype classification

The stability of supertypes and subtypes is an important concern as more alleles will be included and clustered in future studies. To investigate whether the present supertype and subtype classification is reproducible given different sets of alleles, the stability was evaluated with a bootstrapping strategy. The stability of each supertype and subtype was calculated by its average Jaccard index among 100 repetitions (Table 5-3). In previous studies, the stability of supertypes has rarely been reported, as we found only one study that reported the bootstrap values of 12

proposed HLA-A and HLA-B supertypes, the average of which is 0.54 in the range of [0,1] [118]. The average stability for our classification approach is 0.61 for supertypes and 0.75 for subtypes, respectively. In comparison, then, the present classification shows better stability than previous methods and is expected to be more robust. We also simulated a scenario in which new alleles are included, by adding the two already modeled rare alleles, B*08:02 and B*27:09, the pairwise distance matrix calculation and hierarchical clustering of the 451 alleles were performed the same way as the 449 populated alleles. The two alleles, B*08:02 and B*27:09, were clustered into subtypes B08 and B27 respectively, while the classification of other alleles remains the same, demonstrating the stability of present clustering method and corresponding supertypes and subtypes.

5.4 Discussion

HLA genes are extremely polymorphic, resulting in numerous alleles with diverse peptide binding specificities, thus allowing the human adaptive immune system to respond to a very wide range of antigens. However, this polymorphism also poses a significant difficulty in studies including, but not limited to, organ and cell transplantation, disease association studies, and peptide vaccine development. To reduce the conceptual complexity of the HLA system, the concept of supertypes was introduced to cluster alleles that have similar peptide binding specificities. Although, in the past, a number of experiments have been carried out for supertype classification, these cover only a small portion of binding peptide sequence space. For example, if considering only 8-mer, 9-mer, and 10-mer peptides consisting of the 20 canonical amino acids, the binding peptide sequence has $20^8 + 20^9 + 20^{10} \sim 1.1 \times 10^{13}$ possibilities, while HLA-A*02:01, one of the most studied alleles, has binding assay data for only 71,115 unique peptides in the IEDB [91]. With such a massive gap, supertypes can hardly be determined entirely by experiment.

An alternative to experimental methods is to use predicted peptide-HLA affinities, for example, using deep learning models trained by experimental affinities, and this is perhaps the most direct *in silico* approach. However, the coverage and accuracy of peptide-HLA affinity prediction methods are limited, the leading issue being the unsatisfying availability of training data, as mentioned above. Another issue comes from the peptide binding assays themselves. The most widely used measurement of peptide-HLA affinity is the half maximal inhibitory concentration (IC_{50}), which is defined as the concentration of a query peptide that blocks 50% of standard peptide

binding, but different standard peptides have been used for different HLA alleles [143]. As a result, the binding assay data of different alleles are in different bases and thus not, in principle, comparable. However, they have been combined in the training sets of pan specific predictors, which may lead to inaccuracy.

Another alternative to time-consuming experiments is to cluster HLA alleles based on sequence or structural similarity, which recognizes that the peptide binding specificity is determined by the spatial and chemical properties of the peptide binding groove. Sequence/structure-based methods have advantages compared to affinity prediction-based methods. First, the sequences and structures of HLAs are available with less effort than binding assays; second, the methods may be applicable to all HLA alleles and not just those with available binding affinity data (PD), permitting reliable clustering performance on understudied alleles.

Both sequence-based and structure-based approaches have been successfully applied in general protein binding pocket comparison [144-149]. Here, we have demonstrated that both the sequence and structure distances we use for class I HLA are highly correlated with peptide binding specificity distances (*PD*). Furthermore, as structure-based methods should, in principle, incorporate more direct functional detail than simple sequences, and here, indeed our structure-based method outperforms sequence-based methods in that structure distance has higher correlation coefficient with *PD* than sequence distance.

One limitation of structure-based methods, including the *SD*-clustering, is the imperfect quality of HLA crystal structures and predicted models [147]. As illustrated above, natural structural variations occur in crystal structures of the same HLA allele, which can be caused by peptide induced conformational change [150-152], chaperones (for instance, tapasin) [153, 154] or crystal packing [155]. To minimize the impact of imperfect HLA models, here relaxation and coarse graining were applied. Coarse graining decreases the number of degrees of freedom, thus reducing the impact of possible inaccuracies in sidechain orientations and improves its robustness. However, coarse graining may also reduce clustering accuracy through the loss of information that would be present in a fully atomistic structure. Clearly, the development of all-atom models shows promise for improving the present method.

The reasons for the performance differences between supertype classification methods compared in the present study are complex. Apart from the limited data quality available to previous studies,

there are differences in the definition of supertypes. All methods agree that supertypes include functionally similar alleles, although they focus on different aspects, including the selectivity of anchor residues of binding peptides (peptide motif) [117, 118, 126], the interaction profile of the binding groove [121, 129, 156], and, as adopted in the present study, peptide binding specificity [127]. Therefore, these supertype classifications should not be used interchangeably.

In conclusion, we have demonstrated that the present *SD*-clustering has improved clustering performance (cohesion) relative to previous methods. This means that the clusters are better separated from each other. Also, the flexibility is improved, in that clustering at different resolutions is incorporated. Finally, the stability is improved, meaning that users can add more user-defined alleles without perturbing the existing classification structure. Currently, because the details of the interactions between HLA and the peptides are not completely known, on well-studied alleles the sequence and structure-based classification methods are less accurate than prediction-based methods. However, structure-based methods have unquestionable advantage in understudied alleles thus broadening the coverage of reliable supertype classification. With advances in understanding of peptide HLA interactions and the further development of structural modeling approaches, structure-based methods are destined to improve further.

Appendix

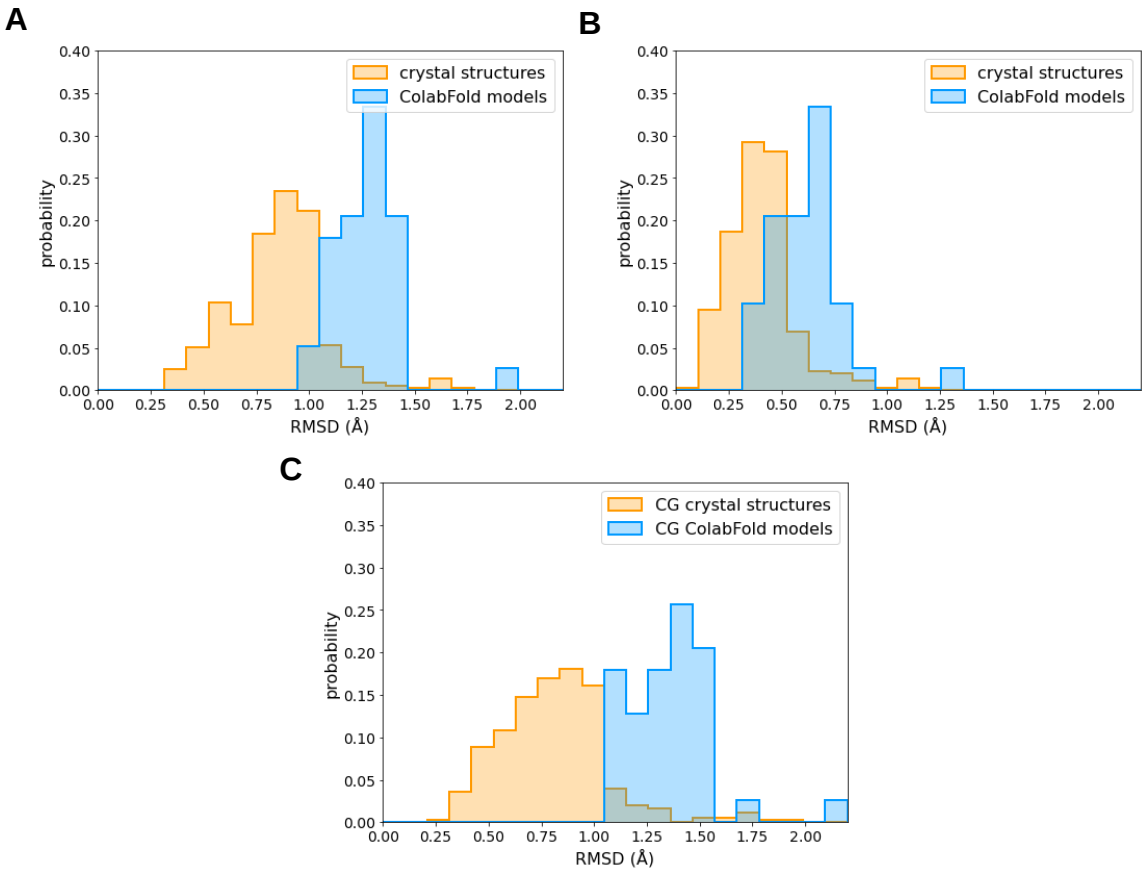


Figure 5-1. Distribution of (A) all-atom, (B) backbone, and (C) coarse-grained RMSDs of crystal structures and ColabFold models compared to the centroid structure of multiple crystal structures (when available) for each allele.

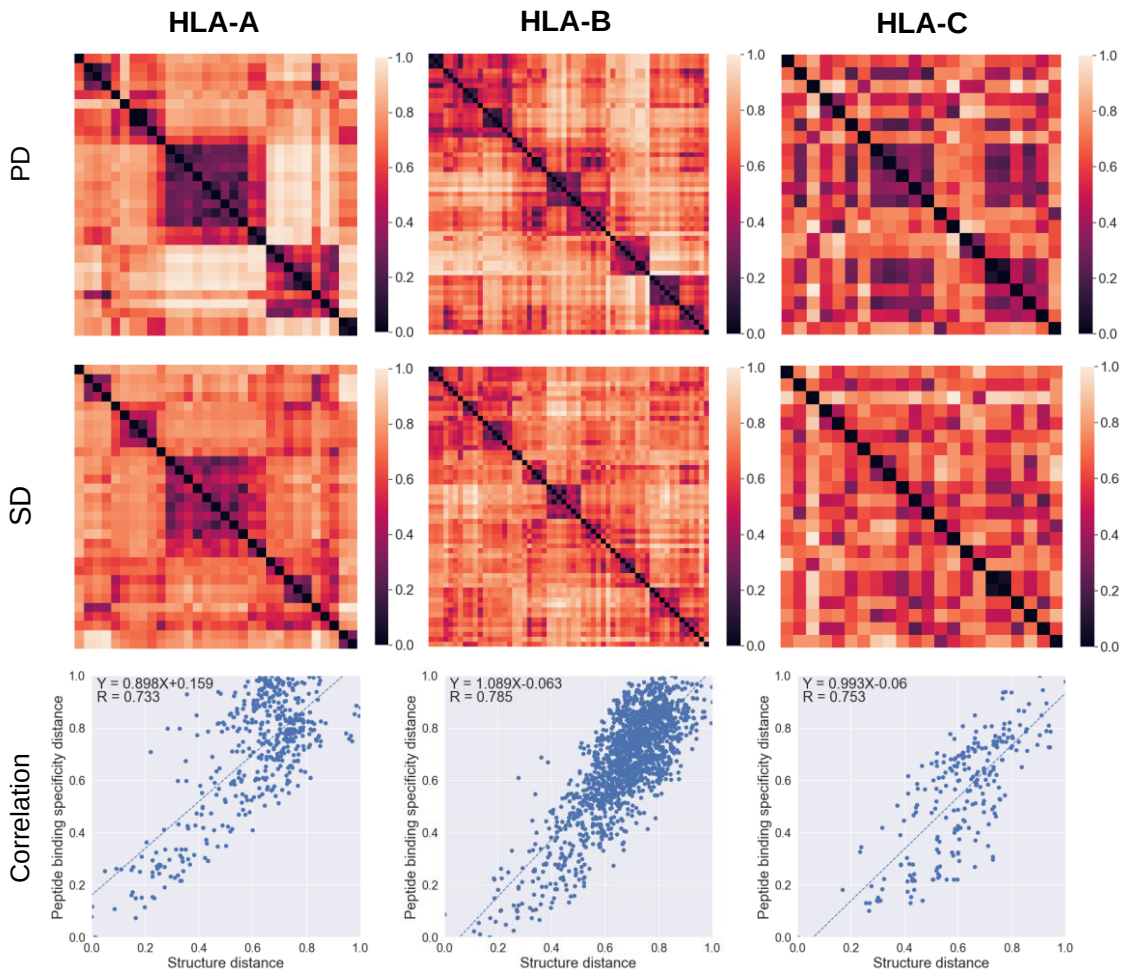


Figure 5-2. Comparison between peptide binding specificity distance *PD* predicted by NetMHCpan and structural distance *SD* defined in the present work. In the correlation plots, the fitted functions between two distances are shown in dotted line. Fitted function and Pearson correlation coefficient (R) are printed out.

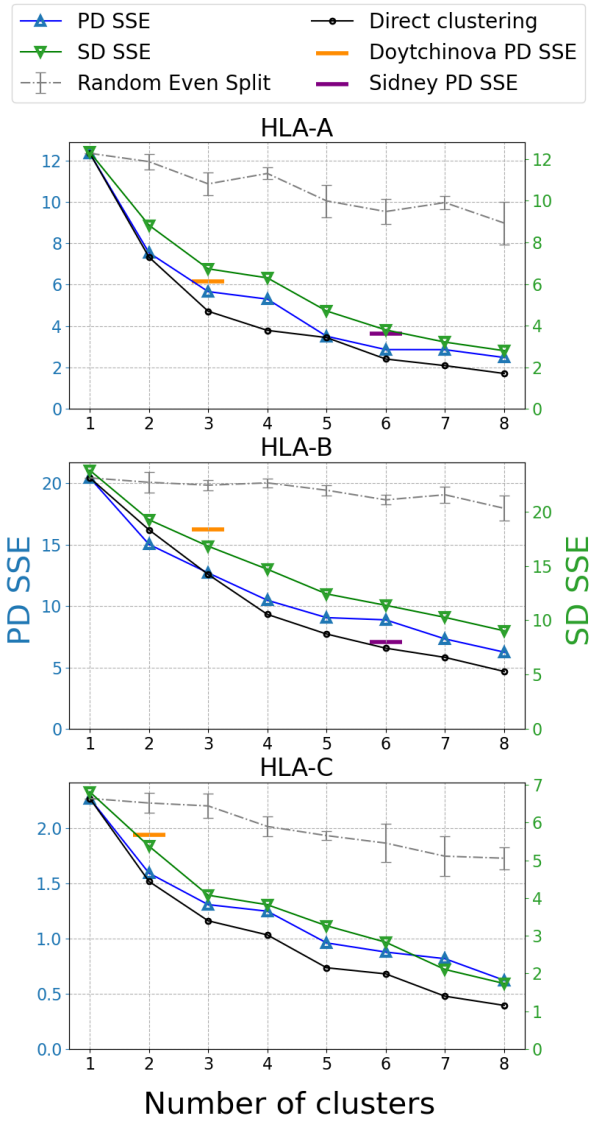


Figure 5-3. Elbow plot showing the sum of squared errors (SSE) of HLA-A, HLA-B and HLA-C reference panel alleles clustering. The SSE of the peptide binding specificity distance (PD SSE) is in blue, and the SSE of the structural distance (SD SSE) is in green. The PD SSE of supertype classifications from Sidney (purple) and Doytchinova (orange) are shown as short horizontal bars.

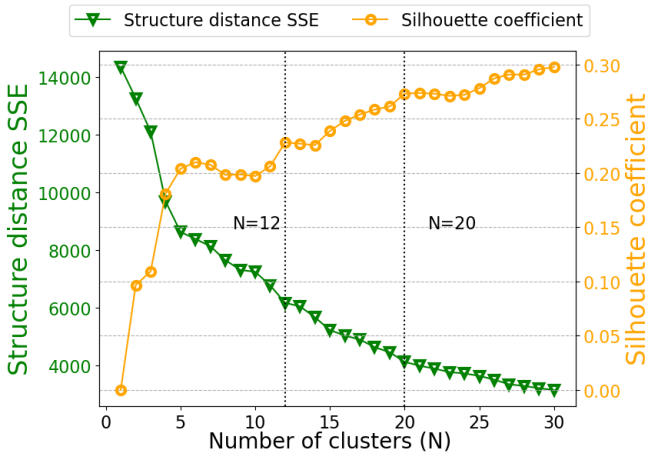


Figure 5-4. Elbow plot and silhouette plot, the sum of squared errors (SSE) and silhouette coefficient (SC) based on *SD* with respect to the number of clusters. Two elbow points and silhouette peaks are shown as vertical dotted lines.

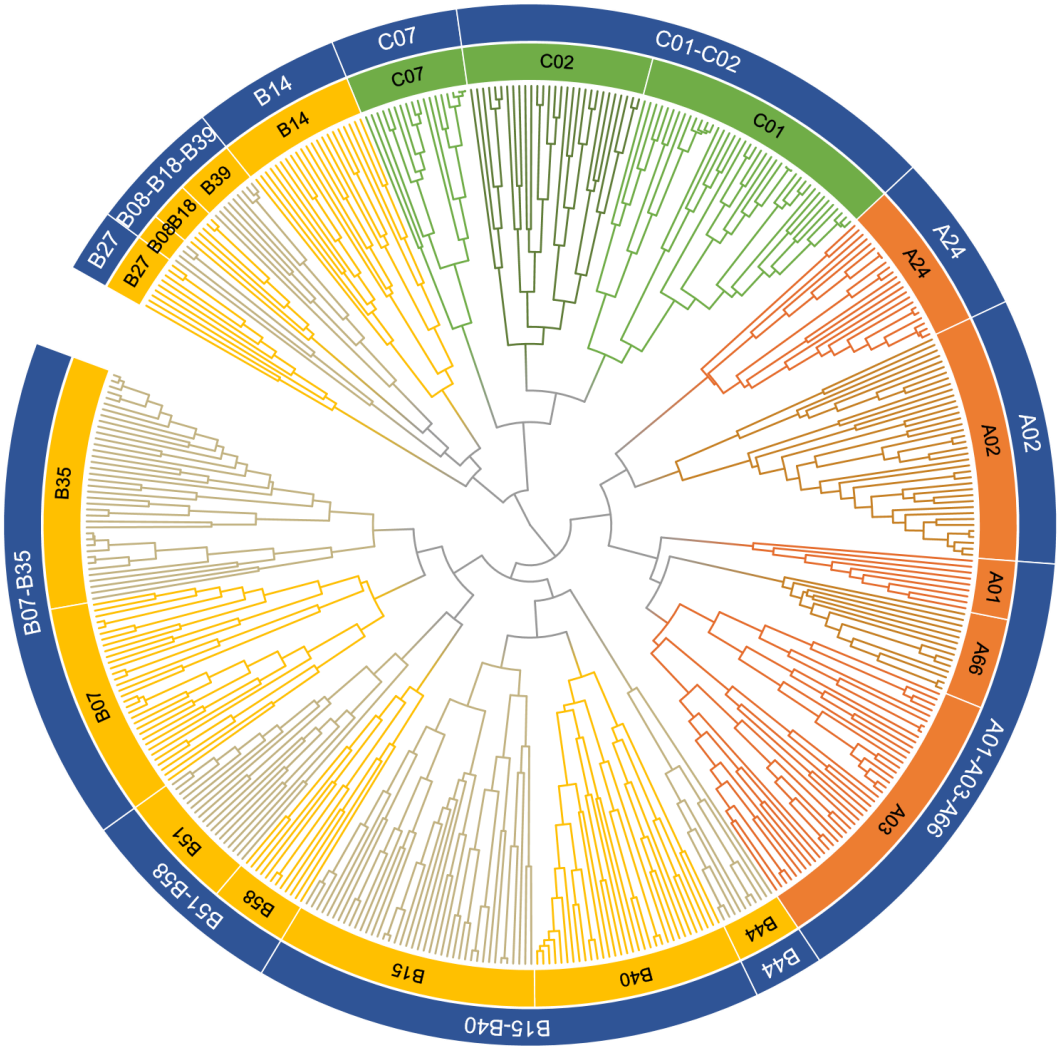


Figure 5-5. Dendrogram of populated HLA supertype hierarchical clustering. Clades of different supertypes are shown in different colors.

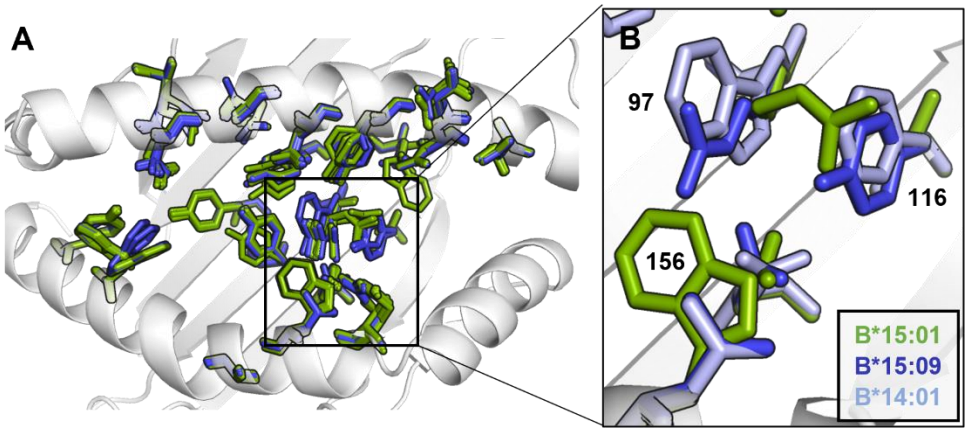


Figure 5-6. Comparing structures of HLA-B*15 alleles that were clustered in subtype B14 and B15. (**A**) Top-down view of the peptide binding groove. Alleles that were clustered in subtype B14 are colored in blue while and B15 alleles are green. The major difference in orientation of key residues is located near F pocket. (**B**) Zoom-in near the F pocket, showing the difference of B*15:09, B*15:01, and B*14:01 in residue 97, 116, and 156. Comparing to B*15:09 and B*14:01, B*15:01 has Trp in position 156 which is bulk in size and pushed Arg 97 towards Ser 116, resulting in the difference in the shape of F pocket.

Table 5-1 Contact residue positions in HLA class I proteins and corresponding weight factor. Positions are listed in the descending order of weight.

Position[a] (*i*)	Overlap of binding specificity[b] (%)	Difference in binding specificity[c] (%)	Weight (w_i)
63	55.4	44.6	9.9
67	65.7	34.3	7.6
116	73.9	26.1	5.8
9	75.6	24.4	5.4
97	78.7	21.3	4.7
152	79.4	20.6	4.6
167	82.8	17.2	3.8
156	83.2	16.8	3.7
74	83.9	16.1	3.6
70	85.6	14.4	3.2
80	86.3	13.7	3.0
171	84.0	13.0	2.9
45	87.3	12.7	2.8
66	88.0	12.0	2.7
77	88.0	12.0	2.7
76	89.4	10.6	2.4
114	89.7	10.3	2.3
99	90.4	9.6	2.1
163	93.1	6.9	1.5
95	93.1	6.9	1.5
59	93.5	6.5	1.4
158	93.8	6.2	1.4
69	94.5	5.5	1.2
24	94.8	5.2	1.2
7	95.5	4.5	1.0

[a] There are no insertions or deletions in the binding domain of most existing HLA I alleles and all alleles considered in this study. Thus, residue positions are constant.

[b] The overlap of binding specificity between altered HLA with single point mutation and wild type.

[c] The difference was calculated as (1 - overlap) %.

Table 5-2. HLA supertype and subtype assignment.

A01-A03-A66											
A01		A03							A66		
A*01:01	A*36:01	A*03:01	A*11:04	A*29:03	A*30:03	A*31:04	A*32:04	A*74:02	A*25:01	A*26:14	
A*01:02	A*80:01	A*03:02	A*11:05	A*29:10	A*30:04	A*31:08	A*32:106	A*74:03	A*26:01	A*26:16	
A*01:03		A*03:08	A*11:06	A*29:11	A*30:10	A*31:09	A*33:01		A*26:02	A*34:01	
A*01:06		A*03:27	A*11:12	A*29:25	A*30:15	A*31:12	A*33:03		A*26:03	A*43:01	
A*01:17		A*11:01	A*11:170	A*29:50	A*31:01	A*31:29	A*33:18		A*26:08	A*66:01	
A*01:23		A*11:02	A*29:01	A*30:01	A*31:02	A*32:01	A*34:02		A*26:12	A*66:02	
A*01:36		A*11:03	A*29:02	A*30:02	A*31:03	A*32:02	A*74:01		A*26:13	A*66:03	

A02										
A02										
A*02:01	A*02:05	A*02:09	A*02:13	A*02:19	A*02:24	A*02:44	A*02:52	A*68:03	A*68:30	
A*02:02	A*02:06	A*02:10	A*02:14	A*02:197	A*02:240	A*02:46	A*02:61	A*68:05	A*69:01	
A*02:03	A*02:07	A*02:11	A*02:16	A*02:20	A*02:34	A*02:48	A*68:01	A*68:06		
A*02:04	A*02:08	A*02:12	A*02:17	A*02:22	A*02:36	A*02:50	A*68:02	A*68:16		

A24											
A24											
A*23:01	A*23:17	A*24:03	A*24:05	A*24:07	A*24:10	A*24:14	A*24:17	A*24:23	A*24:242	A*24:28	A*24:51
A*23:05	A*24:02	A*24:04	A*24:06	A*24:08	A*24:13	A*24:143	A*24:20	A*24:24	A*24:25	A*24:41	

B07-B35											
B07					B35						
B*07:02	B*07:35	B*42:02	B*55:04	B*67:01	B*15:08	B*35:04	B*35:13	B*35:21	B*35:36	B*40:08	
B*07:03	B*07:75	B*42:18	B*55:15	B*78:01	B*15:11	B*35:05	B*35:14	B*35:23	B*35:43	B*53:01	
B*07:05	B*07:91	B*54:01	B*56:01	B*81:01	B*15:29	B*35:06	B*35:15	B*35:24	B*35:44	B*53:03	
B*07:06	B*07:96	B*54:18	B*56:02	B*82:01	B*18:07	B*35:08	B*35:16	B*35:25	B*35:61	B*53:05	
B*07:07	B*39:10	B*55:01	B*56:04	B*82:02	B*35:01	B*35:09	B*35:17	B*35:27	B*35:68		
B*07:08	B*39:20	B*55:02	B*56:05		B*35:02	B*35:10	B*35:18	B*35:29	B*35:77		
B*07:105	B*42:01	B*55:03	B*56:43		B*35:03	B*35:12	B*35:19	B*35:32	B*40:07		

B51-B58										
B51						B58				
B*51:01	B*51:04	B*51:07	B*51:12	B*51:19	B*51:76	B*15:16	B*57:01	B*57:04	B*58:01	
B*51:02	B*51:05	B*51:08	B*51:14	B*51:33	B*52:01	B*15:17	B*57:02	B*57:05	B*58:02	
B*51:03	B*51:06	B*51:10	B*51:15	B*51:61	B*59:01	B*15:67	B*57:03	B*57:25	B*58:06	

B08-B18-B39										
B08		B18			B39					
B*08:01	B*08:04	B*18:01	B*18:03	B*18:06	B*39:01	B*39:04	B*39:09	B*39:24	B*73:01	
B*08:03	B*08:05	B*18:02	B*18:05	B*18:09	B*39:03	B*39:06	B*39:12	B*39:54		

B14										
B14										
B*14:01	B*14:03	B*14:05	B*14:11	B*15:10	B*15:21	B*15:37	B*38:01	B*38:09	B*39:07	B*78:03
B*14:02	B*14:04	B*14:06	B*15:09	B*15:18	B*15:23	B*15:93	B*38:02	B*39:05	B*39:11	

B15-B40										
B15						B40				
B*13:01	B*15:04	B*15:20	B*15:35	B*40:02	B*47:03	B*15:30	B*40:05	B*40:16	B*41:03	B*49:01
B*13:02	B*15:05	B*15:24	B*15:36	B*40:03	B*48:01	B*15:58	B*40:10	B*40:23	B*41:23	B*50:01
B*13:04	B*15:06	B*15:25	B*15:39	B*40:06	B*48:02	B*37:01	B*40:11	B*40:36	B*44:05	B*50:02
B*13:38	B*15:07	B*15:27	B*15:46	B*40:09	B*48:04	B*39:02	B*40:12	B*40:46	B*44:10	B*50:04
B*15:01	B*15:12	B*15:31	B*18:04	B*40:268	B*48:07	B*39:08	B*40:121	B*40:49	B*44:15	
B*15:02	B*15:13	B*15:32	B*35:20	B*40:38		B*40:01	B*40:14	B*41:01	B*45:01	
B*15:03	B*15:15	B*15:34	B*35:28	B*47:01		B*40:04	B*40:15	B*41:02	B*48:03	

B27								
B27								
B*27:01	B*27:02	B*27:03	B*27:04	B*27:05	B*27:06	B*27:07	B*27:08	B*27:14

B44									
B44									
B*44:02	B*44:03	B*44:04	B*44:06	B*44:07	B*44:08	B*44:09	B*44:151	B*44:27	B*44:29

C01-C02											
C01							C02				
C*01:02	C*03:04	C*03:19	C*04:10	C*08:02	C*15:02	C*15:13	B*46:01	C*02:10	C*06:04	C*12:05	C*15:04
C*01:03	C*03:05	C*04:01	C*04:15	C*08:03	C*15:03	C*17:01	B*56:03	C*03:02	C*06:06	C*12:07	C*15:09
C*01:06	C*03:06	C*04:03	C*04:43	C*08:04	C*15:05	C*17:03	C*01:04	C*03:15	C*07:26	C*14:02	C*16:01
C*01:44	C*03:07	C*04:04	C*05:01	C*08:06	C*15:07	C*18:01	C*02:02	C*03:16	C*12:02	C*14:03	C*16:02
C*01:57	C*03:09	C*04:06	C*05:09	C*08:13	C*15:08	C*18:02	C*02:03	C*06:02	C*12:03	C*14:04	C*16:04
C*03:03	C*03:135	C*04:07	C*08:01	C*08:72	C*15:10		C*02:09	C*06:03	C*12:04	C*14:06	

C07									
C07									
C*07:01	C*07:03	C*07:05	C*07:07	C*07:10	C*07:17	C*07:18	C*07:248	C*07:270	
C*07:02	C*07:04	C*07:06	C*07:08	C*07:14	C*07:172	C*07:19	C*07:27	C*07:31	

Table 5-3. Supertype and subtype stability measured by average Jaccard index in bootstrapping.

Supertype	Avg. Jaccard index	Subtype	Avg. Jaccard index
A01-A03-A66	0.65	A01	0.72
		A03	0.70
		A66	0.68
A02	0.58	A02	0.83
A24	0.66	A24	0.89
B07-B35	0.60	B07	0.70
		B35	0.70
B51-B58	0.74	B51	0.91
		B58	0.94
B08-B18-B39	0.25	B08	0.66
		B18	0.66
		B39	0.54
B14	0.68	B14	0.74
B15-B40	0.62	B15	0.54
		B40	0.65
B27	0.68	B27	0.96
B44	0.38	B44	0.64
C01-C02	0.87	C01	0.87
		C02	0.76
C07	0.66	C07	0.93

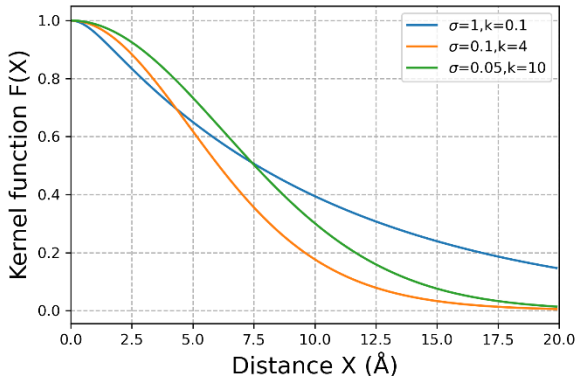


Figure 5-7. Curve of kernel functions with difference set of shape parameters, showing the bell-shape and contribution of the two parameters.

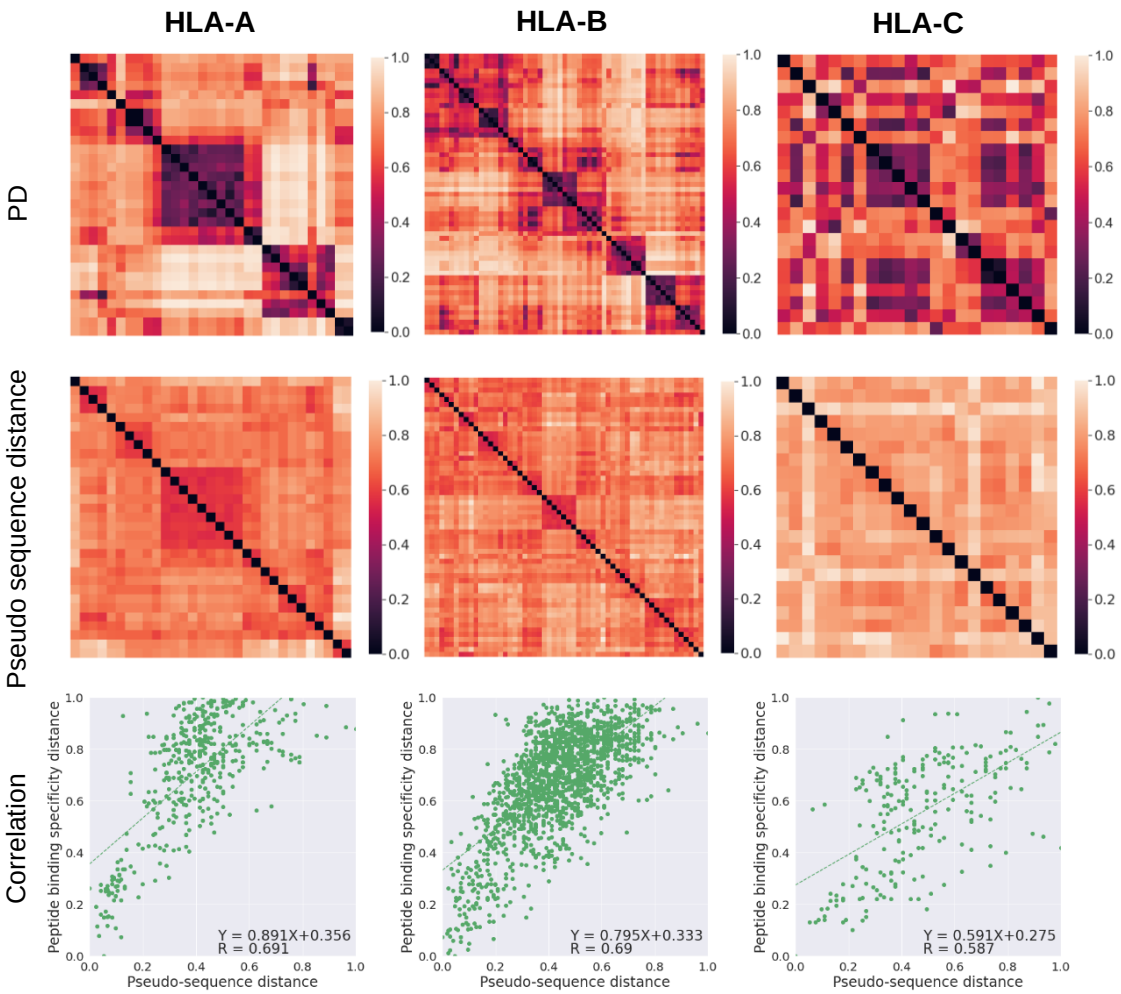


Figure 5-8. Comparison between peptide binding specificity distance PD predicted by NetMHCpan and pseudo sequence distance defined by Reynisson, B., et al. In the correlation plots, the fitted functions between two distances are shown in dotted line. Fitted function and Pearson correlation coefficient (R) are printed out.

Table 5-4. List of populated HLA class I alleles. Alleles that have maximum frequency >0.01 among populations with sample size larger than 50 were selected. Data collected from the Allele Frequency Net Database (AFND) in February 2022.

Allele	Max Frequency	Allele	Max Frequency	Allele	Max Frequency	Allele	Max Frequency	Allele	Max Frequency	Allele	Max Frequency
A*24:02	0.863	A*02:34	0.039	B*39:06	0.239	B*35:25	0.06	B*14:11	0.0192	C*15:02	0.232
A*02:01	0.733	A*31:12	0.0385	B*51:01	0.232	B*39:11	0.059	B*35:44	0.0192	C*05:01	0.228
A*11:01	0.636	A*80:01	0.0382	B*52:01	0.231	B*47:01	0.059	B*40:23	0.019	C*04:04	0.2
A*02:12	0.6	A*23:05	0.034	B*27:04	0.225	B*27:02	0.0579	B*51:12	0.019	C*01:57	0.1996
A*01:01	0.48	A*31:02	0.032	B*44:02	0.216	B*56:03	0.057	B*56:05	0.019	C*05:09	0.179
A*03:01	0.444	A*29:11	0.0317	B*53:01	0.208	B*51:08	0.0558	B*82:01	0.019	C*12:03	0.174
A*34:01	0.44	A*11:05	0.03	B*27:05	0.2	B*40:14	0.055	B*48:04	0.0188	C*12:02	0.156
A*68:02	0.372	A*68:05	0.0275	B*40:06	0.19	B*56:43	0.0544	B*35:61	0.0183	C*17:01	0.151
A*31:01	0.36	A*02:240	0.0272	B*35:08	0.1875	B*55:01	0.052	B*44:151	0.0183	C*07:17	0.147
A*68:01	0.3538	A*26:12	0.026	B*49:01	0.1867	B*39:54	0.0513	B*35:77	0.018	C*03:02	0.138
A*02:06	0.349	A*31:09	0.0251	B*48:03	0.181	B*14:04	0.05	B*51:14	0.018	C*08:02	0.132
A*02:04	0.314	A*02:09	0.025	B*50:01	0.173	B*40:12	0.05	B*67:01	0.018	C*15:03	0.131
A*23:01	0.228	A*24:25	0.025	B*14:05	0.17	B*07:75	0.0477	B*35:09	0.017	C*14:03	0.122
A*02:07	0.227	A*02:13	0.024	B*58:01	0.17	B*07:06	0.043	B*44:15	0.017	C*07:06	0.12
A*26:01	0.218	A*31:04	0.024	B*40:05	0.1674	B*39:12	0.043	B*57:05	0.017	C*08:06	0.114
A*24:07	0.207	A*68:16	0.0238	B*38:01	0.1672	B*44:07	0.043	B*07:07	0.016	C*07:03	0.112
A*03:02	0.1931	A*26:02	0.023	B*18:01	0.167	B*15:08	0.0417	B*13:04	0.016	C*07:04	0.112
A*33:03	0.188	A*29:10	0.023	B*13:02	0.16	B*18:02	0.041	B*15:24	0.016	C*08:03	0.1
A*02:02	0.1855	A*11:170	0.0215	B*35:14	0.16	B*56:04	0.0406	B*35:13	0.016	C*14:02	0.1
A*02:03	0.176	A*02:46	0.021	B*40:10	0.158	B*08:04	0.04	B*40:46	0.016	C*15:07	0.098
A*02:11	0.16	A*24:143	0.0202	B*44:03	0.1574	B*08:05	0.04	B*07:35	0.0158	C*02:10	0.093
A*30:01	0.16	A*02:10	0.02	B*58:02	0.143	B*15:31	0.04	B*50:04	0.0157	C*04:43	0.0821
A*02:22	0.15	A*02:14	0.02	B*42:01	0.138	B*27:06	0.04	B*40:49	0.0155	C*03:135	0.0799
A*30:02	0.1474	A*32:04	0.02	B*15:21	0.135	B*40:03	0.04	B*57:04	0.0153	C*15:04	0.075
A*68:03	0.1376	A*33:18	0.0192	B*51:02	0.135	B*07:03	0.039	B*07:08	0.015	C*15:05	0.0718
A*31:08	0.13	A*68:06	0.0192	B*15:13	0.133	B*14:03	0.039	B*14:06	0.015	C*12:04	0.07
A*11:02	0.127	A*26:08	0.019	B*39:03	0.132	B*35:23	0.039	B*18:06	0.015	C*02:09	0.069
A*33:01	0.122	A*02:16	0.0183	B*35:06	0.129	B*15:16	0.0377	B*35:29	0.015	C*16:04	0.065
A*29:01	0.1202	A*24:13	0.018	B*15:11	0.1287	B*40:121	0.0376	B*40:16	0.015	C*17:03	0.063
A*01:03	0.1185	A*02:48	0.017	B*38:02	0.1272	B*15:15	0.0375	B*44:08	0.015	C*03:06	0.06
A*24:20	0.116	A*02:61	0.017	B*15:07	0.125	B*39:07	0.0372	B*51:04	0.015	C*06:03	0.06
A*32:01	0.114	A*11:04	0.017	B*07:05	0.1217	B*42:18	0.0371	B*58:06	0.015	C*15:08	0.06
A*29:02	0.11	A*30:10	0.017	B*51:10	0.12	B*35:20	0.037	B*35:36	0.0145	C*15:09	0.059
A*24:03	0.1	A*01:23	0.016	B*54:01	0.12	B*15:30	0.036	B*15:23	0.014	C*16:02	0.0577
A*24:17	0.0979	A*26:13	0.016	B*15:32	0.118	B*44:05	0.036	B*15:37	0.014	C*02:03	0.055
A*02:19	0.095	A*24:242	0.0157	B*35:03	0.111	B*18:03	0.035	B*40:15	0.014	C*04:07	0.054
A*74:01	0.095	A*02:20	0.015	B*18:07	0.11	B*35:21	0.035	B*44:27	0.014	C*07:18	0.051
A*02:05	0.087	A*32:106	0.0141	B*45:01	0.11	B*18:09	0.033	B*53:05	0.014	C*08:72	0.0506
A*02:44	0.086	A*02:08	0.013	B*15:03	0.108	B*27:03	0.032	B*55:03	0.0133	C*07:08	0.05
A*11:12	0.086	A*01:17	0.0126	B*14:02	0.1054	B*51:61	0.0319	B*55:15	0.0133	C*18:01	0.05
A*26:03	0.086	A*01:36	0.0126	B*44:06	0.105	B*13:38	0.0315	B*15:29	0.0132	C*18:02	0.05
A*24:04	0.081	A*30:15	0.0126	B*40:04	0.103	B*41:03	0.031	B*38:09	0.0131	C*03:16	0.046
A*01:06	0.08	A*29:03	0.0124	B*50:02	0.1022	B*47:03	0.031	B*15:20	0.013	C*07:31	0.0313
A*02:17	0.08	A*02:36	0.012	B*57:01	0.102	B*48:07	0.031	B*07:105	0.0126	C*01:03	0.031
A*24:05	0.08	A*30:03	0.011	B*44:04	0.1	B*15:09	0.03	B*15:34	0.012	C*07:19	0.031
A*30:04	0.08	A*74:02	0.011	B*35:17	0.099	B*44:09	0.03	B*27:08	0.012	C*07:27	0.031
A*25:01	0.078	A*03:08	0.01	B*15:18	0.0919	B*44:29	0.03	B*39:04	0.012	C*03:07	0.03
A*31:29	0.071	A*24:24	0.01	B*15:10	0.0916	B*48:02	0.03	B*40:07	0.012	C*14:04	0.03
A*02:50	0.07	A*24:28	0.01	B*14:01	0.09	B*55:04	0.03	B*15:58	0.011	C*14:06	0.03
A*36:01	0.07	A*26:14	0.01	B*15:17	0.09	B*39:10	0.029	B*39:20	0.011	C*15:13	0.0288
A*68:30	0.069	A*26:16	0.01	B*37:01	0.09	B*35:68	0.0288	B*57:25	0.0108	C*08:04	0.0282
A*66:01	0.068	A*32:02	0.01	B*53:03	0.09	B*51:76	0.0281	B*41:23	0.0107	C*07:05	0.023
A*02:52	0.067	A*66:03	0.01	B*35:02	0.0894	B*15:46	0.028	B*27:14	0.0104	C*03:09	0.02
A*03:27	0.0608	B*39:01	0.549	B*41:01	0.0863	B*40:268	0.0274	B*15:67	0.01	C*07:14	0.02
A*24:23	0.056	B*35:05	0.515	B*51:06	0.0862	B*39:08	0.027	B*35:10	0.01	C*01:06	0.0192
A*24:14	0.055	B*15:01	0.4007	B*35:27	0.083	B*15:39	0.0263	B*35:15	0.01	C*04:15	0.0192
A*31:03	0.055	B*15:25	0.4	B*39:24	0.082	B*07:91	0.0261	B*35:28	0.01	C*08:13	0.019
A*34:02	0.055	B*40:02	0.398	B*15:93	0.0785	B*35:18	0.026	B*40:36	0.01	C*03:19	0.018
A*24:41	0.0548	B*39:02	0.387	B*15:35	0.0783	B*15:05	0.0257	B*40:38	0.01	C*07:172	0.0164
A*24:10	0.054	B*35:43	0.3846	B*15:04	0.078	B*15:27	0.0257	B*51:03	0.01	C*01:04	0.016
A*24:51	0.0522	B*39:05	0.361	B*40:09	0.078	B*27:07	0.025	B*51:15	0.01	C*04:10	0.015
A*02:197	0.0521	B*15:02	0.358	B*54:18	0.0744	B*51:33	0.025	B*51:19	0.01	C*07:10	0.015
A*11:03	0.051	B*40:01	0.355	B*44:10	0.074	B*82:02	0.025	B*78:03	0.01	C*07:270	0.0145

Table 5-4. Continued.

Allele	Max Frequency	Allele	Max Frequency	Allele	Max Frequency	Allele	Max Frequency	Allele	Max Frequency	Allele	Max Frequency
A*11:06	0.051	B*39:09	0.349	B*15:36	0.073	B*08:03	0.023	C*07:02	0.714	C*06:06	0.014
A*02:24	0.05	B*56:01	0.346	B*18:04	0.073	B*27:01	0.023	C*03:04	0.52	C*04:06	0.0123
A*24:08	0.05	B*08:01	0.3	B*35:19	0.073	B*51:07	0.023	C*01:02	0.5096	C*15:10	0.0122
A*66:02	0.05	B*13:01	0.283	B*41:02	0.0694	B*73:01	0.023	C*04:03	0.459	C*07:07	0.012
A*74:03	0.05	B*15:06	0.273	B*57:03	0.069	B*42:02	0.022	C*07:01	0.424	C*06:04	0.011
A*43:01	0.0493	B*48:01	0.26	B*78:01	0.069	B*35:32	0.021	C*04:01	0.378	C*12:05	0.011
A*29:25	0.0485	B*07:02	0.257	B*40:08	0.0665	B*15:12	0.0207	C*08:01	0.366	C*12:07	0.011
A*24:06	0.048	B*55:02	0.257	B*57:02	0.0664	B*18:05	0.02	C*03:05	0.361	C*07:248	0.0105
A*29:50	0.0477	B*46:01	0.254	B*59:01	0.063	B*35:16	0.02	C*16:01	0.283	C*01:44	0.01
A*01:02	0.047	B*35:01	0.252	B*81:01	0.062	B*35:24	0.02	C*03:03	0.281	C*03:15	0.01
A*23:17	0.0441	B*35:12	0.24	B*35:04	0.0615	B*40:11	0.02	C*02:02	0.244	C*07:26	0.01
A*69:01	0.04	B*56:02	0.24	B*51:05	0.061	B*07:96	0.0192	C*06:02	0.24		

Table 5-5. List of HLA I crystal structures collected from the IMGT/3Dstructure-DB in February 2022.

PDB_id	Allele	Resolution (Å)	PDB_id	Allele	Resolution (Å)	PDB_id	Allele	Resolution (Å)	PDB_id	Allele	Resolution (Å)
3bo8	A*01:01	1.80	6apn	A*02:01	2.22	3rl1	A*03:01	2.00	5ib2	B*27:05	1.44
4nqx	A*01:01	2.00	1eey	A*02:01	2.25	3rl2	A*03:01	2.39	5ib1	B*27:05	1.91
1w72	A*01:01	2.15	1eez	A*02:01	2.30	2xpg	A*03:01	2.60	5ib3	B*27:05	1.91
4nqv	A*01:01	2.39	1s9y	A*02:01	2.30	6eny	A*03:01	5.80	5ib4	B*27:05	1.95
6at9	A*01:01	2.95	2x4o	A*02:01	2.30	1x7q	A*11:01	1.45	6pyj	B*27:05	1.44
6j2a	A*30:03	1.40	2x4p	A*02:01	2.30	2hn7	A*11:01	1.60	4g8i	B*27:05	1.60
6j29	A*30:03	1.60	2x4r	A*02:01	2.30	1qvo	A*11:01	2.22	4g9d	B*27:05	1.60
6j1v	A*30:03	2.00	2x4t	A*02:01	2.30	1q94	A*11:01	2.40	3b6s	B*27:05	1.80
6j1w	A*30:01	1.50	3mgo	A*02:01	2.30	4beo	A*11:01	2.43	3bp4	B*27:05	1.85
3mre	A*02:01	1.10	3v5k	A*02:01	2.31	4uq2	A*11:01	2.43	3lv3	B*27:05	1.94
3d25	A*02:01	1.30	2x4n	A*02:01	2.34	4n8v	A*11:01	2.50	2bst	B*27:05	2.10
6trn	A*02:01	1.35	2c7u	A*02:01	2.38	1hsb	A*68:01	1.90	3dtx	B*27:05	2.10
5n1y	A*02:01	1.39	1qr1	A*02:01	2.40	2hla	A*68:01	2.60	1w0v	B*27:05	2.27
1i4f	A*02:01	1.40	1b0g	A*02:01	2.50	5wwj	A*23:01	2.29	4g9f	B*27:05	1.90
6o53	A*02:01	1.40	1hhi	A*02:01	2.50	5wwu	A*23:01	2.79	4g8g	B*27:05	2.40
4u6y	A*02:01	1.47	1hhj	A*02:01	2.50	5wwi	A*23:01	3.19	5deg	B*27:06	1.83
5ddh	A*02:01	1.50	1hhk	A*02:01	2.50	5xov	A*23:01	2.68	1k5n	B*27:09	1.09
6o4z	A*02:01	1.50	1s9x	A*02:01	2.50	4f7t	A*24:02	1.70	3czf	B*27:09	1.20
2gtw	A*02:01	1.55	5hhm	A*02:01	2.50	4f7p	A*24:02	1.90	1uxw	B*27:09	1.71
6o51	A*02:01	1.55	2x4s	A*02:01	2.55	3vxn	A*24:02	1.95	3d18	B*27:09	1.74
6o4y	A*02:01	1.58	1hhg	A*02:01	2.60	5hgd	A*24:02	2.07	3bp7	B*27:09	1.80
2v2w	A*02:01	1.60	3hla	A*02:01	2.60	5hga	A*24:02	2.20	3b3i	B*27:09	1.86
2v2x	A*02:01	1.60	3to2	A*02:01	2.60	5wxc	A*24:02	2.30	1jgd	B*27:09	1.90
2vll	A*02:01	1.60	1akj	A*02:01	2.65	3nfn	A*24:02	2.39	3hcv	B*27:09	1.95
3bgm	A*02:01	1.60	5mep	A*02:01	2.71	5hgh	A*24:02	2.39	1w0w	B*27:09	2.10
3gso	A*02:01	1.60	1i1f	A*02:01	2.80	3i6l	A*24:02	2.40	1of2	B*27:09	2.20
3pwn	A*02:01	1.60	1i7t	A*02:01	2.80	4f7m	A*24:02	2.40	4o2c	B*39:01	1.80
3v5h	A*02:01	1.63	3i6k	A*02:01	2.80	5hgb	A*24:02	2.40	4o2f	B*39:01	1.90
3bh8	A*02:01	1.65	1b0r	A*02:01	2.90	3vxp	A*24:02	2.50	4o2e	B*39:01	1.98
3fqn	A*02:01	1.65	3gjg	A*02:01	2.90	3vxo	A*24:02	2.61	6mt3	B*18:01	1.21
3kla	A*02:01	1.65	3hae	A*02:01	2.90	2bck	A*24:02	2.80	4xxc	B*18:01	1.43
3pwl	A*02:01	1.65	5hho	A*02:01	2.95	3qzw	A*24:02	2.80	6mt6	B*37:01	1.31
3utq	A*02:01	1.67	1hhh	A*02:01	3.00	5wxd	A*24:02	3.30	6mt4	B*37:01	1.55
3qfd	A*02:01	1.68	4wuu	A*02:01	3.05	3nfj	A*24:02	2.68	6mt5	B*37:01	1.55
4u6x	A*02:01	1.68	6nca	A*02:01	3.30	3vxm	A*24:02	2.50	6mtm	B*37:01	3.00
2git	A*02:01	1.70	1p7q	A*02:01	3.40	3vxs	A*24:02	1.80	6iex	B*40:01	2.31
2gtz	A*02:01	1.70	1hla	A*02:01	3.50	3vxr	A*24:02	2.40	5iek	B*40:02	1.80
3bh9	A*02:01	1.70	3h9h	A*02:01	2.00	3w0w	A*24:02	2.60	1m6o	B*44:02	1.60
3fqr	A*02:01	1.70	3mrg	A*02:01	1.30	3vxu	A*24:02	2.70	3kpm	B*44:02	1.60
3fqx	A*02:01	1.70	3mrb	A*02:01	1.40	6at5	B*07:02	1.50	3dx6	B*44:02	1.70
3o3d	A*02:01	1.70	3mrk	A*02:01	1.40	6avf	B*07:02	2.03	3l3i	B*44:02	1.70
3pwj	A*02:01	1.70	3mrr	A*02:01	1.60	6avg	B*07:02	2.60	3l3d	B*44:02	1.80
2gt9	A*02:01	1.75	3mrd	A*02:01	1.70	4u1h	B*07:02	1.59	3kpl	B*44:02	1.96
6opd	A*02:01	1.79	3mrc	A*02:01	1.80	3vcl	B*07:02	1.70	3l3g	B*44:02	2.10
1duz	A*02:01	1.80	3mrj	A*02:01	1.87	5eo0	B*07:02	1.70	3l3j	B*44:02	2.40
1i7u	A*02:01	1.80	3mrm	A*02:01	1.90	5eo1	B*07:02	1.85	3l3k	B*44:02	2.60
1tvb	A*02:01	1.80	3mr9	A*02:01	1.93	4u1k	B*07:02	2.09	3dx7	B*44:03	1.60
1tvh	A*02:01	1.80	3mri	A*02:01	2.10	1xh3	B*35:01	1.48	1n2r	B*44:03	1.70
3fqt	A*02:01	1.80	3mrp	A*02:01	2.10	2fyy	B*35:01	1.50	3kpn	B*44:03	2.00
3fqu	A*02:01	1.80	3mrq	A*02:01	2.20	1zhk	B*35:01	1.60	3kpo	B*44:03	2.30
3ft2	A*02:01	1.80	3mrf	A*02:01	2.30	4prn	B*35:01	1.65	1sys	B*44:03	2.40
3gsu	A*02:01	1.80	3mrn	A*02:01	2.30	1zsd	B*35:01	1.70	6d29	B*57:01	1.88
3o3a	A*02:01	1.80	3mro	A*02:01	2.35	2cik	B*35:01	1.75	6d2t	B*57:01	1.90
3gsw	A*02:01	1.81	3mrh	A*02:01	2.40	3lko	B*35:01	1.80	6bxp	B*57:01	1.45
1jf1	A*02:01	1.85	3mrl	A*02:01	2.41	3lkp	B*35:01	1.80	6bxq	B*57:01	1.58
3o3e	A*02:01	1.85	4jfo	A*02:01	2.11	3lkq	B*35:01	1.80	5vue	B*57:01	1.80
5enw	A*02:01	1.85	4jfp	A*02:01	1.91	4pr5	B*35:01	1.80	6d2r	B*57:01	1.83
3h7b	A*02:01	1.88	4jfq	A*02:01	1.90	4pra	B*35:01	1.85	5vuf	B*57:01	1.90
5mer	A*02:01	1.88	2j8u	A*02:01	2.88	2h6p	B*35:01	1.90	5vud	B*57:01	2.00
3myj	A*02:01	1.89	2uwe	A*02:01	2.40	3lks	B*35:01	1.90	3vh8	B*57:01	1.80
1t1z	A*02:01	1.90	2jcc	A*02:01	2.50	2axg	B*35:01	2.00	3x12	B*57:01	1.80

Table 5-5. Continued.

PDB_id	Allele	Resolution (Å)	PDB_id	Allele	Resolution (Å)	PDB_id	Allele	Resolution (Å)	PDB_id	Allele	Resolution (Å)
2guo	A*02:01	1.90	2p5e	A*02:01	1.89	3lkn	B*35:01	2.00	3upr	B*57:01	2.00
2x4q	A*02:01	1.90	2p5w	A*02:01	2.20	3lkr	B*35:01	2.00	3wuw	B*57:01	2.00
3ft4	A*02:01	1.90	3hg1	A*02:01	3.00	4lnr	B*35:01	2.00	6d2b	B*57:01	2.04
3gjf	A*02:01	1.90	1lp9	A*02:01	2.00	1a9e	B*35:01	2.50	5b38	B*57:01	2.30
3gsv	A*02:01	1.90	3utt	A*02:01	2.60	4qrr	B*35:01	3.00	2rfx	B*57:01	2.50
3o3b	A*02:01	1.90	3uts	A*02:01	2.71	1a9b	B*35:01	3.20	5b39	B*57:01	2.50
3rew	A*02:01	1.90	2bnq	A*02:01	1.70	2nx5	B*35:01	2.70	5t6x	B*57:01	1.69
5hhp	A*02:01	1.90	2bnr	A*02:01	1.90	3mv7	B*35:01	2.00	5t6y	B*57:01	1.76
6pte	A*02:01	1.90	2f54	A*02:01	2.70	3mv8	B*35:01	2.10	5t6w	B*57:01	1.90
3fqw	A*02:01	1.93	2f53	A*02:01	2.10	4prp	B*35:01	2.50	5t6z	B*57:01	2.00
3ft3	A*02:01	1.95	3qfj	A*02:01	2.29	3mv9	B*35:01	2.70	5t70	B*57:01	2.10
3gsr	A*02:01	1.95	2gj6	A*02:01	2.56	4u1m	B*42:01	1.18	5v5m	B*57:01	2.88
1t1y	A*02:01	2.00	1ao7	A*02:01	2.60	4u1j	B*42:01	1.38	5vwh	B*58:01	1.65
2clr	A*02:01	2.00	3pwp	A*02:01	2.69	1.00E+27	B*51:01	2.20	5vwj	B*58:01	2.00
2x70	A*02:01	2.00	3h9s	A*02:01	2.70	4mji	B*51:01	2.99	5ind	B*58:01	2.13
3giv	A*02:01	2.00	1qrn	A*02:01	2.80	3spv	B*08:01	1.30	5im7	B*58:01	2.50
3hpj	A*02:01	2.00	1qse	A*02:01	2.80	4qrs	B*08:01	1.40	5inc	B*58:01	2.88
3v5d	A*02:01	2.00	1qsf	A*02:01	2.80	4qrt	B*08:01	1.40	5txs	B*15:01	1.70
4k7f	A*02:01	2.00	3d3v	A*02:01	2.80	6p2c	B*08:01	1.40	1xr9	B*15:01	1.79
5swq	A*02:01	2.00	3d39	A*02:01	2.81	6p2f	B*08:01	1.48	3c9n	B*15:01	1.87
5hhn	A*02:01	2.03	6rsy	A*02:01	2.95	6p27	B*08:01	1.59	1xr8	B*15:01	2.30
2x4u	A*02:01	2.10	3o4l	A*02:01	2.54	4qru	B*08:01	1.60	5vz5	B*15:01	2.59
3gsx	A*02:01	2.10	1bd2	A*02:01	2.50	6p23	B*08:01	1.60	4lcy	B*46:01	1.60
3ixa	A*02:01	2.10	2pye	A*02:01	2.30	6p2s	B*08:01	1.65	5w1w	C*01:02	3.10
5hhq	A*02:01	2.10	3qeq	A*02:01	2.59	4qrq	B*08:01	1.70	5w1v	C*01:02	3.31
4uq3	A*02:01	2.10	3qdm	A*02:01	2.80	3skm	B*08:01	1.80	1efx	C*03:04	3.00
3gsq	A*02:01	2.12	3qdj	A*02:01	2.30	3x13	B*08:01	1.80	6jtn	C*08:02	1.90
1duy	A*02:01	2.15	3qdg	A*02:01	2.69	1m05	B*08:01	1.90	6jtp	C*08:02	1.90
1jht	A*02:01	2.15	6tro	A*02:01	3.00	3x14	B*08:01	2.00	5xos	C*16:04	1.70
6ptb	A*02:01	2.15	1oga	A*02:01	1.40	1agd	B*08:01	2.05	5xot	C*16:04	2.79
1t21	A*02:01	2.19	2vlr	A*02:01	2.30	1agc	B*08:01	2.10	1qqd	C*04:01	2.70
1i1y	A*02:01	2.20	2vlj	A*02:01	2.40	1agb	B*08:01	2.20	1im9	C*04:01	2.80
1i7r	A*02:01	2.20	2vlk	A*02:01	2.50	1agf	B*08:01	2.20	6jto	C*05:01	1.70
1im3	A*02:01	2.20	6rpb	A*02:01	2.50	1age	B*08:01	2.30	5w6a	C*06:02	1.74
1qew	A*02:01	2.20	6rpa	A*02:01	2.56	3sko	B*08:01	1.60	5w69	C*06:02	2.80
1s8d	A*02:01	2.20	6rp9	A*02:01	3.12	3ffc	B*08:01	2.80	6joz	C*15:10	1.35
1s9w	A*02:01	2.20	3gsn	A*02:01	2.80	4qrp	B*08:01	2.90	6jp3	C*15:10	1.66
1t1w	A*02:01	2.20	5tez	A*02:01	1.70	1mi5	B*08:01	2.50	6pbh	C*15:10	1.89
1t1x	A*02:01	2.20	5yxn	A*02:01	2.03	3sjv	B*08:01	3.10	6id4	C*15:10	2.40
1t20	A*02:01	2.20	5yxu	A*02:01	2.70	3bxn	B*14:02	1.86	5wjn	C*15:10	2.85
1t22	A*02:01	2.20	3ox8	A*02:03	2.16	3bvn	B*14:02	2.55	5wjl	C*15:10	3.15
3bhb	A*02:01	2.20	3oxr	A*02:06	1.70	6pyl	B*27:03	1.52	5wkf	C*15:10	2.95
3i6g	A*02:01	2.20	6p64	A*02:06	3.05	6pz5	B*27:03	1.53	5wkh	C*15:10	3.20
3mgt	A*02:01	2.20	3oxs	A*02:07	1.75	5def	B*27:04	1.60	4nt6	C*08:01	1.84

Table 5-6. The reference panel alleles and supertype classifications from previous studies.

Allele	Sidney et al.	Doytchinova et al.	Allele	Sidney et al.	Doytchinova et al.	Allele	Sidney et al.	Doytchinova et al.
A*01:01	A01	A3	B*42:01	B07	B7	B*44:03	B44	B44
A*26:01	A01	A2	B*51:01	B07	B44	B*45:01	B44	B27
A*26:02	A01	A2	B*51:02	B07	B44	B*15:16	B58	B27
A*26:03	A01	A2	B*51:03	B07	B44	B*15:17	B58	B27
A*29:02	A01 A24	A3	B*53:01	B07	B44	B*57:01	B58	B44
A*30:01	A01 A03	A3	B*54:01	B07	B7	B*57:02	B58	B44
A*30:02	A01	A3	B*55:01	B07	B7	B*58:01	B58	B44
A*30:03	A01	A3	B*55:02	B07	B7	B*58:02	B58	B44
A*30:04	A01	A3	B*56:01	B07	B7	B*15:01	B62	B27
A*32:01	A01	A3	B*67:01	B07	B7	B*15:02	B62	B7
A*02:01	A02	A2	B*78:01	B07	B7	B*15:12	B62	B27
A*02:02	A02	A2	B*08:01	B08	B7	B*15:13	B62	B27
A*02:03	A02	A2	B*08:02	B08	B44	B*46:01	B62	B27
A*02:04	A02	A2	B*14:02	B27	B7	B*52:01	B62	B44
A*02:05	A02	A2	B*15:03	B27	B27	C*01:02		C1
A*02:06	A02	A2	B*15:09	B27	B7	C*03:02		C1
A*02:07	A02	A2	B*15:10	B27	B7	C*07:02		C1
A*02:14	A02	A2	B*15:18	B27	B7	C*08:01		C1
A*02:17	A02	A2	B*27:02	B27	B44	C*12:02		C1
A*68:02	A02	A2	B*27:03	B27	B27	C*14:02		C1
A*69:01	A02	A2	B*27:04	B27	B27	C*16:01		C1
A*03:01	A03	A3	B*27:05	B27	B27	C*16:04		C1
A*11:01	A03	A3	B*27:06	B27	B27	C*02:02		C2
A*31:01	A03	A3	B*27:07	B27	B27	C*03:07		C2
A*33:01	A03	A3	B*27:09	B27	B27	C*03:15		C2
A*33:03	A03	A3	B*38:01	B27	B44	C*04:01		C2
A*66:01	A03	A2	B*39:01	B27	B7	C*05:01		C2
A*68:01	A03	A3	B*39:02	B27	B27	C*06:02		C2
A*74:01	A03	A3	B*39:09	B27	B7	C*07:07		C2
A*23:01	A24	A24	B*48:01	B27	B27	C*12:04		C2
A*24:02	A24	A24	B*73:01	B27	B7	C*12:05		C2
B*07:02	B07	B7	B*18:01	B44	B7	C*15:02		C2
B*07:03	B07	B7	B*37:01	B44	B27	C*16:02		C2
B*07:05	B07	B7	B*40:01	B44	B27	C*17:01		C2
B*15:08	B07	B7	B*40:02	B44	B27	C*18:01		C2
B*35:01	B07	B7	B*40:06	B44	B27			
B*35:03	B07	B7	B*44:02	B44	B44			

Table 5-7. Selecting optimal shape parameters and residue similarity matrix. Sum of SSEs were colored green-red while Sum of Pearson's Rs were colored red-green, and green indicates better performance.

Shape parameter grid		Sum of SSE [a]			Sum of Pearson's R [a]		
sigma	**k**	**Grantham**	**PMBEC**	**SM_THREAD_NORM**	**Grantham**	**PMBEC**	**SM_THREAD_NORM**
0.075	0.8	175.38	151.25	171.13	1.87	2.12	1.92
0.1	0.8	169.02	150.90	176.28	1.96	2.18	1.99
0.2	0.8	152.86	165.67	164.05	2.19	2.29	2.18
0.3	0.8	147.24	167.50	150.00	2.26	2.29	2.23
0.5	0.8	157.20	158.18	152.76	2.27	2.23	2.22
0.075	1	177.31	150.99	185.41	1.91	2.15	1.95
0.1	1	168.73	155.15	168.86	2.00	2.21	2.03
0.2	1	154.87	174.63	151.07	2.22	2.30	2.21
0.3	1	142.49	163.41	155.86	2.27	2.27	2.23
0.5	1	148.14	157.42	150.84	2.26	2.21	2.20
0.075	2	175.01	159.85	165.56	2.04	2.23	2.06
0.1	2	159.85	160.48	161.76	2.14	2.28	2.15
0.2	2	142.77	164.98	157.04	2.27	2.27	2.23
0.3	2	153.00	158.37	150.00	2.26	2.22	2.21
0.5	2	140.89	155.34	158.70	2.21	2.17	2.16
0.075	4	158.22	163.92	162.68	2.17	2.29	2.17
0.1	4	147.66	174.91	149.53	2.24	2.29	2.22
0.2	4	152.83	160.78	154.72	2.26	2.21	2.21
0.3	4	142.05	159.39	151.47	2.23	2.18	2.17
0.5	4	150.76	149.14	153.26	2.16	2.14	2.11
0.075	8	147.21	165.17	149.91	2.25	2.29	2.23
0.1	8	142.77	160.29	153.47	2.26	2.25	2.23
0.2	8	149.83	158.72	151.47	2.23	2.18	2.17
0.3	8	164.42	149.24	160.72	2.18	2.15	2.13
0.5	8	153.09	155.85	150.17	2.12	2.11	2.07
0.075	16	146.96	162.14	152.01	2.26	2.24	2.23
0.1	16	150.09	160.78	152.26	2.25	2.21	2.20
0.2	16	164.42	149.24	162.18	2.19	2.15	2.14
0.3	16	155.59	151.65	147.71	2.13	2.13	2.09
0.5	16	151.90	159.87	147.93	2.06	2.08	2.02

[a] Sum of values of HLA-A. HLA-B, and HLA-C

Table 5-8. Residue similarity matrix S adapted from Grantham distance matrix, derived as described in Methods section 2.1.

Arg	Leu	Pro	Thr	Ala	Val	Gly	Ile	Phe	Tyr	Cys	His	Gln	Asn	Lys	Asp	Glu	Met	Trp	
0.49	0.33	0.66	0.73	0.54	0.42	0.74	0.34	0.28	0.33	0.48	0.59	0.68	0.79	0.44	0.70	0.63	0.37	0.18	**Ser**
1.00	0.53	0.52	0.67	0.48	0.55	0.42	0.55	0.55	0.64	0.16	0.87	0.80	0.60	0.88	0.55	0.75	0.58	0.53	**Arg**
	1.00	0.54	0.57	0.55	0.85	0.36	0.98	0.90	0.83	0.08	0.54	0.47	0.29	0.50	0.20	0.36	0.93	0.72	**Leu**
		1.00	0.82	0.87	0.68	0.80	0.56	0.47	0.49	0.21	0.64	0.65	0.58	0.52	0.50	0.57	0.60	0.32	**Pro**
			1.00	0.73	0.68	0.73	0.59	0.52	0.57	0.31	0.78	0.80	0.70	0.64	0.60	0.70	0.62	0.40	**Thr**
				1.00	0.70	0.72	0.56	0.47	0.48	0.09	0.60	0.58	0.48	0.51	0.41	0.50	0.61	0.31	**Ala**
					1.00	0.49	0.87	0.77	0.74	0.11	0.61	0.55	0.38	0.55	0.29	0.44	0.90	0.59	**Val**
						1.00	0.37	0.29	0.32	0.26	0.54	0.60	0.63	0.41	0.56	0.54	0.41	0.14	**Gly**
							1.00	0.90	0.85	0.08	0.56	0.49	0.31	0.53	0.22	0.38	0.95	0.72	**Ile**
								1.00	0.90	0.05	0.53	0.46	0.27	0.53	0.18	0.35	0.87	0.81	**Phe**
									1.00	0.10	0.61	0.54	0.33	0.60	0.26	0.43	0.83	0.83	**Tyr**
										1.00	0.19	0.28	0.35	0.06	0.28	0.21	0.09	0.00	**Cys**
											1.00	0.89	0.68	0.85	0.62	0.81	0.60	0.47	**His**
												1.00	0.79	0.75	0.72	0.87	0.53	0.40	**Gln**
													1.00	0.56	0.89	0.80	0.34	0.19	**Asn**
														1.00	0.53	0.74	0.56	0.49	**Lys**
															1.00	0.79	0.26	0.16	**Asp**
																1.00	0.41	0.29	**Glu**
																	1.00	0.69	**Met**

Table 5-9. Random peptide set for measuring peptide binding specificity distance.

Peptide length	Natural relative abundance [a]	Recalculated ratio (%) [b]	Number of peptides	Random sequence length [c]
8	0.207	8.2	4120	4127
9	1	39.8	19904	19912
10	0.422	16.8	8400	8409
11	0.366	14.6	7285	7295
12	0.244	9.7	4857	4868
13	0.179	7.1	3563	3575
14	0.094	3.7	1871	1884
15	0.065	-	-	-

[a] The relative abundance compared to 9-mer as reported in Ref 60.

[b] The ratio was recalculated as NetMHCpan only accept residue length from 8 to 14.

[c] The input of NetMHCpan is a long protein sequence, then the NetMHCpan server split the sequence into peptides of certain length using sliding window algorithm.

Table 5-10 Anchor alleles defined for each supertype and subtype for nearest neighbor clustering. Clusters that have arbitrary anchor were highlighted in orange.

Anchor alleles	Subtype	Supertype
A*01:01	A01	A01-A03-A66
A*03:01	A03	
A*11:01		
A*30:01		
A*66:01	**A66**	
A*02:01	A02	A02
A*02:03		
A*02:06		
A*02:07		
A*68:01		
A*24:02	A24	A24
B*07:02	B07	B07-B35
B*42:01		
B*35:01	B35	
B*08:01	B08	B08-B18-B39
B*18:01	B18	
B*39:01	B39	
B*14:02	B14	B14
B*15:01	B15	B15-B40
B*40:02		
B*40:01	B40	
B*27:05	B27	B27
B*44:02	B44	B44
B*44:03		
B*51:01	B51	B51-B58
B*57:01	B58	
B*58:01		
C*04:01	C01	C01-C02
C*05:01		
C*08:02		
B*46:01	C02	
C*06:02		
C*07:01	**C07**	**C07**

Chapter 6

Prediction of Pseudomonas aeruginosa permeability to small molecules using random forest models

My contribution to this project is to perform data cleaning, cheminformatics analysis, and Random Forest classification.

A version of this chapter was originally published as:

Leus, I.V., Weeks, J.W., Bonifay, V., Shen, Y., Yang, L., Cooper, C.J., Nash, D., Duerfeldt, A.S., Smith, J.C., Parks, J.M., Rybenkov, V.V., Zgurskaya, H. I. (2022). Property space mapping of Pseudomonas aeruginosa permeability to small molecules. *Scientific Reports*, 12(1), 1-17.

Author contribution:

All authors read the manuscript and contributed to its writing. Leus, I.V. and Weeks, J.W. measured antibacterial activities and performed accumulation experiments, Bonifay, V. developed LC-MS method and carried out compound quantification experiments, Shen, Y. carried out cheminformatics analyses and Random Forest classification, Yang, L. performed accumulation experiments with radioactive compounds, Cooper, C.J. calculated physicochemical descriptors and carried out machine learning analyses, Nash, D. synthesized compounds, Duerfeldt, A.S. designed compounds and synthetic routes, Smith, J.C. and Parks, J.M. designed and supervised cheminformatics and Random Forest experiments, Rybenkov, V.V. designed and supervised the studies, developed Gnomic algorithms, carried out machine learning experiments, Zgurskaya, H. I. designed and supervised the studies and performed intracellular accumulation studies and antibacterial analyses.

Abstract

Two membrane cell envelopes act as selective permeability barriers in Gram-negative bacteria, protecting cells against noxious molecules. Significant efforts are being directed toward understanding how small molecules permeate these barriers. In this study, we measured the accumulation of a library of structurally diverse compounds in four isogenic strains of Pseudomonas aeruginosa, an important human pathogen notorious for multidrug resistance, with varied permeability properties using mass spectrometry. Based on the physicochemical properties and accumulation of these compounds, we further developed a random forest model to predict good permeators. We posit that this approach can be used for more detailed mapping of the property space and for rational design of compounds with high Gram-negative permeability.

6.1 Introduction

Pseudomonas aeruginosa is a challenging pathogen that causes a variety of infections in humans and animals. Only a few therapeutic options are available for *P. aeruginosa* infections and those fail against multidrug-resistant strains [157, 158]. This challenge is exacerbated by a lack of success in discovering new small molecules with antibacterial activities against this pathogen [159].

The major reason why *P. aeruginosa* presents such a formidable therapeutic challenge is that the cells are protected by an effective drug permeability barrier composed of two lipid membranes and reinforced by active multidrug efflux pumps [9]. Currently, major efforts are being directed toward understanding and overcoming this drug permeation barrier and optimizing drug potencies against *P. aeruginosa* and other Gram-negative pathogens [160-162].

Drug permeation studies in bacteria, analyses of intracellular accumulation of a broad range of chemical structures have been enabled by liquid chromatography-tandem mass spectrometry (LC-MS/MS). A recent study conducted a larger analysis of a diversity library of 100 compounds built around five major natural product-derived scaffolds and found only 12 positively charged compounds that accumulated significantly in E. coli, and, within this subset, no correlation was observed with hydrophobicity or molecular weight [163, 164], in contrast to some early studies of radioactively labeled fluoroquinolones [165, 166]. However, all 12 compounds contained amines, 8 of which were primary amines. Calculations of physicochemical descriptors, synthesis of

additional analogues, and application of a random forest classification model to 68 compounds revealed that rotatable bonds and globularity were negative predictors of accumulation while the amphiphilic moment, the distance between hydrophobic and hydrophilic portions of a molecule, was a positive predictor.

A similar study found no correlations between the size and hydrophobicity and the intracellular accumulation of these inhibitors [167]. In agreement with early studies of small sets of compounds [168, 169], they also found poor correlation between the overall level of bacteria-associated compound and antibacterial activity in compounds with matched biochemical activities.

The varying significance of physicochemical properties such as hydrophobicity and molecular weight in the intracellular accumulation of compounds is a consequence of the complex nature of the Gram-negative permeability barrier. Active efflux pumps, which transport compounds across cell envelope, act synergistically with the membranes, together forming a permeability barrier for any given compound [170, 171]. This synergistic relationship manifests as highly non-linear changes in intracellular accumulation levels depending on the extracellular concentration of compounds and the species-specific compositions of efflux pumps and the OM. An additional difficulty is that the above studies analyze small sample sizes with limited diversity compared to the entirety of chemical property space.

Most of the previous drug permeation studies have been limited to *E. coli*, however, these data are not fully applicable to all pathogens. In contrast, the cell envelope of *P. aeruginosa* is notably less permeable to certain compounds than that of *E. coli* [172]. The general porins located in the *E. coli* OM enable diffusion of amphiphilic molecules up to 600 Da in size [173], while *P. aeruginosa* only have substrate-specific porins that limit natural uptake of amino acids, simple sugars, and short peptides [174]. Interestingly, inactivation of all characterized and predicted porin-like proteins in *P. aeruginosa* led to only minor changes in antibiotic susceptibility, suggesting that almost all antibiotics and diverse hydrophilic nutrients can bypass porins and instead permeate directly through the OM lipid bilayer [174]. In addition, *P. aeruginosa* cells express several multidrug efflux pumps with partially overlapping substrate specificities that pump out compounds that do manage to permeate slowly across the OM [175, 176].

The contributions of the OM barrier and active efflux in antibacterial activities of compounds can be separated by comparing these activities in wild-type, efflux-deficient, hyperporinated, and

'barrierless' (i.e., both hyperporinated and efflux-deficient) strains [177, 178]. In combination with machine learning techniques, such analyses identified physicochemical properties of compounds predictive of their interactions with proteins and lipids of the OM and active efflux pumps and showed that each pathogen presents its own unique challenges. Both barriers (i.e., OM permeability and efflux avoidance) must be considered during antibiotic development efforts, because optimizing the properties of compounds to overcome one barrier tends to result in detrimental properties against the other barrier [177, 179].

In this study, the accumulation of drug-like libraries composed of 83 compounds in P. aeruginosa was measured by an LC-MS approach. The "good" permeators were defined as compounds that accumulate in wild type cells linearly with increasing concentration but that are not affected by compromising permeation barriers and have low non-specific retention on control filters. We further calculated the physicochemical properties of 66 molecules, and selected a descriptor set with the minimal synonymity between its members. Finally, Random Forest model was trained to predict permeability of small molecule from its physicochemical features. This strategy can be used for further, more detailed mapping of property space and for the rational design of compounds with optimized Gram-negative permeability.

6.2 Methods

6.2.1 Data used in study

The intracellular accumulation level of 83 compounds in four isogenic strains of *P. aeruginosa*, at four concentration and two time points, was measure by Inga V. Leus, Jon W. Weeks of the University of Oklahoma. Among the tested molecules, 17 were excluded as they failed the blank control, leaving 66 for further analysis.

6.2.2 Cheminformatics

The chemical diversity of our compounds was compared to that of five other libraries and intracellular accumulation studies, including the Shared Platform for Antibiotic Research and Knowledge (SPARK) Public Projects Data [157] (archived by Collaborative Drug Discovery Vault), FDA-approved drugs [180], and small molecule permeability in bacteria studies by Richter et al. [163], Davis et al. [181], and Iyer et al [167]. In this comparison, nine features were used to describe each molecule: the molecular weight, logP, the number of hydrogen bond donors, the

number of hydrogen bond acceptors, log D at pH 7.4, the topological polar surface area, Fsp3, the heavy atom count and the number of rotatable bonds. Molecular features for molecules from the four intracellular accumulation studies were calculated using the Marvin suite from ChemAxon [182]. All molecules were filtered to retain only those with molecular weights between 200 and 1,500 g/mol and to remove ions and peptides. Standardization and principal component analysis were then applied to the data points using scikit-learn version 0.24.1 [183]. To illustrate how much each feature contributes to a particular principal component, PCA loadings were labeled, which was defined as the coefficients of the linear combination of the original variables from which the principal components are constructed. Figures were generated by matplotlib (version 3.3.4) [184] and seaborn (version 0.11.1) [185].

6.2.3 Physicochemical property calculation

Tautomers and protonation states at pH 7.4 for each compound were predicted using Marvin from ChemAxon [182]. A total of 217 physicochemical properties, 2D and 3D, were then calculated by Connor J. Cooper and Jerry M. Parks of Center for Molecular Biophysics, ORNL.

6.2.4 Selection of descriptors

The initial list of 217 descriptors was manually inspected to select only those that represent integral physicochemical or topological properties of a molecule or generic molecular fingerprints. The selected 47 descriptors were then clustered and analyzed by a modified elbow method. Namely, the unexplained variance was calculated as a function of the number of clusters, fit to a three-segmented straight line, and the larger elbow used as a cutoff. Descriptors closest to the centers of the identified 27 clusters were subsequently used in ML analyses.

6.2.5 Random forest classification

A procedure consisting of data standardization, oversampling, and random forest (RF) classification was developed for this task. All procedures were performed with Python 3.8.8 in a Jupyter Notebook environment (Available at https://github.com/yshen25/EPI_ML) with scikit-learn version 0.24.1 [183] and imbalanced-learn version 0.8.0 [186]. Of the 66 compounds, 25 were identified as permeators and the remaining 41 were labeled as non-permeators. Because the dataset is imbalanced, the use of an oversampling technique was expected to be beneficial. Based on the mapping result (Fig. 6-4), the permeators belong to six different clusters in feature space,

so common techniques such as SMOTE [187] and ADASYN [188] may produce noise points because synthetic data points may be situated in overlapping class regions. To overcome this issue, k-means SMOTE [189] was applied. By giving more weight to sparse minority areas and avoiding overlapping zones, k-means SMOTE is expected to be better at overcoming skewness. Physicochemical properties were standardized to have mean and standard deviation of 0 and 1.

To obtain increased accuracy in the RF model, three hyperparameters were tuned using an exhaustive grid search. Hyperparameters included the number of trees in the random forest (n_estimators), the minimum number of samples required at a leaf node (min_samples_leaf), and the maximum number of features considered at each split (max_features). The hyperparameter combination with the highest accuracy was then selected for model training and testing. The performance of the final model was evaluated by leave-one-out cross validation. The 95% confidence interval was calculated by performing five-fold stratified cross validation repeated five times. Model performance was evaluated by the following scoring metrics. The accuracy describes the portion of samples that are correctly classified and is defined as follows:

$$accuracy = \frac{true\ positives + \ true\ negatives}{all\ samples}$$

Precision is the proportion of true positives among all predicted positives, also called positive predictive value:

$$precision = \frac{true\ positives}{true\ positive + false\ positives}$$

Recall is the proportion of true positives among all actual positives, also called sensitivity.

$$recall = \frac{true\ positives}{true\ positives + false\ negatives}$$

The F1 score assesses the balance between precision and recall:

$$F1 = 2 \times \frac{precision \times recall}{precision + recall}$$

The receiver operating characteristic (ROC) curve plots the true positive rate versus the false positive rate. The area under the ROC curve, AUROC, quantifies the ability of the model to

discriminate between classes, with 1.0 corresponding to a perfect classifier and 0.5 corresponding to random split.

6.3 Results

6.3.1 Properties of the focused library of compounds for analysis

Previous studies have shown that the physicochemical space of antibiotics is broader than other drugs and includes more hydrophilic compounds, such as beta-lactams, which penetrate the cells through water-filled OM porins, as well as some large compounds, such as vancomycin, which act on cell surfaces and do not need to penetrate the cell envelope [190]. In addition, the presence of primary amines and positive charges are in general associated with increased penetration into some bacterial cells [191]. From a focused library of 12,000 commercially available compounds, 220 compounds with PSA 50 +, MW < 2000 Da, and $cLogD_{7.4} < 5$ were selected for further analyses.

Approximately 10% of the 220 purchased compounds possessed antibacterial activities in at least one of the four *P. aeruginosa* strains, including representatives of fluoroquinolones, sulfonamides, cyclines and linezolid (data not shown). After elimination of insoluble compounds and optimization of the LC-MS method, 98 compounds were selected for further analyses, out of which the intracellular accumulation of 83 compounds was measured and for 66 compounds it was quantified.

We calculated nine physicochemical properties of the analyzed 66 compounds, including the molecular weight, logP, the number of hydrogen bond donors and acceptors, $clogD_{7.4}$, the topological polar surface area, the fraction of sp^3 hybridized carbon atoms (Fsp3), the heavy atom count, and the number of rotatable bonds for the analyzed compounds. These properties were also calculated for the SPARK compound library [192], FDA-approved drugs [180], and compounds whose accumulation in *E. coli* was analyzed in previous studies [167, 181, 191]. The dataset of 66 compounds is relatively small, but it is comparable to previous studies [163, 167]. Principal component analysis showed that the chemical space covered in this work is broader than the space covered by previous studies [167, 181, 191] (Fig. 6-1). The distribution of properties of this library surrounds that of Iyer et al. [167] in every feature and that of Richter et al. [191] in principal component 1 (PC1), which is represented by molecular weight, number of hydrogen bond donors and acceptors, topological polar surface area, rotatable bounds, and heavy atom count (Fig. 6-2).

The distribution of the compounds analyzed by Richter et al. [191] is wider in the Fsp3 property space but narrower in logP and logD, resulting in a slightly narrower distribution in principal component 2 (PC2). The partial overlap in PC1 between our compounds and the Richter et al. set is due to an offset in the range of values for the topological polar surface area and the number of hydrogen bond acceptors. For example, the number of hydrogen bond acceptors ranges from 0-8 in the Richter et al. set and from 2-12 in the present study.

6.3.2 Classification of compounds

The measured levels of compound accumulation were analyzed using machine learning to determine which features of the compounds are most important for enabling their cellular accumulation and to determine their optimal values. For modeling, we used compound concentration slopes of the accumulation levels observed after 1 min and 40 min incubations, which were approximated as straight lines with y-intercept set to zero. To accentuate the permeability aspect, the accumulation slopes in efflux-deficient and hyperporinated strains were normalized to those in the parental PAO1 strain. In several cases, the slopes were lower in the permeabilized cells than in wild type cells, presumably due to data scatter or low signal. Such changes cannot be attributed to the porination state or efflux deficiency of the cell. Therefore, such cases were attributed to the lack of an increase in permeability and the ratios were set to 1.

To curate accumulation data for training binary classification ML models, permeators and non-permeators were defined based on three criteria. First, the accumulation of permeators should be equal to or higher than outside the cell, which was indicated by the slope at 40 min larger than 1. Second, permeators should be rarely influenced by porination state or efflux deficiency of the cell, and the total barrier ratio (PΔ3-Pore/PAO1) should be less than 1.3. Third, molecules with concentration of cell-free filter higher than in-cell accumulation were classified as non-permeators. A total of 25 compounds satisfied these criteria and were classified as good permeators.

6.3.3 Random Forest classification

We trained a random forest model to distinguish between permeators and non-permeators using 27 2D and 3D physicochemical descriptors for each compound that were pre-selected from a larger descriptor pool by clustering analysis (see Methods). The model achieved only 61% accuracy, 48% recall and 46% precision when tested on leave-one-out cross-validation (Fig. 6-3A). The area

under the receiver-operator characteristic (AUROC) curve was only 0.53. The broad confidence interval (Fig. 6-3B) indicates that the performance of the model is largely dictated by which compounds were included in the training set for a given cross-validation step. This behavior implies that a larger compound set would likely improve the performance of the model. Descriptors derived from molecular dynamics (MD) simulations are prevalent among the top ten, suggesting that molecular shape, conformational flexibility, and dynamics are important for predicting permeability (Fig. 6-3C). Examples include shape-based descriptors such as the Hall-Kier kappa3 molecular shape descriptor, the average acylindricity and asphericity, i.e., the deviation from cylindrical and spherical symmetries, and the radius of gyration (average Rg and average smallest principal Rg). Other top descriptors associated with favorable uptake include lipophilicity (MolLogP) and properties related to molecule topology (second principal moment of inertia, minimum absolute E-state index, and Morgan fingerprint density).

A Gnomic classifier developed using MATLAB 2019a with the Statistics and Machine Learning toolbox was demonstrated to have accuracy between 85% and 97% in predicting the node of compound from property (Fig 6-4, detailed data not shown). It employed 7 descriptors that were derived from several rounds of selection. By using the 7 descriptors selected by Gnomic classifier in RF model instead of all the 27 descriptors mentioned above, the performance improved significantly (Fig 6-5). The accuracy increased to 0.80 and the AUROC increased to 0.66. It illustrated that the removal of synonymous or noisy descriptors could help classification and prediction of molecule property. These 7 descriptors comprise a reference set in the property space for inspecting compounds. The top descriptors of good permeators mainly define topological, constitutional, and molecular properties of compounds (Table 6-1). Among these, shape (avg_asphericity) and number of rotatable bonds (NumRotatableBonds) were highlighted in previously developed models based on the studies of either intracellular accumulation or antibacterial activities [163, 177, 181]. However, a closer inspection shows that the numerical values of these descriptors vary broadly between different permeation nodes and do not clearly distinguish them.

6.4 Discussion

In this study, the permeation of small molecules into *P. aeruginosa* was measured and the associated physicochemical properties were analyzed. A total of 66 structurally diverse

compounds were tested, among which, 25 was identified as good permeators. Also, among the 217 calculated molecular properties, 7 was finally demonstrated to be the most important.

The molecular signature of good permeators consisted of a mixture of distinct properties including the size and shape of the molecule, polarity, topology, and a handful of molecular fingerprints such as the number of aromatic rings or rotatable bonds (Fig. 6-4C). These properties are often found in this type of analysis and can be readily traced to the mechanism of small molecule penetration into bacteria [163, 177, 181]. The true challenge was to find numeric values of these parameters that are conducive to penetration.

To solve this problem, we performed ML analysis. First, we trained an RF model in order to predict good permeators from the properties. However, due to the limited size and broad coverage of chemical space in the dataset, the performance metrics for the RF model were unsatisfactory (Fig. 6-2C). Thus, a Gnomic algorithm was applied to select descriptors with minimal information overlap by ranking descriptors according to their unpredictability. This approach would create an acceptable compromise between absorbing unique molecular signatures available in diverse descriptors and avoiding the synonymity between them. As a result, 7 descriptors were selected. By using the stripped set of descriptors instead of all 27 descriptors, the performance of RF model improved significantly. Clearly, the use of synonymous descriptors in a model only expands the dimensionality of the solution without improving its predictive power.

Traditionally, ML methods address this issue through the use of curvature tests, which eliminate highly correlated descriptors. The descriptor clustering step builds on this idea and sets an adaptive threshold for the acceptable correlation. However, these steps capture only pairwise correlations between descriptors and often overlook the instances when the information contributed by one descriptor is fully contained in a group of others. Gnomic searches for such instances by building regression models that predict a given descriptor through the rest. A low misclassification error in such models signifies a low informational contribution from said descriptor. This approach can be viewed as an ML equivalent of the linear algebra concept of a linear independence between vectors. Unlike in linear algebra, the predictability can recapture instances of non-linear interaction between descriptors. This feature makes it better suited for the analysis of seemingly stochastic distributions such as the one examined here.

The inspection of structures of good permeators did not yield a single functional group consistently present in all compounds. However, some of the moieties such as trifluoro- or diaminomethyl-substitutions are known to improve permeation across bacterial membranes [178, 193]. Our results also show that the contribution of such chemical modifications to permeation across *P. aeruginosa* and other Gram-negative cell envelopes will be scaffold dependent. Further analyses of intracellular permeation and modeling of larger chemical libraries around several biologically active scaffolds using hit expansion techniques will lead to a more detailed mapping of the space of good permeators and could potentially reveal scaffold-dependent heuristics.

Appendix

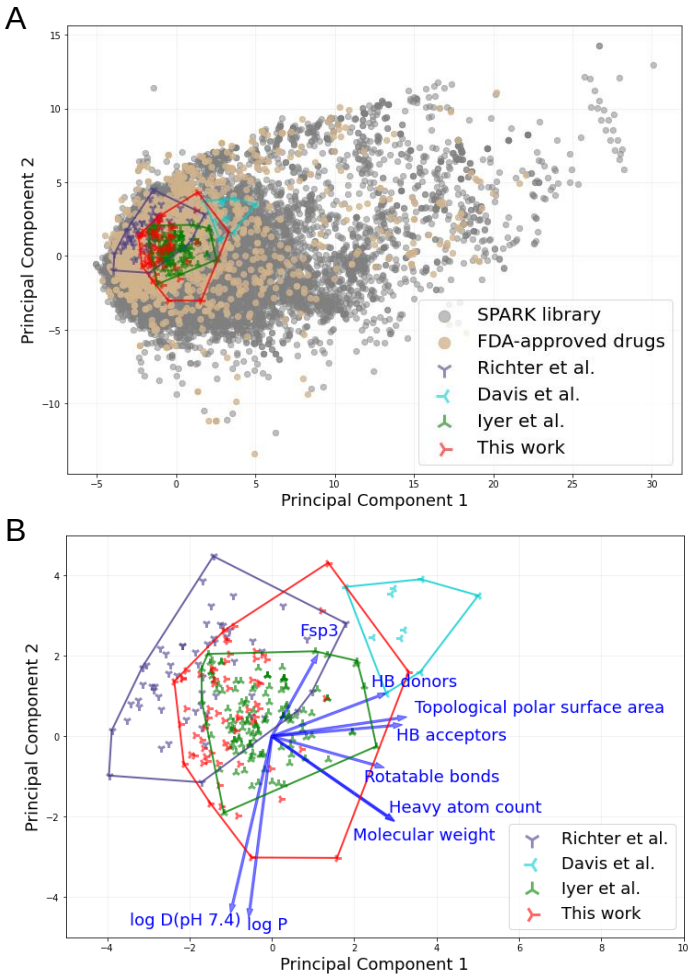


Figure 6-1. Properties of the library of 66 compounds analyzed in this study. (A) Principal component analysis of diversity in the chemical space of molecules from SPARK library [192], Richter et al. [163], Davis et al. [181], Iyer et al., [167], FDA approved drugs (extracted from [180]) and this study. The outermost data points from the four studies are linked to show the relative size of the sampled space. (B) Zoom-in view of (A) and the loading vectors for this plot. Data points for the SPARK library and FDA-approved drugs are hidden for clarity. Molecular weight, hydrogen bond donors and acceptors, topological polar surface area, rotatable bounds and heavy atom count are the major contributing factors to PC1, whereas LogP, logD and the fraction of sp3 hybridized carbon atoms (Fsp3) are major contributors to PC2. Loading vectors were scaled by a factor of 7 for clarity.

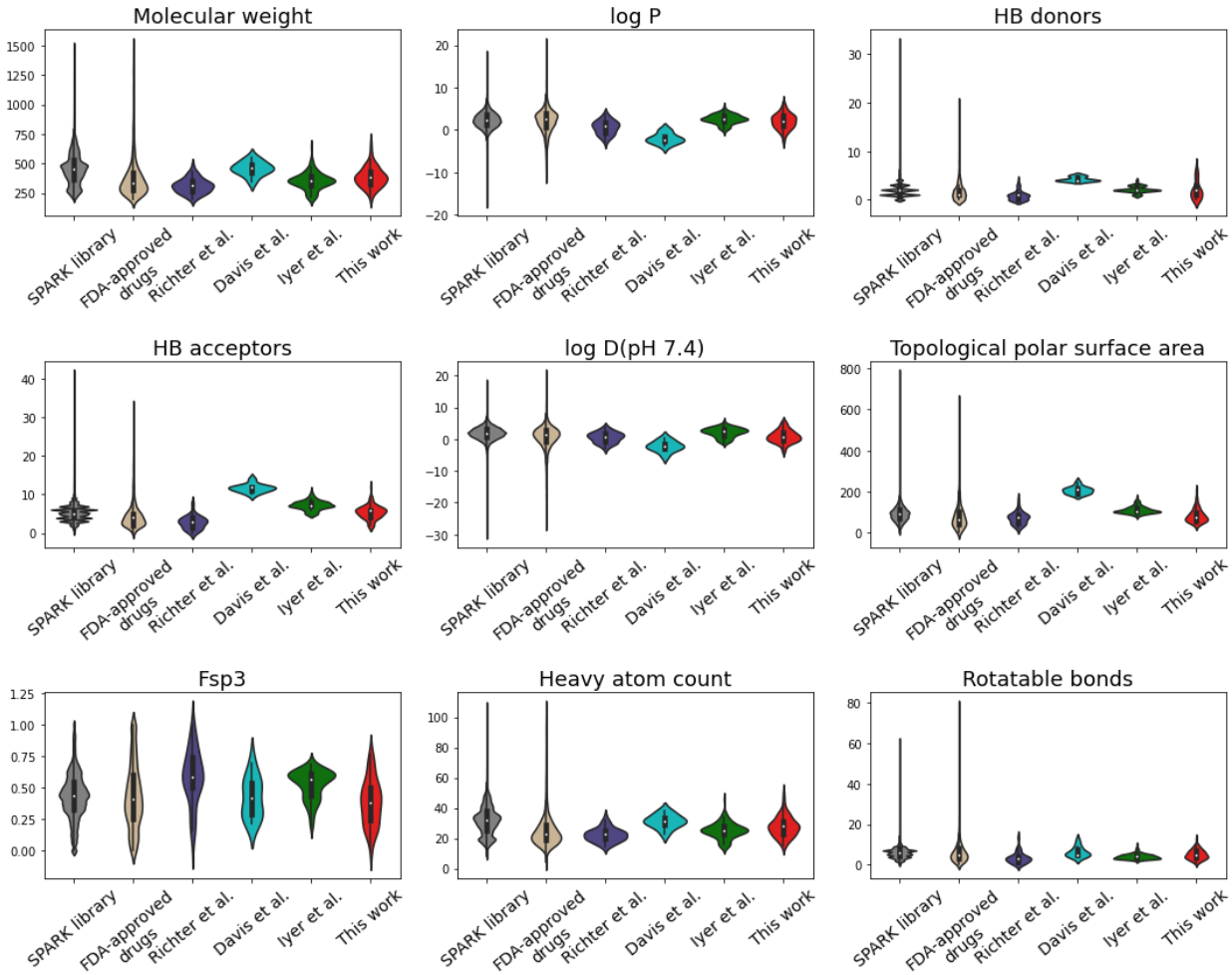


Figure 6-2. Violin plots of compounds from two chemical libraries and four studies showing the distribution of molecules among each of nine molecular features used to describe chemical space diversity.

A

Precision	Recall	F1	Accuracy	AUROC
0.458	0.478	0.44	0.606	0.53

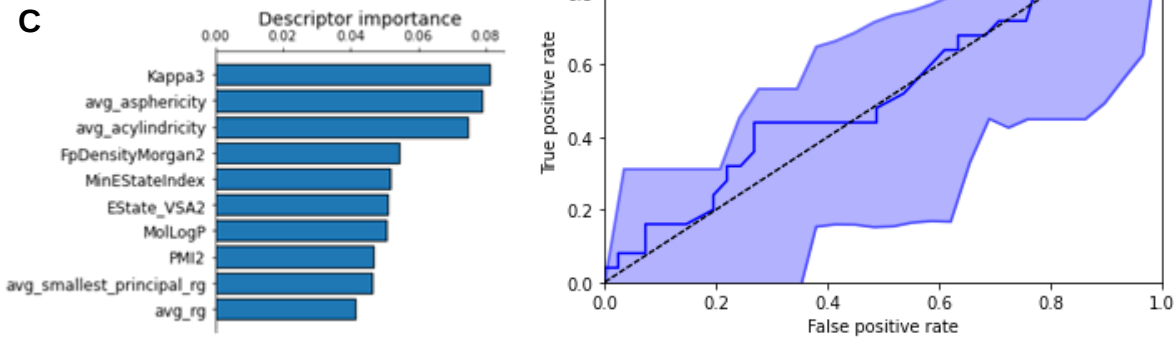


Figure 6-3. Performance of random forest model using 27 features. (A) Performance metrics. (B) ROC curve derived from leave-one-out cross validation with 95% confidence bands. (C) Importance bar plot of top ten features.

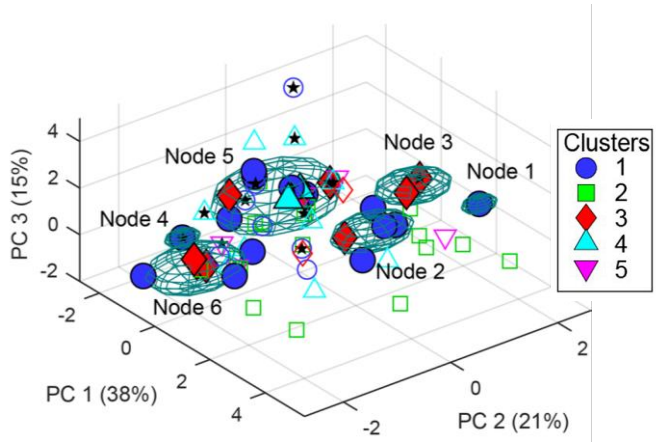
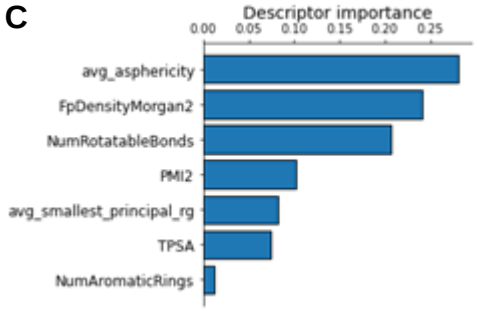


Figure 6-4. Distribution of compounds in the property space of 7 features. Clusters were assigned according to permeability in four strains (data not shown). Good permeators were filled marks while bad permeators were hollow. Figure was generated by Valentin V. Rybenkov.

A

Precision	Recall	F1	Accuracy	AUROC
0.60	0.45	0.51	0.80	0.66



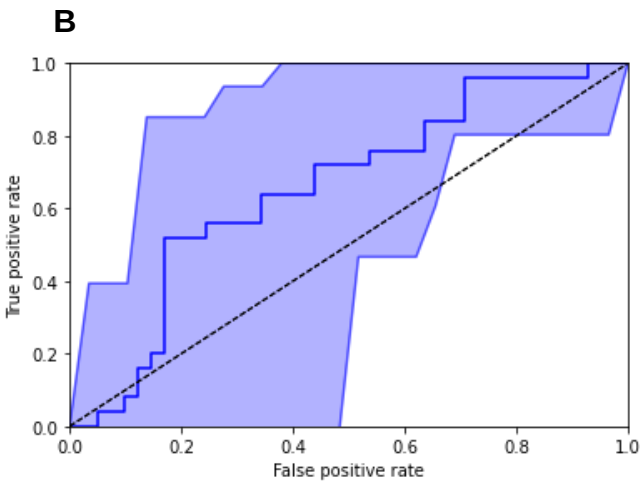


Figure 6-5. Performance of random forest model using 7 features. (A) Performance metrics. (B) ROC curve derived from leave-one-out cross validation with 95% confidence bands. (C) Importance bar plot.

Table 6-1. The 7 descriptors selected by Gnomic algorithm and their definitions.

Name	Type	Definition [a]	Class
NumRotatableBonds	1D	Number of Rotatable Bonds	Constitutional descriptor
NumAromaticRings	1D	Number of Aromatic rings	Constitutional descriptor
PMI2	2D	Second (largest) principal moment of inertia	Topological descriptor
FpDensityMorgan2	2D	Morgan fingerprint density	Topological descriptor
TPSA	2D	Topological polar surface area	Molecular property descriptor
avg_asphericity	3D	average asphericity over 100 ns MD trajectory	MD
avg_smallest_principal_rg	3D	average smallest principal radius of gyration over 100 ns MD trajectory	MD

[a] RDKit descriptor definitions adapted from https://cb.imsc.res.in/deduct/descriptors/eJaFhpBqbWtuZWt9

Chapter 7 Conclusion and future works

7.1 Conclusion

In this project, we explored the application of machine learning methods in the development of new therapies, peptide-based MHC I epitope vaccines and small antimicrobial molecules, to fight against the emerging threats towards human health, including cancer, pandemic outbreak, and antimicrobial resistance. Multiple machine learning methods, sequence-based and structure-based, were applied during study and were demonstrated to contribute to the development of new therapies.

The peptide-based vaccines show unique advantages in safety and producibility and were considered a promising technique in fighting against cancer and acute infectious diseases. In chapter 2 and 3, we developed peptide-based vaccines, including personalized pancreatic cancer vaccine and SARS-Cov-2 vaccine for the general public. The general concepts and approaches of peptide-based vaccine design, which could be easily applied to the studies about other diseases, were illustrated. First, the sequences of disease-related proteins were derived from gene sequencing. For pancreatic cancer vaccines, the protein regions that carry tumor specific point mutations were used as target sequences. And for SARS-CoV-2 vaccines, the target sequences are early-stage viral proteins. Then, peptide candidates that tightly bind to MHC alleles carried by target population were selected, the peptide: MHC affinity was predicted using artificial neural network models. Two types of affinity predictors were applied: allele-specific predictors that have high accuracy but low allele coverage and pan-specific predictors that have good coverage but unsatisfying accuracy. Usually, peptide-based cancer vaccines are personalized that require MHC mapping to identify patient's MHC haplotype, while virus vaccines need to consider covering populated alleles. Finally, peptide candidates were further filtered either using knowledge-based or sequence- and structure-based methods. For example, the cancer neoantigens were selected to be derived from oncogenes, and the favored neoantigens have agretopic indices > 4. In another example, SARS-CoV-2 vaccines were selected from regions of low mutation entropy, and the strong binders were validated using flexible docking methods.

To improve the accuracy of peptide: MHC binding prediction, we applied structure-based methods. In chapter 3, the development of peptide-based COVID-19 vaccine, strong peptide binders were validated using method based on flexible docking. Thus, in chapter 4, we developed

automated HLA modeling pipeline to provide high quality HLA structures for this method. The pipeline was built on the Rosetta CM homology modeling protocol, operations such as template selection, HPC job submission, and file management, were automated using Python scripts. In chapter 5, we further demonstrated that the peptide binding specificity is mainly determined by the landscape of peptide binding groove. Also, based on the structural similarity, populated HLA class I alleles were classified into 12 supertypes and 20 subtypes. The use of supertypes would facilitate the supertype-specific peptide: MHC affinity predictors that fill in the gap between allele-specific and pan-specific predictors, also contribute to HLA-disease association studies.

Small molecule drugs are widely used in the treatment of bacterial infections, however, the rapid spread of antimicrobial resistance is blocking available treatment options. Gram-negative bacteria have permeability barrier, outer membrane (OM), and efflux pumps that decrease concentration of antimicrobials inside the cell, demonstrating multidrug resistance. In order to develop new drugs or rejuvenate outdated antimicrobials against resistant gram-negative strains, permeability of candidate molecules need to be tested. In section 6, the accumulation of various small molecules inside *P. aeruginosa* were measured, and we developed random forest (RF) models aiming at identifying the decisive physicochemical features that govern permeability. Initially, 217 features were calculated. After several rounds of dimension reduction, 27 features were used to establish the first RF model, which showed poor performance. The successful Gmonic model illustrated that the activity group could be predicted from 7 features. By using the same 7 features, we built the second RF model and got better performance than the first model.

The limitation of above-mentioned machine learning methods mainly lies in accuracy. The 14 predicted neoantigen pancreatic cancer vaccine was tested in mice, and only 1 showed protective activity. Also, the RF model that tried to predict small molecule permeability achieved AUROC less than 0.7, showing unsatisfying performance. Clearly, the major reason is the lack of experimental data to train the model, on the other hand, it is also the major motivation to apply *in silico* methods rather than entirely relying on experiments. Thus, careful data curation and model building would be a key factor of successful application of ML methods. The importance of data curation was well illustrated by the development of permeability RF model, as the different set of features used for building the model results in different performance. Also, an advanced machine learning approach could better overcome imperfect data and achieve higher performance, for

example the use of structural modeling plus structure distance metric showed higher correlation with peptide binding specificity than simple sequence comparison, while both approaches start from HLA protein sequences. With the development of machine learning algorithms and accumulation of high-quality data, the accuracy would be improved gradually.

Another limitation is the comprehensiveness of ML models used in specific study. Due to the limited available training data and specific purpose of each project, the derived ML models are focused on solving a specific question. As a result, the ML model may not be suitable for comprehensive purposes. For example, it is obvious that the development of peptide-based vaccines targeting cancer and virial infection are largely different, though rely on the same mechanism. We used allele-specific affinity predictor NetMHC when predicting neoantigens from mice tumor and used pan-specific predictor NetMHCpan when predicting SARS-Cov-2 epitopes and peptide binding specificity. The reason is that the allele-specific predictors have better performance than pan specific predictors but narrower coverage on supported alleles. The MHC system is much simpler than human HLA system, and all studied MHC alleles carried by C57BL/6J were included in NetMHC predictor, while the studied HLA alleles are not fully covered. Another example is the RF model and Gnomic model in chapter 6. The poor performance and instability of RF models were largely due to the diversity of studied molecules. Illustrated by the Fig. 6-4 that good permeators are separated in multiple clusters, the physicochemical rule that decides the permeability is not unique. Plus, the number of molecules is far from enough, as some good permeator clusters only have a few molecules, the decision boundary could not be clearly revealed, which further determined that the performance of RF model varied significantly in the cross validation. In comparison, the Gnomic method was fitted to predict 5 clusters in the activity space, which are also close in the feature space, and resulted in much better performance. It well demonstrated that the comprehensiveness of ML methods is closely related to the training data, which should be kept in mind when applying pre-developed methods in other projects. In response to such considerations, we included the stability test in chapter 5 to show the extensibility of our method.

Explainability is an important concern when applying ML models. As mentioned above, the limitations of ML methods used in a study determines that the model itself may have limited added value, while understanding the process of model building may provide insights into the question.

In chapter 5, we reported both the clustering result and the dendrogram in order to illustrate relationship between supertypes and subtypes, which was used to validate the inter-locus functional overlap. Also, the clustering result could be explained using structural comparison, which directly supported the hypothesis that the peptide binding specificity is mainly determined by the landscape of peptide binding groove, also showed the credibility of clustering result. In chapter 6, although the RF models failed to accurately predict molecule permeability from its physicochemical property, we were still able to acquire the feature importance information that guide future studies. The negative example comes from the peptide: MHC affinity predictors that used neural networks. The neural networks usually have outstanding performance compared to other ML approaches, however, is known for being a black box algorithm that is hard to interpret. It leads to the problem that the abstract prediction process provides little information for better understanding the given data, nor justifies the prediction result. Thus, explainable machine learning methods is a hot topic and would benefit future studies.

Although advances have been made by applying ML and other in silico methods, the experimental assay is still the gold standard that cannot be bypassed. Generally, a treatment development process includes discovery, pre-clinical research, clinical trial, and regulatory approval, and takes 14 years to complete on average. Among these steps, ML methods focus on accelerating the discovery, while other steps must be done by experiments. In addition, within the discovery phase, ML methods still need to be supported by experiments. For example, the in silico development of peptide-based vaccines could suggest candidates that are more likely to be effective, which increases the efficiency significantly compared to random selection, while the efficacy has to be further verified by immunogenicity assays. It is because of two reasons: 1) the inaccuracy of peptide: MHC affinity predictors. 2) the complexity of peptide presentation and T cell recognition. The former issue could be improved by the development of ML methods, the latter one could not and need to combine other approaches.

In conclusion, we demonstrated that the application of machine learning methods could benefit treatment development. The advantages of structure-based methods over sequence-based methods have been demonstrated throughout the project. With the development of structural modeling algorithms, for instance AlphaFold, and increase in capacity of HPCs, this disadvantage will surely be overcome, declaring the dawn of post-genomics. Because the sequence determines structure,

and structure determines function. In the genomics era, a large number of genome sequence were made available, based on which the bioinformatics tools that predict or compare protein functions were developed, significantly promotes the studies in treatment development. Nowadays, the protein structures could be predicted easily and accurately, so that the structural informatics is likely to replicate the impact as of bioinformatics.

7.2 Future work

Many simulations and tests were left undone due to the limited time for performing research. Also, as an early section in treatment development, predictions made by ML methods need to be further validated by experiments, many of which are still ongoing. Meanwhile, there are new insights brought by ML methods from both prediction process and result worth further investigation. Thus, future works will be focused on three aspects:

First, to improve the accuracy and explainability of ML methods used in this project, refinement will be carried out by enlarging training data set and implementing advanced algorithms. One of the future works is to broaden the scope of application of the structure-based docking method that applied to validate strong binder peptide prediction. This method is limited to 9 mer peptides and certain MHC alleles. We have established pipeline to model HLA alleles supporting such method, which solved the issue of limited number of available HLA structures, so that the structure-based binder validation method would be applied in more alleles. Also, efforts should be made to study the relationship between structural flexibility of peptide-MHC complex and their affinity, in order to extend the applicable peptide length. In addition, the RF model for predicting small molecule permeability discussed in chapter 6 showed poor performance largely due to the unsatisfying dataset. Correspondingly, by incorporating more molecules into study, the performance of RF model would improve, and more useful information could be revealed.

Second, feedback from experimental assays will be incorporated and used to guide the refinement of ML methods. Our collaborators have performed immunogenicity assay that tested the 14 putative pancreatic cancer vaccine discussed in chapter 2, among which 1 epitope has been reported to strongly correlate with anti-tumor T cell activity. This valuable feedback facilities further investigation of the mechanism behind peptide presentation of MHC molecules, and possible methods to improve the true positive rate of peptide-based vaccine prediction. Another possibility is to analyze the possible correlation between predicted SARS-CoV-2 peptide epitopes

and efficacy of COVID-19 vaccines, which could be considered as clinical review and used to validate the prediction result. Also, it would in return provide a potential strategy to improve existing vaccines by refining the composition of protein particle or nucleic acid, such as removing regions that cause allergic response and modifying mutations that cause immune escape.

Finally, there are new findings reveled during the studies that worth further investigation. In chapter 5, we demonstrated that the defined structure distance metric is highly correlated with peptide binding specificity. Such finding supports the hypothesis that the peptide binding pattern is mainly determined by the landscape of peptide binding groove. Following this trace, we could examine the molecular mechanism of the peptide binding process using molecular dynamics simulation, which could reveal the major interactions that contribute to binding affinity and have the potential to develop new methods to calculate the peptide: MHC affinity.

References

1. Sung, H., et al., *Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries.* CA Cancer J Clin, 2021. **71**(3): p. 209-249.
2. Global Burden of Disease Cancer, C., et al., *Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life Years for 29 Cancer Groups From 2010 to 2019: A Systematic Analysis for the Global Burden of Disease Study 2019.* JAMA Oncol, 2022. **8**(3): p. 420-444.
3. *WHO COVID-19 Dashboard* [cited 2020; Available from: https://covid19.who.int/.
4. Fosgerau, K. and T. Hoffmann, *Peptide therapeutics: current status and future directions.* Drug Discov Today, 2015. **20**(1): p. 122-8.
5. Blass, E. and P.A. Ott, *Advances in the development of personalized neoantigen-based therapeutic cancer vaccines.* Nat Rev Clin Oncol, 2021. **18**(4): p. 215-229.
6. Heitmann, J.S., et al., *A COVID-19 peptide vaccine for the induction of SARS-CoV-2 T cell immunity.* Nature, 2022. **601**(7894): p. 617-622.
7. Nikaido, H., *Prevention of drug access to bacterial targets: permeability barriers and active efflux.* Science, 1994. **264**(5157): p. 382-8.
8. Zgurskaya, H.I., et al., *Trans-envelope multidrug efflux pumps of Gram-negative bacteria and their synergism with the outer membrane barrier.* Res Microbiol, 2018. **169**(7-8): p. 351-356.
9. Zgurskaya, H.I. and V.V. Rybenkov, *Permeability barriers of Gram-negative pathogens.* Ann N Y Acad Sci, 2020. **1459**(1): p. 5-18.
10. Paul, S.M., et al., *How to improve R&D productivity: the pharmaceutical industry's grand challenge.* Nat Rev Drug Discov, 2010. **9**(3): p. 203-14.
11. Brogi, S., et al., *Editorial: In silico Methods for Drug Design and Discovery.* Front Chem, 2020. **8**: p. 612.
12. Vamathevan, J., et al., *Applications of machine learning in drug discovery and development.* Nat Rev Drug Discov, 2019. **18**(6): p. 463-477.
13. Klein, J. and A. Sato, *The HLA system.* New England Journal of Medicine, 2000. **343**(10): p. 702-709.
14. Hewitt, E.W., *The MHC class I antigen presentation pathway: strategies for viral immune evasion.* Immunology, 2003. **110**(2): p. 163-169.
15. Hennecke, J. and D.C. Wiley, *T cell receptor-MHC interactions up close.* Cell, 2001. **104**(1): p. 1-4.
16. Koutsakos, M., et al., *Downregulation of MHC Class I Expression by Influenza A and B Viruses.* Front Immunol, 2019. **10**: p. 1158.
17. Cornel, A.M., I.L. Mimpen, and S. Nierkens, *MHC Class I Downregulation in Cancer: Underlying Mechanisms and Potential Targets for Cancer Immunotherapy.* Cancers (Basel), 2020. **12**(7).
18. Matzaraki, V., et al., *The MHC locus and genetic susceptibility to autoimmune and infectious diseases.* Genome Biol, 2017. **18**(1): p. 76.
19. Wieczorek, M., et al., *Major histocompatibility complex (MHC) class I and MHC class II proteins: conformational plasticity in antigen presentation.* Frontiers in immunology, 2017. **8**: p. 292.
20. Wu, Y., et al., *Structural Comparison Between MHC Classes I and II; in Evolution, a Class-II-Like Molecule Probably Came First.* Front Immunol, 2021. **12**: p. 621153.
21. Nakamura, T., et al., *The Role of Major Histocompatibility Complex in Organ Transplantation- Donor Specific Anti-Major Histocompatibility Complex Antibodies Analysis Goes to the Next Stage.* Int J Mol Sci, 2019. **20**(18).
22. Barker, D.J., et al., *The IPD-IMGT/HLA Database.* Nucleic Acids Res, 2022.
23. World Health, O., *Global action plan on antimicrobial resistance*. 2015, Geneva: World Health Organization.
24. Melnyk, A.H., A. Wong, and R. Kassen, *The fitness costs of antibiotic resistance mutations.* Evol Appl, 2015. **8**(3): p. 273-83.
25. Vogwill, T. and R.C. MacLean, *The genetic basis of the fitness costs of antimicrobial resistance: a meta-analysis approach.* Evol Appl, 2015. **8**(3): p. 284-95.
26. Alonso, A., et al., *Overexpression of the multidrug efflux pump SmeDEF impairs Stenotrophomonas maltophilia physiology.* J Antimicrob Chemother, 2004. **53**(3): p. 432-4.
27. Olivares, J., et al., *Overproduction of the multidrug efflux pump MexEF-OprN does not impair Pseudomonas aeruginosa fitness in competition tests, but produces specific changes in bacterial regulatory networks.* Environ Microbiol, 2012. **14**(8): p. 1968-81.

28. Martinez, J.L., *General principles of antibiotic resistance in bacteria.* Drug Discov Today Technol, 2014. **11**: p. 33-9.
29. Pages, J.M., C.E. James, and M. Winterhalter, *The porin and the permeating antibiotic: a selective diffusion barrier in Gram-negative bacteria.* Nat Rev Microbiol, 2008. **6**(12): p. 893-903.
30. Beceiro, A., M. Tomas, and G. Bou, *Antimicrobial resistance and virulence: a successful or deleterious association in the bacterial world?* Clin Microbiol Rev, 2013. **26**(2): p. 185-230.
31. Reygaert, W., *Methicillin-resistant Staphylococcus aureus (MRSA): molecular aspects of antimicrobial resistance and virulence.* Clin Lab Sci, 2009. **22**(2): p. 115-9.
32. Blair, J.M., et al., *Molecular mechanisms of antibiotic resistance.* Nat Rev Microbiol, 2015. **13**(1): p. 42-51.
33. Ramirez, M.S. and M.E. Tolmasky, *Aminoglycoside modifying enzymes.* Drug Resist Updat, 2010. **13**(6): p. 151-71.
34. Robicsek, A., et al., *Fluoroquinolone-modifying enzyme: a new adaptation of a common aminoglycoside acetyltransferase.* Nat Med, 2006. **12**(1): p. 83-8.
35. Bush, K. and P.A. Bradford, *beta-Lactams and beta-Lactamase Inhibitors: An Overview.* Cold Spring Harb Perspect Med, 2016. **6**(8).
36. Pfeifer, Y., A. Cullik, and W. Witte, *Resistance to cephalosporins and carbapenems in Gram-negative bacterial pathogens.* Int J Med Microbiol, 2010. **300**(6): p. 371-9.
37. Blair, J.M., G.E. Richmond, and L.J. Piddock, *Multidrug efflux pumps in Gram-negative bacteria and their role in antibiotic resistance.* Future Microbiol, 2014. **9**(10): p. 1165-77.
38. Kumar, A. and H.P. Schweizer, *Bacterial resistance to antibiotics: active efflux and reduced uptake.* Adv Drug Deliv Rev, 2005. **57**(10): p. 1486-513.
39. Poole, K., *Efflux pumps as antimicrobial resistance mechanisms.* Ann Med, 2007. **39**(3): p. 162-76.
40. Bray, F., et al., *Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.* CA Cancer J Clin, 2018. **68**(6): p. 394-424.
41. Siegel, R.L., K.D. Miller, and A. Jemal, *Cancer statistics, 2019.* CA Cancer J Clin, 2019. **69**(1): p. 7-34.
42. Kanda, M., et al., *Presence of somatic mutations in most early-stage pancreatic intraepithelial neoplasia.* Gastroenterology, 2012. **142**(4): p. 730-733 e9.
43. Yurgelun, M.B., et al., *Germline cancer susceptibility gene variants, somatic second hits, and survival outcomes in patients with resected pancreatic cancer.* Genet Med, 2019. **21**(1): p. 213-223.
44. Ino, Y., et al., *Immune cell infiltration as an indicator of the immune microenvironment of pancreatic cancer.* Br J Cancer, 2013. **108**(4): p. 914-23.
45. Li, J., et al., *Tumor Cell-Intrinsic Factors Underlie Heterogeneity of Immune Cell Infiltration and Response to Immunotherapy.* Immunity, 2018. **49**(1): p. 178-193 e7.
46. Collins, M.A., et al., *Oncogenic Kras is required for both the initiation and maintenance of pancreatic cancer in mice.* J Clin Invest, 2012. **122**(2): p. 639-53.
47. Masugi, Y., et al., *Characterization of spatial distribution of tumor-infiltrating CD8(+) T cells refines their prognostic utility for pancreatic cancer survival.* Mod Pathol, 2019. **32**(10): p. 1495-1507.
48. Balachandran, V.P., G.L. Beatty, and S.K. Dougan, *Broadening the Impact of Immunotherapy to Pancreatic Cancer: Challenges and Opportunities.* Gastroenterology, 2019. **156**(7): p. 2056-2072.
49. Liu, L., et al., *Low intratumoral regulatory T cells and high peritumoral CD8(+) T cells relate to long-term survival in patients with pancreatic ductal adenocarcinoma after pancreatectomy.* Cancer Immunol Immunother, 2016. **65**(1): p. 73-82.
50. Karamitopoulou, E., et al., *High tumor mutational burden (TMB) identifies a microsatellite stable pancreatic cancer subset with prolonged survival and strong anti-tumor immunity.* Eur J Cancer, 2022. **169**: p. 64-73.
51. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.
52. Van der Auwera, G.A., et al., *From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline.* Curr Protoc Bioinformatics, 2013. **43**: p. 11 10 1-11 10 33.
53. McLaren, W., et al., *The Ensembl Variant Effect Predictor.* Genome Biol, 2016. **17**(1): p. 122.
54. Krupke, D.M., et al., *The Mouse Tumor Biology Database: A Comprehensive Resource for Mouse Models of Human Cancer.* Cancer Res, 2017. **77**(21): p. e67-e70.
55. Hundal, J., et al., *pVACtools: A Computational Toolkit to Identify and Visualize Cancer Neoantigens.* Cancer Immunol Res, 2020. **8**(3): p. 409-420.
56. Andreatta, M. and M. Nielsen, *Gapped sequence alignment using artificial neural networks: application to the MHC class I system.* Bioinformatics, 2016. **32**(4): p. 511-7.

57. Lawrence, M.S., et al., *Mutational heterogeneity in cancer and the search for new cancer-associated genes.* Nature, 2013. **499**(7457): p. 214-218.
58. Duan, F., et al., *Genomic and bioinformatic profiling of mutational neoepitopes reveals new rules to predict anticancer immunogenicity.* J Exp Med, 2014. **211**(11): p. 2231-48.
59. Feig, C., et al., *Targeting CXCL12 from FAP-expressing carcinoma-associated fibroblasts synergizes with anti-PD-L1 immunotherapy in pancreatic cancer.* Proc Natl Acad Sci U S A, 2013. **110**(50): p. 20212-7.
60. Waddell, N., et al., *Whole genomes redefine the mutational landscape of pancreatic cancer.* Nature, 2015. **518**(7540): p. 495-501.
61. Balachandran, V.P., et al., *Identification of unique neoantigen qualities in long-term survivors of pancreatic cancer.* Nature, 2017. **551**(7681): p. 512-516.
62. Ajina, R. and L.M. Weiner, *T-Cell Immunity in Pancreatic Cancer.* Pancreas, 2020. **49**(8): p. 1014-1023.
63. Bailey, P., et al., *Exploiting the neoantigen landscape for immunotherapy of pancreatic ductal adenocarcinoma.* Sci Rep, 2016. **6**: p. 35848.
64. Pon, J.R. and M.A. Marra, *Driver and passenger mutations in cancer.* Annu Rev Pathol, 2015. **10**: p. 25-50.
65. McFarland, C.D., et al., *The Damaging Effect of Passenger Mutations on Cancer Progression.* Cancer Res, 2017. **77**(18): p. 4763-4772.
66. Calina, D., et al., *Towards effective COVID-19 vaccines: Updates, perspectives and challenges (Review).* Int J Mol Med, 2020. **46**(1): p. 3-16.
67. Azmi, F., et al., *Recent progress in adjuvant discovery for peptide-based subunit vaccines.* Hum Vaccin Immunother, 2014. **10**(3): p. 778-96.
68. Lu, R., et al., *Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding.* Lancet, 2020. **395**(10224): p. 565-574.
69. Lan, J., et al., *Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor.* Nature, 2020. **581**(7807): p. 215-220.
70. Shang, J., et al., *Structural basis of receptor recognition by SARS-CoV-2.* Nature, 2020. **581**(7807): p. 221-224.
71. Walls, A.C., et al., *Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein.* Cell, 2020. **183**(6): p. 1735.
72. Wang, Q., et al., *Structural and Functional Basis of SARS-CoV-2 Entry by Using Human ACE2.* Cell, 2020. **181**(4): p. 894-904 e9.
73. Li, W., et al., *Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus.* Nature, 2003. **426**(6965): p. 450-4.
74. Hofmann, H., et al., *Human coronavirus NL63 employs the severe acute respiratory syndrome coronavirus receptor for cellular entry.* Proc Natl Acad Sci U S A, 2005. **102**(22): p. 7988-93.
75. Yadav, R., et al., *Role of Structural and Non-Structural Proteins and Therapeutic Targets of SARS-CoV-2 for COVID-19.* Cells, 2021. **10**(4).
76. V'Kovski, P., et al., *Coronavirus biology and replication: implications for SARS-CoV-2.* Nat Rev Microbiol, 2021. **19**(3): p. 155-170.
77. Kim, D., et al., *A high-resolution temporal atlas of the SARS-CoV-2 translatome and transcriptome.* Nat Commun, 2021. **12**(1): p. 5120.
78. Li, W., et al., *Stability of SARS-CoV-2-Encoded Proteins and Their Antibody Levels Correlate with Interleukin 6 in COVID-19 Patients.* mSystems, 2022. **7**(3): p. e0005822.
79. Mautner, L., et al., *Replication kinetics and infectivity of SARS-CoV-2 variants of concern in common cell culture models.* Virol J, 2022. **19**(1): p. 76.
80. Hadfield, J., et al., *Nextstrain: real-time tracking of pathogen evolution.* Bioinformatics, 2018. **34**(23): p. 4121-4123.
81. Korber, B., et al., *Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus.* Cell, 2020. **182**(4): p. 812-827 e19.
82. Yurkovetskiy, L., et al., *Structural and Functional Analysis of the D614G SARS-CoV-2 Spike Protein Variant.* Cell, 2020. **183**(3): p. 739-751 e8.
83. Deng, X., et al., *Transmission, infectivity, and neutralization of a spike L452R SARS-CoV-2 variant.* Cell, 2021. **184**(13): p. 3426-3437 e8.
84. Starr, T.N., et al., *Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding.* Cell, 2020. **182**(5): p. 1295-1310 e20.
85. Choi, B., et al., *Persistence and Evolution of SARS-CoV-2 in an Immunocompromised Host.* N Engl J Med, 2020. **383**(23): p. 2291-2293.

86. Greaney, A.J., et al., *Complete Mapping of Mutations to the SARS-CoV-2 Spike Receptor-Binding Domain that Escape Antibody Recognition.* Cell Host Microbe, 2021. **29**(1): p. 44-57 e9.
87. Fehr, A.R. and S. Perlman, *Coronaviruses: an overview of their replication and pathogenesis.* Methods Mol Biol, 2015. **1282**: p. 1-23.
88. Wu, F., et al., *A new coronavirus associated with human respiratory disease in China.* Nature, 2020. **579**(7798): p. 265-269.
89. Shang, J., et al., *Compositional diversity and evolutionary pattern of coronavirus accessory proteins.* Brief Bioinform, 2021. **22**(2): p. 1267-1278.
90. Shannon, C.E., *A mathematical theory of communication.* The Bell system technical journal, 1948. **27**(3): p. 379-423.
91. Vita, R., et al., *The Immune Epitope Database (IEDB): 2018 update.* Nucleic Acids Res, 2019. **47**(D1): p. D339-D343.
92. O'Donnell, T.J., et al., *MHCflurry: open-source class I MHC binding affinity prediction.* Cell systems, 2018. **7**(1): p. 129-132. e4.
93. Shao, X.M., et al., *High-Throughput Prediction of MHC Class I and II Neoantigens with MHCnuggets.* Cancer Immunol Res, 2020. **8**(3): p. 396-408.
94. Reynisson, B., et al., *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 Res, 2020. **48**(W1): p. W449-W454.
95. Zhang, H., O. Lund, and M. Nielsen, *The PickPocket method for predicting binding specificities for receptors based on receptor pocket similarities: application to MHC-peptide binding.* Bioinformatics, 2009. **25**(10): p. 1293-9.
96. Peters, B. and A. Sette, *Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method.* BMC Bioinformatics, 2005. **6**: p. 132.
97. Kim, Y., et al., *Derivation of an amino acid similarity matrix for peptide: MHC binding and its application as a Bayesian prior.* BMC Bioinformatics, 2009. **10**: p. 394.
98. Aranha, M.P., et al., *Combining Three-Dimensional Modeling with Artificial Intelligence to Increase Specificity and Precision in Peptide-MHC Binding Predictions.* J Immunol, 2020. **205**(7): p. 1962-1977.
99. Paiva, V.A., et al., *Protein structural bioinformatics: An overview.* Comput Biol Med, 2022. **147**: p. 105695.
100. Song, Y., et al., *High-resolution comparative modeling with RosettaCM.* Structure, 2013. **21**(10): p. 1735-42.
101. Kaas, Q., M. Ruiz, and M.P. Lefranc, *IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data.* Nucleic Acids Res, 2004. **32**(Database issue): p. D208-10.
102. Schrodinger, LLC, *The PyMOL Molecular Graphics System, Version 1.8*. 2015.
103. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
104. McGuffin, L.J., K. Bryson, and D.T. Jones, *The PSIPRED protein structure prediction server.* Bioinformatics, 2000. **16**(4): p. 404-5.
105. Alford, R.F., et al., *The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design.* J Chem Theory Comput, 2017. **13**(6): p. 3031-3048.
106. Lee, K.H., et al., *Connecting MHC-I-binding motifs with HLA alleles via deep learning.* Commun Biol, 2021. **4**(1): p. 1194.
107. Di Marco, M., et al., *Unveiling the Peptide Motifs of HLA-C and HLA-G from Naturally Presented Peptides and Generation of Binding Prediction Matrices.* J Immunol, 2017. **199**(8): p. 2639-2651.
108. Jin, P. and E. Wang, *Polymorphism in clinical immunology–from HLA typing to immunogenetic profiling.* Journal of translational medicine, 2003. **1**(1): p. 1-11.
109. Kishore, A. and M. Petrek, *Next-generation sequencing based HLA typing: deciphering immunogenetic aspects of sarcoidosis.* Frontiers in genetics, 2018. **9**: p. 503.
110. Sidney, J., et al., *Several HLA alleles share overlapping peptide specificities.* The Journal of Immunology, 1995. **154**(1): p. 247-259.
111. del Guercio, M.-F., et al., *Binding of a peptide antigen to multiple HLA alleles allows definition of an A2-like supertype.* The journal of Immunology, 1995. **154**(2): p. 685-693.
112. Fruci, D., et al., *Anchor residue motifs of HLA class-I-binding peptides analyzed by the direct binding of synthetic peptides to HLA class I α chains.* Human immunology, 1993. **38**(3): p. 187-192.

113. Sidney, J., et al., *The HLA-A0207 peptide binding repertoire is limited to a subset of the A0201 repertoire.* Human immunology, 1997. **58**(1): p. 12-20.
114. Sidney, J., et al., *Definition of an HLA-A3-like supermotif demonstrates the overlapping peptide-binding repertoires of common HLA molecules.* Human immunology, 1996. **45**(2): p. 79-93.
115. Sidney, J., et al., *Specificity and degeneracy in peptide binding to HLA-B7-like class I molecules.* The Journal of Immunology, 1996. **157**(8): p. 3480-3490.
116. Sidney, J., et al., *Practical, biochemical and evolutionary implications of the discovery of HLA class I supermotifs.* Immunology today, 1996. **17**(6): p. 261-266.
117. Sette, A. and J. Sidney, *Nine major HLA class I supertypes account for the vast preponderance of HLA-A and-B polymorphism.* Immunogenetics, 1999. **50**(3): p. 201-212.
118. Lund, O., et al., *Definition of supertypes for HLA molecules using clustering of specificity matrices.* Immunogenetics, 2004. **55**(12): p. 797-810.
119. Kobayashi, H., J. Lu, and E. Celis, *Identification of helper T-cell epitopes that encompass or lie proximal to cytotoxic T-cell epitopes in the gp100 melanoma tumor antigen.* Cancer research, 2001. **61**(20): p. 7577-7584.
120. Panigada, M., et al., *Identification of a promiscuous T-cell epitope in Mycobacterium tuberculosis Mce proteins.* Infection and immunity, 2002. **70**(1): p. 79-85.
121. Doytchinova, I.A. and D.R. Flower, *In silico identification of supertypes for class II MHCs.* The Journal of Immunology, 2005. **174**(11): p. 7085-7095.
122. Cano, P., B. Fan, and S. Stass, *A geometric study of the amino acid sequence of class I HLA molecules.* Immunogenetics, 1998. **48**(5): p. 324-334.
123. McKenzie, L., et al., *Taxonomic hierarchy of HLA class I allele sequences.* Genes & Immunity, 1999. **1**(2): p. 120-129.
124. Chelvanayagam, G., *A roadmap for HLA-A, HLA-B, and HLA-C peptide binding specificities.* Immunogenetics, 1996. **45**(1): p. 15-26.
125. Zhang, C., A. Anderson, and C. DeLisi, *Structural principles that govern the peptide-binding motifs of class I MHC molecules.* Journal of molecular biology, 1998. **281**(5): p. 929-947.
126. Sidney, J., et al., *HLA class I supertypes: a revised and updated classification.* BMC immunology, 2008. **9**(1): p. 1-15.
127. Thomsen, M., et al., *MHCcluster, a method for functional clustering of MHC molecules.* Immunogenetics, 2013. **65**(9): p. 655-665.
128. Reche, P.A. and E.L. Reinherz, *Definition of MHC supertypes through clustering of MHC peptide-binding repertoires*, in *Immunoinformatics*. 2007, Springer. p. 163-173.
129. Doytchinova, I.A., P. Guan, and D.R. Flower, *Identifiying human MHC supertypes using bioinformatic methods.* The Journal of Immunology, 2004. **172**(7): p. 4314-4323.
130. Tong, J.C., T.W. Tan, and S. Ranganathan, *In silico grouping of peptide/HLA class I complexes using structural interaction characteristics.* Bioinformatics, 2007. **23**(2): p. 177-183.
131. Bonsack, M., et al., *Performance Evaluation of MHC Class-I Binding Prediction Tools Based on an Experimentally Validated MHC-Peptide Binding Data Set.* Cancer Immunol Res, 2019. **7**(5): p. 719-736.
132. N'Diaye, A., et al., *Single Marker and Haplotype-Based Association Analysis of Semolina and Pasta Colour in Elite Durum Wheat Breeding Lines Using a High-Density Consensus Map.* PLoS One, 2017. **12**(1): p. e0170941.
133. Abed, A. and F. Belzile, *Comparing Single-SNP, Multi-SNP, and Haplotype-Based Approaches in Association Studies for Major Traits in Barley.* Plant Genome, 2019. **12**(3): p. 1-14.
134. Mirdita, M., et al., *ColabFold: making protein folding accessible to all.* Nature Methods, 2022.
135. Duvaud, S., et al., *Expasy, the Swiss Bioinformatics Resource Portal, as designed by its users.* Nucleic Acids Res, 2021. **49**(W1): p. W216-W227.
136. Trolle, T., et al., *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, 2016. **196**(4): p. 1480-1487.
137. Jumper, J., et al., *Highly accurate protein structure prediction with AlphaFold.* Nature, 2021. **596**(7873): p. 583-589.
138. Henikoff, S. and J.G. Henikoff, *Amino acid substitution matrices from protein blocks.* Proc Natl Acad Sci U S A, 1992. **89**(22): p. 10915-9.
139. Mathieu, A., et al., *The interplay between the geographic distribution of HLA-B27 alleles and their role in infectious and autoimmune diseases: A unifying hypothesis.* Autoimmunity Reviews, 2009. **8**(5): p. 420-425.

140. Elahi, S., et al., *Protective HIV-specific CD8+ T cells evade Treg cell suppression.* Nat Med, 2011. **17**(8): p. 989-95.
141. Nielsen, C.M., et al., *RTS,S malaria vaccine efficacy and immunogenicity during Plasmodium falciparum challenge is associated with HLA genotype.* Vaccine, 2018. **36**(12): p. 1637-1642.
142. Barber, L.D., et al., *The inter-locus recombinant HLA-B* 4601 has high selectivity in peptide binding and functions characteristic of HLA-C.* The Journal of experimental medicine, 1996. **184**(2): p. 735-740.
143. Sette, A., et al., *Peptide binding to the most frequent HLA-A class I alleles measured by quantitative molecular binding assays.* Molecular immunology, 1994. **31**(11): p. 813-822.
144. Shulman-Peleg, A., R. Nussinov, and H.J. Wolfson, *SiteEngines: recognition and comparison of binding sites and protein-protein interfaces.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W337-41.
145. Gao, M. and J. Skolnick, *APoc: large-scale identification of similar protein pockets.* Bioinformatics, 2013. **29**(5): p. 597-604.
146. Yu, D.J., et al., *Designing template-free predictor for targeting protein-ligand binding sites with classifier ensemble and spatial clustering.* IEEE/ACM Trans Comput Biol Bioinform, 2013. **10**(4): p. 994-1008.
147. Maheshwari, S. and M. Brylinski, *Predicting protein interface residues using easily accessible on-line resources.* Briefings in Bioinformatics, 2015. **16**(6): p. 1025-1034.
148. Lee, H.S. and W. Im, *G-LoSA: An efficient computational tool for local structure-centric biological studies and drug design.* Protein Sci, 2016. **25**(4): p. 865-76.
149. Simonovsky, M. and J. Meyers, *DeeplyTough: Learning Structural Comparison of Protein Binding Sites.* J Chem Inf Model, 2020. **60**(4): p. 2356-2366.
150. Leger, C., et al., *Ligand-induced conformational switch in an artificial bidomain protein scaffold.* Sci Rep, 2019. **9**(1): p. 1178.
151. Zarutskie, J.A., et al., *A conformational change in the human major histocompatibility complex protein HLA-DR1 induced by peptide binding.* Biochemistry, 1999. **38**(18): p. 5878-87.
152. Kumar, P., et al., *Conformational changes within the HLA-A1:MAGE-A1 complex induced by binding of a recombinant antibody fragment with TCR-like specificity.* Protein Sci, 2009. **18**(1): p. 37-49.
153. Sadegh-Nasseri, S., et al., *The convergent roles of tapasin and HLA-DM in antigen presentation.* Trends Immunol, 2008. **29**(3): p. 141-7.
154. Sieker, F., S. Springer, and M. Zacharias, *Comparative molecular dynamics analysis of tapasin-dependent and -independent MHC class I alleles.* Protein Sci, 2007. **16**(2): p. 299-308.
155. Eyal, E., et al., *The limit of accuracy of protein modeling: influence of crystal packing on protein structure.* J Mol Biol, 2005. **351**(2): p. 431-42.
156. Tong, J.C., T.W. Tan, and S. Ranganathan, *In silico grouping of peptide/HLA class I complexes using structural interaction characteristics.* Bioinformatics, 2007. **23**(2): p. 177-83.
157. Buhl, M., S. Peter, and M. Willmann, *Prevalence and risk factors associated with colonization and infection of extensively drug-resistant Pseudomonas aeruginosa: a systematic review.* Expert Rev Anti Infect Ther, 2015. **13**(9): p. 1159-70.
158. Behzadi, P., Z. Barath, and M. Gajdacs, *It's Not Easy Being Green: A Narrative Review on the Microbiology, Virulence and Therapeutic Prospects of Multidrug-Resistant Pseudomonas aeruginosa.* Antibiotics (Basel), 2021. **10**(1).
159. Azam, M.W. and A.U. Khan, *Updates on the pathogenicity status of Pseudomonas aeruginosa.* Drug Discov Today, 2019. **24**(1): p. 350-359.
160. Zgurskaya, H.I., *An Old Problem in a New Light: Antibiotic Permeation Barriers.* ACS Infect Dis, 2020. **6**(12): p. 3090-3091.
161. Lewis, K., *The Science of Antibiotic Discovery.* Cell, 2020. **181**(1): p. 29-45.
162. Silver, L.L., *Challenges of antibacterial discovery.* Clin Microbiol Rev, 2011. **24**(1): p. 71-109.
163. Richter, M.F., et al., *Predictive compound accumulation rules yield a broad-spectrum antibiotic.* Nature, 2017. **545**(7654): p. 299-304.
164. Huigens, R.W., 3rd, et al., *A ring-distortion strategy to construct stereochemically complex and structurally diverse compounds from natural products.* Nat Chem, 2013. **5**(3): p. 195-202.
165. Chopra, I. and K. Hacker, *Uptake of minocycline by Escherichia coli.* J Antimicrob Chemother, 1992. **29**(1): p. 19-25.
166. McCaffrey, C., et al., *Quinolone accumulation in Escherichia coli, Pseudomonas aeruginosa, and Staphylococcus aureus.* Antimicrob Agents Chemother, 1992. **36**(8): p. 1601-5.
167. Iyer, R., et al., *Evaluating LC-MS/MS To Measure Accumulation of Compounds within Bacteria.* ACS Infect Dis, 2018. **4**(9): p. 1336-1345.

168. Asuquo, A.E. and L.J. Piddock, *Accumulation and killing kinetics of fifteen quinolones for Escherichia coli, Staphylococcus aureus and Pseudomonas aeruginosa.* J Antimicrob Chemother, 1993. **31**(6): p. 865-80.
169. Piddock, L.J., Y.F. Jin, and D.J. Griggs, *Effect of hydrophobicity and molecular mass on the accumulation of fluoroquinolones by Staphylococcus aureus.* J Antimicrob Chemother, 2001. **47**(3): p. 261-70.
170. Rybenkov, V.V., et al., *The Whole Is Bigger than the Sum of Its Parts: Drug Transport in the Context of Two Membranes with Active Efflux.* Chem Rev, 2021. **121**(9): p. 5597-5631.
171. Saha, P., et al., *Drug Permeation against Efflux by Two Transporters.* ACS Infect Dis, 2020. **6**(4): p. 747-758.
172. Tamber, S. and R.E. Hancock, *The outer membranes of pseudomonads*, in *Pseudomonas*. 2004, Springer. p. 575-601.
173. Nikaido, H., *Molecular basis of bacterial outer membrane permeability revisited.* Microbiol Mol Biol Rev, 2003. **67**(4): p. 593-656.
174. Ude, J., et al., *Outer membrane permeability: Antimicrobials and diverse nutrients bypass porins in Pseudomonas aeruginosa.* Proc Natl Acad Sci U S A, 2021. **118**(31).
175. Westfall, D.A., et al., *Bifurcation kinetics of drug uptake by Gram-negative bacteria.* PLoS One, 2017. **12**(9): p. e0184671.
176. Krishnamoorthy, G., et al., *Synergy between Active Efflux and Outer Membrane Diffusion Defines Rules of Antibiotic Permeation into Gram-Negative Bacteria.* mBio, 2017. **8**(5).
177. Cooper, C.J., et al., *Molecular Properties That Define the Activities of Antibiotics in Escherichia coli and Pseudomonas aeruginosa.* ACS Infect Dis, 2018. **4**(8): p. 1223-1234.
178. Mehla, J., et al., *Predictive Rules of Efflux Inhibition and Avoidance in Pseudomonas aeruginosa.* mBio, 2021. **12**(1).
179. Zhao, S., et al., *Defining new chemical space for drug penetration into Gram-negative bacteria.* Nat Chem Biol, 2020. **16**(12): p. 1293-1302.
180. *Collaborative Drug Discovery*, in *www.collaborativedrug.com* Burlingame, CA.
181. Davis, T.D., C.J. Gerry, and D.S. Tan, *General platform for systematic quantitative evaluation of small-molecule permeability in bacteria.* ACS Chem Biol, 2014. **9**(11): p. 2535-44.
182. ; Available from: www.collaborativedrug.com.
183. Pedregosa, F., et al., *Scikit-learn: Machine learning in Python.* the Journal of machine Learning research, 2011. **12**: p. 2825-2830.
184. Hunter, J.D., *Matplotlib: A 2D graphics environment.* Computing in science & engineering, 2007. **9**(03): p. 90-95.
185. Waskom, M.L., *Seaborn: statistical data visualization.* Journal of Open Source Software, 2021. **6**(60): p. 3021.
186. Lemaître, G., F. Nogueira, and C.K. Aridas, *Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning.* The Journal of Machine Learning Research, 2017. **18**(1): p. 559-563.
187. Chawla, N.V., et al., *SMOTE: synthetic minority over-sampling technique.* Journal of artificial intelligence research, 2002. **16**: p. 321-357.
188. Haibo, H., Yang, B., Garcia, E. A. & Shutao, L., in *IEEE International Joint Conference on Neural Networks*. 2008. p. 1322–1328.
189. Douzas, G., F. Bacao, and F. Last, *Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE.* Information Sciences, 2018. **465**: p. 1-20.
190. O'Shea, R. and H.E. Moser, *Physicochemical properties of antibacterial compounds: implications for drug discovery.* J Med Chem, 2008. **51**(10): p. 2871-8.
191. Richter, M.F. and P.J. Hergenrother, *The challenge of converting Gram-positive-only compounds into broad-spectrum antibiotics.* Ann N Y Acad Sci, 2019. **1435**(1): p. 18-38.
192. Thomas, J., et al., *Shared Platform for Antibiotic Research and Knowledge: A Collaborative Tool to SPARK Antibiotic Discovery.* ACS Infect Dis, 2018. **4**(11): p. 1536-1539.
193. Mansbach, R.A., et al., *Machine Learning Algorithm Identifies an Antibiotic Vocabulary for Permeating Gram-Negative Bacteria.* J Chem Inf Model, 2020. **60**(6): p. 2838-2847.

Vita

Yue Shen acquired Bachelor of Science degree in Biotechnology from Shanghai Jiao Tong University. After that, he decided to step further into science and chose to attend University of Tennessee, Knoxville to pursue a Doctor of Philosophy degree in Life Science. His study concentrated on the in-silico development of drugs and peptide-based vaccines. Along with the path to a PhD degree, Yue also pursued a master's degree in Statistics to better apply Machine Learning technique in his research.

Publications

Shen, Y., Parks, J.M., & Smith, J.C. (2023) HLA Class I Supertype Classification Based on Structural Similarity. *The Journal of Immunology*, 210(1), 103-114. doi: 10.4049/jimmunol.2200685

Leus, I.V., Weeks, J.W., Bonifay, V., Shen, Y. …& Zgurskaya, H. I. (2022). Property space mapping of Pseudomonas aeruginosa permeability to small molecules. *Sci Rep* 12, 8220. doi: 10.1038/s41598-022-12376-1

Ajina, R., Malchiodi, Z. X., Fitzgerald, A. A., Zuo, A., Wang, S., Moussa, M., Cooper, S. J., Shen, Y., ... & Weiner, L. M. (2021). Antitumor T-cell Immunity Contributes to Pancreatic Cancer Immune Resistance. *Cancer Immunol Res*, 9(4), 386-400. doi: 10.1158/2326-6066.CIR-20-0272

Acharya, A., Agarwal, R., Baker, M. B., Baudry, J., Bhowmik, D., Boehm, S., ... & Zanetti-Polzi, L. (2020). Supercomputer-based ensemble docking drug discovery pipeline with application to COVID-19. *Journal of chemical information and modeling*, 60(12), 5832-5852. doi: 10.1021/acs.jcim.0c01010

Ajina, R., Zuo, A., Wang, S., Moussa, M., Cooper, C. J., Shen, Y., ... & Weiner, L. M. (2020). Immune Selection Pressure Contributes to Pancreatic Cancer Immune Evasion. *bioRxiv*. doi: 10.1101/2020.06.15.151274

Xie, Y., Wei, Y., Shen, Y., Li, X., Zhou, H., Tai, C., ... & Ou, H. Y. (2018). TADB 2.0: an updated database of bacterial type II toxin–antitoxin loci. *Nucleic Acids Research*, 46(Database issue), D749. doi:10.1093/nar/gkx1033.