UNIVERSITY OF CALIFORNIA
SANTA CRUZ

STRUCTURE-BASED MODELING OF PEPTIDE/MHC-I COMPLEXES

A dissertation submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY
in
BIOMOLECULAR ENGINEERING AND BIOINFORMATICS

by

Santrupti Nerli

December 2021

The Dissertation of Santrupti Nerli is approved:

Professor David Haussler, Chair

Professor Rebecca DuBois

Professor Nikolaos G. Sgourakis

Peter Biehl
Vice Provost and Dean of Graduate Studies

Contents

List of Figures

List of Tables

STRUCTURE-BASED MODELING OF PEPTIDE/MHC-I COMPLEXES

Santrupti Nerli

Abstract

Major Histocompatibility Complex (MHC) class I molecules present intracellular peptides for surveillance by cytotoxic T cells. To fully understand the mechanism of T cell recognition and prioritize different therapeutic targets, it is of practical relevance to first characterize the surface features of peptide/MHC (pMHC) complexes which requires high-resolution structural models. In this work, I present RosettaMHC, a high-throughput pMHC homology modeling approach, which utilizes structural templates from the Protein Data Bank (PDB) selected using several strategies. The accuracy of a peptide backbone is primarily dependent on the selection of a closest template. Here, we have explored template selection based on (i) the peptide sequence, and (ii) the MHC groove sequence which allow us to obtain models within 1.5 Å and 1 Å respectively from their native X-ray structures. In addition, we have added Artificial Neural Network (ANN) models to filter structures that are potentially inaccurate thereby increasing throughput while helping us select models within 1 Å from their corresponding natives. We applied RosettaMHC to model structures of (i) anaplastic lymphoma kinase neoepitopes found in neuroblastoma cancer patients in complex with 2,904 HLA alleles to identify putative alleles that may display these peptides, and (ii) SARS-CoV-2 peptide/HLA-A*02:01 complexes that can be utilized as candidates to generate pMHC tetramers to probe T cells and possibly as vaccine targets. Through this work, we want to establish that, for the most frequent alleles, we have sufficient knowledge from the existing databases to predict sub-angstrom accuracy models thereby reducing the efforts required to experimentally determine pMHC structures.

Alongside RosettaMHC, I (i) helped develop resonance assignment methods that are used to interpret NMR data, (ii) worked with alternative protocols in Rosetta to solve structures of biologically important protein targets of varying sizes and complexities with the help of sparse NMR data, and (iii) built a method to estimate side chain conformational entropy using boolean satisfiability and applied it to understand dynamics of the therapeutic cytokine, Interleukin-2, which helped characterize a minor conformational state that is relevant for drug design.

Although the Rosetta energy function together with ANN-based scores help us eliminate

inaccurate peptide backbones, poor templates may still be locally optimal, leading to imprecise modeling results in some cases. In the future, we want to employ our satisfiability-based side chain packing method towards a more accurate peptide backbone identification. I believe that this work can provide a rational approach for high-throughput modeling of pMHC targets with desired recognition features which are relevant for cancer immunotherapy and infectious diseases.

Dedicated to my husband, Prasannakumar, who supported me throughout my tenure at UCSC; to my son, Rian, the light of my life; and my parents, without whom I wouldn't be here.

Acknowledgements

I would like to extend special thanks to my advisor Prof. Nikolaos G. Sgourakis for introducing me to the field of NMR and computational biology, chemistry; teaching me everything including how to think about a problem, writing research papers, presenting my work, and most importantly about being resilient. I cannot imagine myself here after 5 years without Nik's constant support. I would also like to acknowledge my committee members, Prof. David Haussler for his help through the final years of my PhD, and Prof. Rebecca DuBois for teaching me the fundamentals of protein engineering, which has fueled more appreciation for the work I do. In addition to my committee members, I would like to express my gratitude to Andrew McShan and Sarah Overall for insightful discussions over the years, reviewing my works, helping me improve my presentation and writing skills and above all, making me appreciate myself often. I would like to thank Viviane Silva De Paula and Hau Truong for sharing their knowledge on various topics of NMR and their encouragement to push through my PhD. I would like to thank Prof. Sami Khuri and Prof. Natalia Khuri for motivating and preparing me to pursue PhD. Lastly, I would like to thank all the people in the list below who have helped me in one way or another during my PhD.

Family Members

Shanta Yabannavar, Rudragouda Patil, Kasturi Patil, Vijaylaxmi Kelageri, Nagaraj Kelageri, Deepa Sangolli, Vishwanath Patil, Apoorva Karekal, Sandeep Nerli, and Pooja Nerli.

Lab Members

Yi Sun, Sagar Gupta, Mohamad Alani, Tyler Florio, and Nico Gonzalez.

Collaborators

Thomas Evangelidis, Enrique Marcos, Jason Fernandes, Maximilian Haeussler, Aimee Marceau, Seth Rubin, Mark Yarmarkovich, John Maris, Haotian Du, PoSsu Huang, David Baker, Andrew Leaver-Fay, and Frank DiMiao.

Members at UCSC

Theo-Alyce Gordon, Alison Lindberg, Eric Shell, Ramon Berger and ITS/ADC Staff at UCSC.

Other members

David Flores Solis, Kostas Zampitakis, Ben Sherman, Marco Chammoro, Hailey Wallace, Joey Toor, Dimitris Achlioptas, and Eleni Dimitripolou.

Chapter 1

Introduction

The adaptive immune system in jawed vertebrates comprises of cell types that specialize in eliminating distinct pathogenic microorganisms [53]. The main cell types in the adaptive immune system are T cells and B cells. These cells utilize receptors (T cell receptors or TCRs and B cell receptors or BCRs) expressed on their surface to recognize foreign antigens [2]. It is estimated that there are approximately 10 million unique TCR and BCR molecules in every human, each capable of recognizing moderately dissimilar antigens [2]. T cells differentiate into either CD4+ T helper cells (Th) or CD8+ cytotoxic T cells (CTLs or cytotoxic T lymphocytes) to provide cell-mediated immunity, whereas B cells differentiate into plasma cells that produce antibodies thus providing antibody-mediated immunity. The TCRs on the surface of Th or CTLs recognize pathogenic antigens cleaved into short peptides and presented to them by major histocompatibility complex (MHC) molecules, also referred to as human leukocyte antigens (HLA) in humans. There are two types of MHC molecules (coded by the MHC loci), class I and class II MHC molecules. Endogenous antigens within the cell are presented on all nucleated cells using class I MHC molecules, whereas exogenous antigens originating from extracellular self or foreign proteins are displayed by class II MHC molecules. TCRs of CTLs and Th cells recognize peptides presented by class I and II MHC molecules, respectively. Since MHC molecules play a central role in deriving immune responses, understanding their structure, function and peptide repertoire is the key to developing peptide-based vaccines and therapies that utilize host's immune system to fight infectious diseases and cancer. This thesis discusses only human MHC class I (or HLA-I) molecules, and I use the terms MHC and HLA synonymously.

1.1 MHC Class I Antigen Processing and Presentation Pathway

The processing and presentation of peptides on the surface of the cell by class I MHC molecules involves a series of events highlighted in Figure 1.1 [69]. In this pathway, proteins expressed within a cell will be degraded in the cytosol by the proteasome. The proteasome digests polypeptides into smaller peptides of 5–25 amino acids in length. The transporter associated with antigen presentation (TAP) molecule selects peptides of lengths up to 15 amino acids and transports the peptides into the endoplasmic reticulum (ER). Within the ER, endoplasmic reticulum aminopeptidase (ERAP) further cleaves the peptides at the N-terminal regions to form smaller fragments of lengths typically from 9-12 amino acids (larger or smaller length peptides can escape ERAP filter) [133]. Here, the partially folded MHC molecules stabilized by the ER-hosted chaperones, will complete folding by binding specific high-affinity peptides. The peptide/MHC (pMHC) molecule is then transported to the surface of the cell through the golgi apparatus for surveillance by the T cells. While all the events in processing and presentation pathway play an important role in displaying a peptide on the cell surface, peptide binding to the MHC molecule is the most selective event among them [131].

1.2 MHC restriction and T cell immunity

MHC molecules are highly polymorphic since they are constantly under evolutionary pressure to acquire new variants that modify their function of antigen presentation thus altering their peptide repertoire [90]. Currently, there over 13,000 class I HLA alleles (encoded by HLA-A, -B and -C genes) identified in the International Immunogenetics Project (IMGT)/HLA database [97]. The A, B and C alleles can be clustered into 12 supertypes based on their overlapping peptide specificities as shown by the unrooted phylogenetic tree in Figure 1.2 [96].

The $\alpha\beta$ TCR molecule (which recognizes class I MHC molecules) is a dimer generated from somatic gene rearrangement of variable (V), diversity (D), joining (J) and constant (C) gene segments [105]. Such a rearrangement introduces nucleotide insertions or deletions at the V(D)J junctions leading to a large TCR repertoire. This diversity in the repertoire is contributed by six complementarity determining regions (CDRs) in TCRs shown by the colored loops in

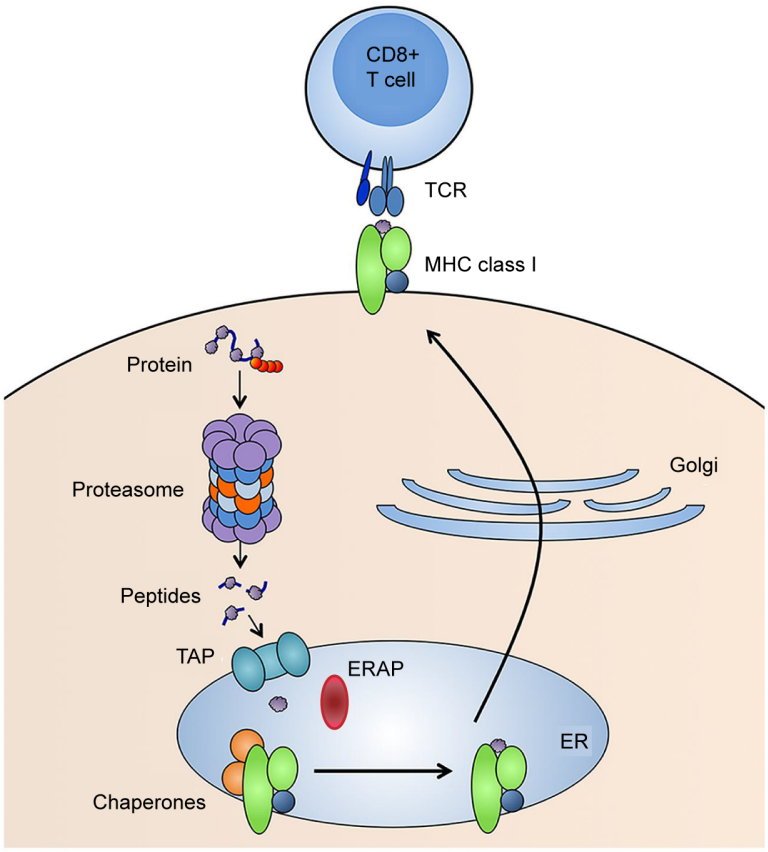


Figure 1.1: MHC class I antigen processing and presentation pathway. Figure adopted from [69] and modified to include relevant information.

Figure 1.3A. The CDRs associate with both the peptide and the MHC molecule (Figure 1.3). The MHC class I molecules generally binds peptides of 8–14 amino acids in length (Figure 1.3B). There are typically two conserved sequence positions which anchor the peptide into the MHC molecule [105]. The longer peptides bulge out in the center (Figure 1.3B). One of the important feature of TCRs is that they do not randomly recognize any combination of the pMHC complex, but rather complexes formed by the host's HLA haplotype (or a set of HLA genes inherited by the hosts from their parents; more specifically every human inherits two HLA haplotypes consisting of three HLA alleles from each parent). [43].

T cells originate in the bone marrow and mature in the Thymus. Due to positive selection, cells which are able to bind MHC class I molecules with at least a weak affinity survive (Figure 1.4) [28]. During a process called 'death by neglect', T cells which would be non-functional due to an inability to bind MHC molecules are eliminated (Figure 1.4) [28]. T cells that exhibit

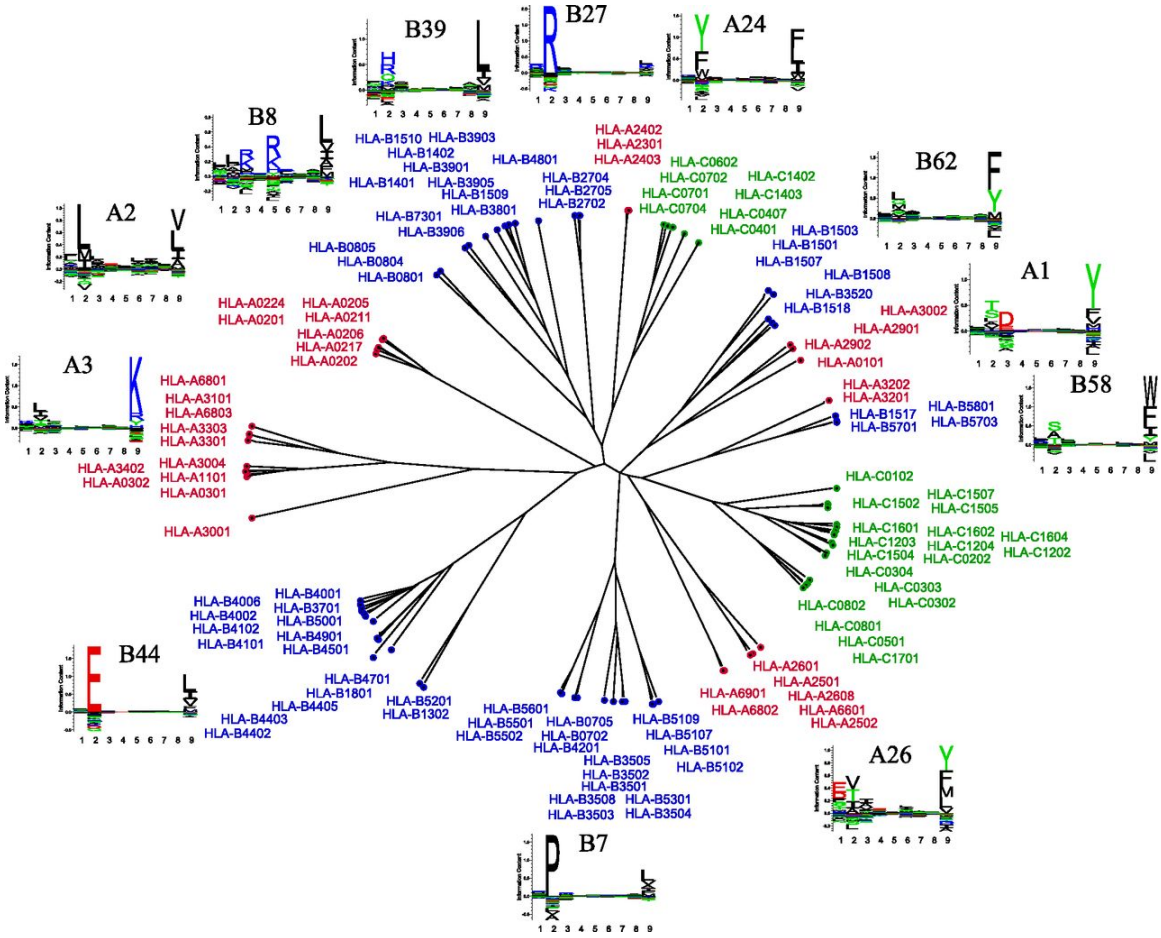


Figure 1.2: Clustering of 121 HLA-A, -B and -C alleles according to their binding motifs using NetMHCpan [45]. Figure adopted from [96].

high affinity for self peptides or MHC molecules are also eliminated due to negative selection since, these cells would direct immune responses towards self-proteins in the periphery (Figure 1.4) [28]. A T cell recognizes the self MHC molecules (from the host's haplotype) when they bind non-self peptides, but it mounts a response only when the peptide is bound to a particular alleleic MHC molecule, and this phenomenon is referred to as MHC restriction [43]. Through this mechanism, a T cell repertoire tolerates self peptides but is prepared to identify foreign peptides. Therefore, characterizing the pMHC-TCR repertoire becomes important to understand adaptive immunity.

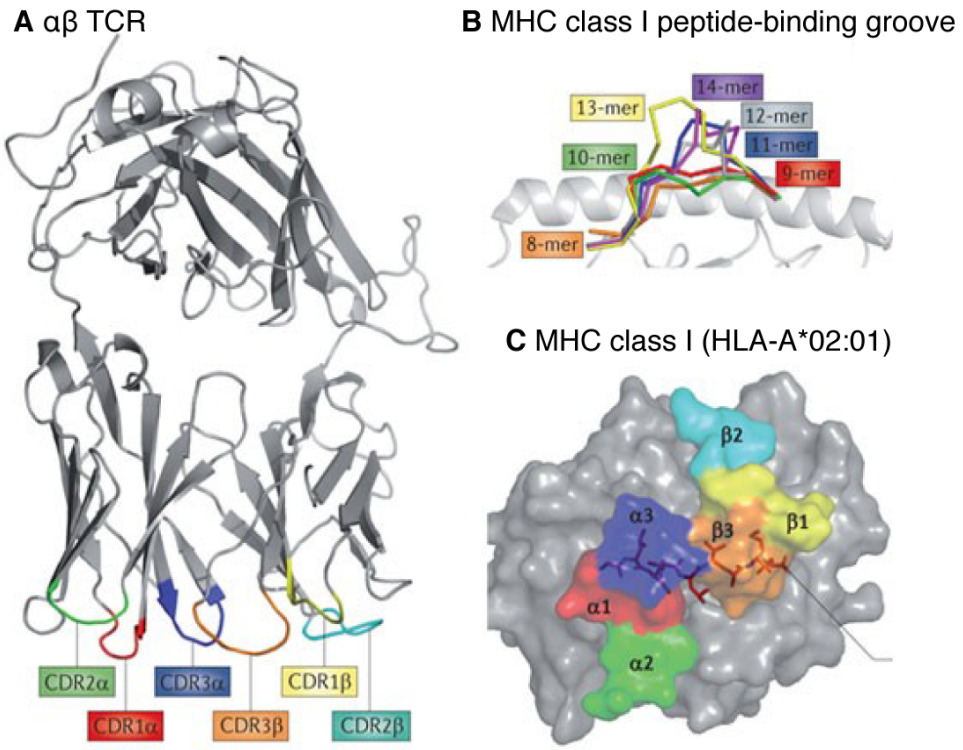


Figure 1.3: Diversity of T cell receptors. (A) Structure of $\alpha\beta$ T cell receptor showing six variable regions or complementarity determining regions (CDRs) (colored loops). (B) Structure of the MHC binding groove highlighting that the MHC class I molecules can accommodate peptides of varying lengths. The restricted ends of the MHC molecule causes the peptides to extend the conformation in the center (bulge). The colors indicate different lengths of the peptides. (C) Surface of HLA-A*02:01 (gray) presenting GLCTLVAML peptide (sticks) from Epstein-Barr virus. The colored regions on the surface of pMHC molecule indicate contacts made by the six corresponding CDR loops of the T cell receptor, AS01 TCR96. Figure adopted from [105] and modified.

1.3 Structural insights into peptide/MHC binding and TCR recognition

The MHC class I molecule is a heterodimer and consists of a glycosylated amino acid heavy chain (α chain), and a light chain (β2-microglobulin or β2m) (Figure 1.5). The MHC-I molecules display peptides of specific lengths (8-14 amino acids) (Figure 1.5). The heavy chain consists of α1, α2, α3 extracellular domains each comprising approximately 90 residues, transmembrane and intracellular (cytoplasmic tail) regions (Figure 1.5). The α1 and α2 domains form the binding groove where a peptide is nestled within six distinct pockets (A-F) inside the MHC binding groove

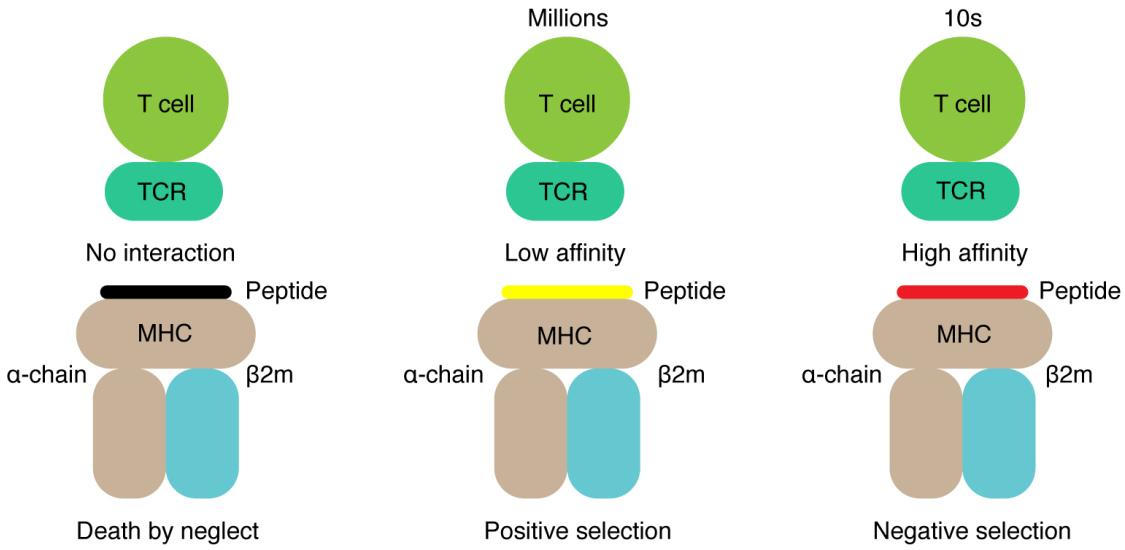


Figure 1.4: Selection and maintenance of TCR repertoire. Interactions between pMHC and TCRs facilitate T cell survival, maturation and death. If there are no interactions between the pMHC molecules and the TCR, then the non-functional T cell undergoes death (a process called death by neglect) (Left). Millions of T cells that have TCRs which exhibit low affinity interaction with pMHC survive and proliferate (Middle). Due to negative selection, T cells that bind pMHC molecules with high affinity are eliminated (Right). The color coding of the units is: MHC α-chain: wheat, MHC β2m: cyan, Peptides: black, yellow and red, T cell: green and TCR: dark green. Figure inspired from [28].

as shown in Figure 1.6 [15]. The B and F pockets in the binding groove are of special importance since the peptide anchors deeply within these pockets thereby reducing variability of preferred amino acids at these positions (Figure 1.6). Typically, peptides of lengths 8, 9 and 10 (commonly displayed by HLA-I) have their anchors at positions 1 and 8 (8 mer), 2 and 9 (9 mer), and 2 and 10 (10 mer) (N-terminal residues of the peptides fit in the B pocket whereas C-terminal residues fit in the F-pocket). The α3 domain closer to the transmembrane region is an important binding site for the CD8 co-receptor (Figure 1.5), which is responsible for co-stimulation of the immune response.

Each HLA allele can potentially display 10^3 to 10^6 peptides. TCRs that recognize these pMHC complexes are highly cross reactive meaning a single TCR can recognize many pMHC molecules [132, 12]. For instance, for a typical peptide of length 9, there are half a trillion possible peptides but only a few million TCRs, so every TCR must recognize many peptides making TCRs inherently cross-reactive. Therefore, the space of pMHC-TCRs is large. However, there are hundreds of pMHC and pMHC-TCR structures available in the databases [13]. From these

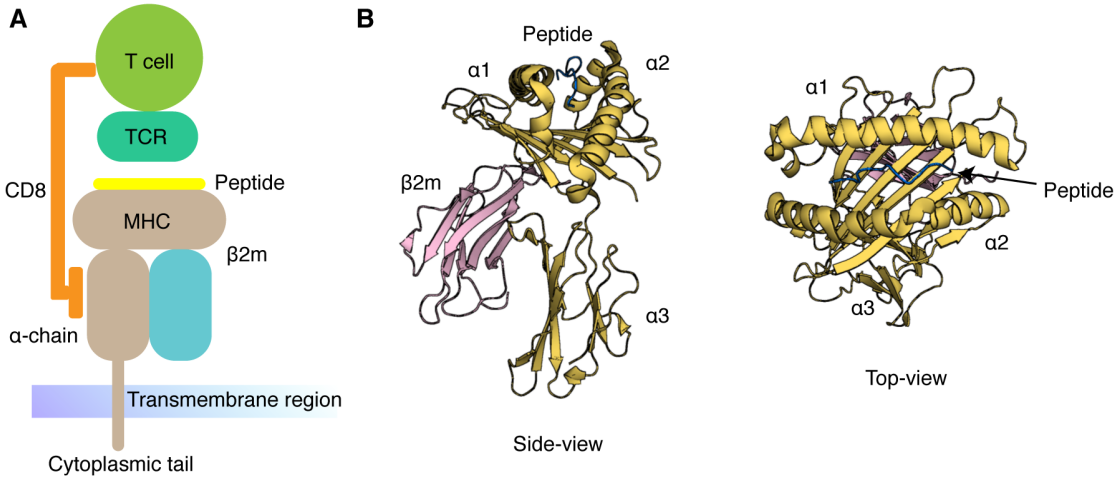
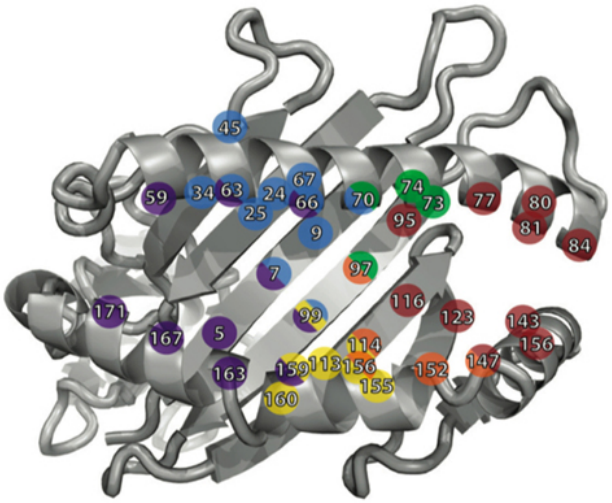


Figure 1.5: Structure of MHC class I molecule. (A) Cartoon representation of MHC class I molecule and its interaction with TCR and CD8 molecules of T cells. The color coding of the units is: MHC α-chain and cytoplasmic tail: wheat, transmembrane region: blue, MHC β2m: cyan, Peptide: yellow, T cell: green, CD8: orange, and TCR: dark green. (B) Side-view (Left) and top-view (right) of a structure of class I MHC molecule (PDB ID: 6AT9). α1, α2 and α3 domains (or α chain) is highlighted in yellow. β2m and peptide are shown in pink and blue respectively.

databases, we have a general understanding of how pMHC-TCRs interact thereby helping us in identifying residues in the peptide that anchor the TCR. From previous studies, we have learned that TCRs are oriented diagonally on the pMHC surface as shown in Figure 1.3C (colored regions) [13]. Using the structural information and surface of the pMHC complex, we can learn about TCR cross-reactivity which becomes useful when engineering new TCRs to fight diseases.

1.4 Peptide-based vaccines

Identification and characterization of a single or a pool of immunodominant peptides (or epitopes) derived from foreign pathogens or cancer tissues is an important step in designing peptide-based vaccines to fight cancer and infectious diseases [65]. Generally, synthetic peptides are administered to healthy individuals with the help of adjuvants, which are engulfed by the dendriatic cells (DCs) in the periphery [65]. The DCs then trigger the activation of T cells by presenting the peptides via MHC molecules.



Pocket	Peptide Position accomodated
Pocket A	1
Pocket B	2
Pocket C	6
Pocket D	3
Pocket E	7
Pocket F	C-terminus

	B pocket specificity		F pocket specificity	
Supertype	**Description**	**a.a. residues***	**Description**	**a.a. residues***
A01	Small and aliphatic	ATSVLIMQ	Aromatic and large hydrophobic	FWYLIM
A02	Small and aliphatic	ATSVLIMQ	Aliphatic and small hydrophobic	ATSVLIMQ
A03	Small and aliphatic	ATSVLIMQ	Basic	RHK
A24	Aromatic and aliphatic	FWYLIVMQ	Aromatic, aliphatic and hydrophobic	FWYLIVM
A01 A03	Small and aliphatic	ATSVLIMQ	Aromatic and basic	YRK
A01 A24	Small, aliphatic and aromatic	ASTVLIMQFWY	Aromatic and large hydrophobic	FWYLIM
B07	Proline	P	Aromatic, aliphatic and hydrophobic	FWYLIVM
B08	Undefined		Aromatic, aliphatic and hydrophobic	FWYLIVM
B27	Basic	RHK	Aromatic, aliphatic, basic and hydrophobic	RHKFWYLIVM
B44	Acidic	DE	Aromatic, aliphatic and hydrophobic	FWYLIVM
B58	Small	AST	Aromatic, aliphatic and hydrophobic	FWYLIVM
B62	Aliphatic	LIVMQ	Aromatic, aliphatic and hydrophobic	FWYLIVM

*a.a.residues = amino acid residues

Figure 1.6: MHC binding pockets (A-F) are shown by colored residues on the α1 and α2 domains. The color codes and corresponding peptide residues that fit inside each pocket are shown on the right. The amino acid preference for each supertype within the most conserved B and F pockets are shown in the table below. Figure adopted from [15].

The diversity of HLA types in human population poses a significant challenge in designing a universal peptide-based vaccine candidate/s for any disease. However, previous studies have demonstrated that there is significant overlap in the peptide-binding specificities of HLA supertypes (Figure 1.2). The HLA supertypes A02, A03 and B07 are known to cover 90% of the world population [134]. Therefore, plausible vaccine candidates are peptides that bind HLA supertype families covering large population and elicit effective immune responses.

1.5 Thesis statement

MHC class I antigens present endogenous peptides from tumor or infected cells to the cytotoxic T lymphocytes for surveillance. To design and develop vaccines and therapies to treat cancer and infectious diseases, we need to understand the mechanism of T-cell receptor recognition and peptide immunogenicity. We can identify and elucidate TCR recognition features with the help of high-resolution peptide/MHC-I structures. However, solving peptide/MHC structures experimentally using X-ray crystallography or NMR spectroscopy is tedious, time-consuming and has limitations. Furthermore, modeling of peptide/MHC structures with high accuracy is still a challenge in the field despite the availability of a number of docking approaches. This thesis describes a high-throughput peptide/MHC homology modeling method using several template selection and scoring strategies that generates sub-angstrom structures for a vast majority of the cases shown in the future sections.

Chapter 2

Background

Sequence-based methods are widely applied to identify and characterize immunodominant epitopes. Complementary to sequence-based approaches, structure-based methods can provide detailed molecular basis for the pMHC and pMHC-TCR interactions. Here, we review and highlight the limitations of popular sequence-based and structure-based approaches which are used to predict binding affinities, free energies and structural models.

2.1 Sequence-based methods

Early methods to predict binding of peptides to MHC molecules utilized motifs or scoring matrices derived from sequences. A well-known motif-based prediction model is provided by SYFPEITHI [94], which estimates binding affinities using MHC class I eluted-ligand data (EL; eluted ligands are peptides processed and presented naturally and can be eluted from the antigen presenting cells; such ligands can be characterized using mass spectrometry methods and then used as training data for prediction methods) and binding assay data obtained by single amino acid substitution at each peptide position. These methods were subsequently replaced by more sophisticated statistical algorithms and machine learning approaches trained on large complementary experimental datasets and are used extensively in the field [78].

The MixMHCpred [10] computational framework utilizes data generated by mass spectrometry analysis of eluted pMHC peptides. Subsequently, the eluted ligands are used to construct multiple position weight matrices of specified lengths (9 and 10) from which peptide motifs for

each HLA allele can be drawn. While MixMHCpred uses a linear matrix approach, alternative state-of-the-art methods such as NetMHCpan4.0 [45] and MHCflurry [83] apply simple single layer feed forward neural network models trained using eluted ligand data and binding affinity data available in the databases such as Immune Epitope Database (IEDB) [123] which has 1,200,000 pMHC epitopic peptides out of which 300,000 are from binding and the remaining are from ligand elution assays. The major limitations of sequence-based approaches are biases introduced by the availability of limited data containing (i) cysteine residues in the peptide sequences, (ii) low prevalence of a vast majority of MHC allotypes and hence little to no experimental data for such allotypes and (iii) post-translational modifications in the MHC binding groove, leads to low predictive power of these methods [78].

2.2 Structure-based methods

An advantage of structure-based methods is that they do not require allotype specific training datasets but rather they rely on physicochemical properties to provide information about pMHC interactions [6]. Several methods are available that can either predict models of pMHC or binding affinities from first principles. These approaches utilize general rules including (i) sampling of peptide backbone (ii) optimizing amino acid side chain degrees of freedom, and (iii) scoring using physically realistic energy functions. [6]

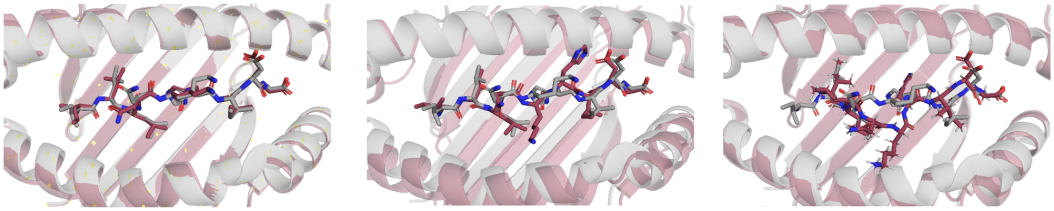
A number of structure-based approaches implement molecular docking, where a peptide is docked inside the binding groove of the MHC molecule [6]. The difficulty of obtaining an accurate model from these approaches stem from the flexibility of peptide backbones and side chains that needs to be accounted for. Therefore, to sample diverse peptide conformations, these methods utilize (i) rigid docking (graph-based algorithms to identify shape complementarity between peptide and MHC molecule), (ii) sampling of all degrees of freedom of the peptide backbone and side chains (computationally expensive), or (iii) heuristic approaches that employ Monte Carlo-based search-like methods [6]. The sampled conformations are ranked using scoring functions and can be used to guide the search for more accurate pMHC structures. Approaches such as pDOCK [51] and FlexPepDock [61] start with structural templates or homology models followed by sampling of the peptide backbones and refinement using Monte Carlo procedure. Similarly, MHCSim [113] and DockTope [70] utilize closest templates based on pMHC sequence to per-

form docking using either molecular dynamics (MD) simulations or AutoDock Vina (that employs genetic algorithms) [118]. A landmark study (henceforth referred to as Yanover and Bradley) [129] performs flexible peptide backbone docking simulations which consist of low-resolution and high-resolution modeling phases. During the low resolution phase, the peptide backbone is constructed outwards from the anchor positions which are explored during docking moves followed by loop closure. The docked peptide is refined in the high-resolution phase in the Rosetta force field. While Yanover and Bradley method can be used to predict peptide binding landscapes for HLA-A and -B alleles, it is not high-throughput and lacks the resolution necessary to predict sub-angstrom accuracy models. Alternative approaches such as MFPred [101] and GradDock [54] use scoring functions enhanced by mean field theory or redefined weights for score terms to rank peptides. GradDock is shown to perform better compared to Yanover and Bradley method. All these methods aim to predict pMHC complexes within 1.5 to 2 Å and peptide backbones within 1.5 Å all-atom or Cα RMSD from the native X-ray structure (RMSD or Root-Mean Squared Deviation is a measure to show how close a model is to its corresponding X-ray structure, a value of 0 indicates that the backbone and side chain conformations are recovered completely in the model relative to the X-ray structure, higher values indicate deviations from the X-ray structure; see Figure 2.1). Here, we aim to model sub-angstrom pMHC structures since models that have RMSD values above 1 Å have an altered TCR surface due to inaccuracies in the peptide backbone and side chain conformations (Figure 2.1).

Recently, structure-based methods have been used in conjunction with sequence-based approaches to improve the prediction of pMHC binding specificities and modeling accuracy [7]. One such study [7] applies NetMHCpan4.0 [45] to select high-affinity peptides and homology models them using MODELLER [24] followed by all-atom refinement using FlexPepDock [61].

2.3 Conclusions

Sequence-based methods are being developed continuously to predict peptides binding to MHC-I molecules or the immunogenicity of pMHC-I complexes. The accuracy of these methods are improving due to the availability of experimental data originating from elution assays, binding affinity assays, and mass spectrometry and the use of advanced machine learning techniques. Despite their high accuracy, these methods cannot provide molecular insights into the pMHC



RMSD: 0.3 Å
Correct peptide backbone and minor changes in side chain conformations

RMSD: 1.0 Å
Closer peptide backbone and some incorrect side chain conformations

RMSD: 2.3 Å
Incorrect peptide backbone

Figure 2.1: RMSD values capturing variations in peptide backbone and side chain conformations. Overlay for structural model (magenta) and X-ray structure (grey) for showing RMSD values of (Left) 0.3 Å which indicates that the peptide backbones are very close and that most of the side chain conformations are correct, (Middle) 1.0 Å representing peptide backbones that are close and have some incorrect side chain conformations, and (Right) 2.3 Å which highlights that the peptide backbones are incorrect. RMSD values are calculated for heavy atoms, N, C, CA, O, of the peptide backbone.

binding which is an important step in early discovery of drug targets or therapies. Alternatively, structure-based methods can offer details at a molecular level, although their development is limited relative to sequence-based methods due to their time complexity. Currently available approaches including docking, homology modeling or loop closing, model peptide backbone conformations with RMSD values up to 2 Å. However, a 2 Å pMHC model can showcase a different surface compared to a native structure and hence limit our ability to study T-cell receptor recognition in silico (Figure 2.1). Therefore, in this thesis, we develop a new structure-based modeling method that attempts to address accuracy and throughput.

Chapter 3

Protein Structure Modeling using Rosetta and NMR Data

[Some of the text and figures in this chapter have been published with the following citations:

- 4D-CHAINS: Thomas Evangelidis, **Santrupti Nerli**, Jiří Nováček, Andrew E. Brereton, P. Andrew Karplus, Rochelle R. Dotas, Vincenzo Venditti, Nikolaos G. Sgourakis, and Konstantinos Tripsianes. Automated NMR resonance assignments and structure determination using a minimal set of 4D spectra. *Nature Communications*, 9(1):384, January 2018.
- MAUS: **Santrupti Nerli**, Viviane S. De Paula, Andrew C. McShan, and Nikolaos G. Sgourakis. Backbone-independent NMR resonance assignments of methyl probes in large proteins. 98 *Nature Communications*, 12(1):691, January 2021.
- CS-Rosetta: **Santrupti Nerli** and Nikolaos G. Sgourakis. CS-ROSETTA. *Methods in Enzymology*, 614:321–362, 2019
- **Santrupti Nerli**, Andrew C. McShan, and Nikolaos G. Sgourakis. Chemical shift-based methods in NMR structure determination. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 106-107:1–25, June 2018.

]

In this chapter, I introduce NMR spectroscopy, resonance assignment problem and my contributions to two methods 4D-CHAINS and MAUS that are used to interpret NMR data. I describe in detail, our previous work carried out using an open-source molecular modeling software suite, Rosetta where, we solved structures of seven important protein targets with the help of sparse NMR data. Lastly, I show a solution NMR structure of a pMHC complex that was deposited recently and how we can use modeling methods to obtain accurate pMHC structures thereby reducing the need to determine structures experimentally.

3.1 Nuclear Magnetic Resonance Spectroscopy

Proteins can be studied using NMR spectroscopy, for instance, we can determine their 3D-structures, and study their interactions with ligands and drug-like molecules [39, 3]. The nuclei of atoms from isotopes such as ^{1}H, ^{13}C and ^{15}N possess magnetic spin properties. To examine proteins using NMR, isotopically labelled protein sample is placed in an external magnetic field. A separate radio frequency is applied perpendicular to the external magnetic field which causes the nuclear spins to precess at characteristic frequencies. The differences in these frequencies from a reference signal gives rise to a measure called chemical shift (or resonance) [39].

Chemical shifts inform us about the local environment of nuclei using which we can predict the secondary structures of peptide backbone, model orientation of side chains, decipher protein dynamics and hydrogen bonding network [74]. Currently, many methods use chemical shift measurements to determine the solution structures of proteins [74]. Typically, such methods sample small peptide fragments derived from known structures and filter them by comparing experimental and back-calculated chemical shift values. The fragments are assembled using optimization protocols, and force fields biased by chemical shift values. Alongside chemical shifts, additional information from NMR experiments such as nuclear Overhauser effect (NOEs obtained from NOESY experiments) measurements provide distance restraints between atoms that are closer in space, and residual dipolar coupling (RDC) values which help determine motion/orientation of domains in multi-domain proteins [74].

3.1.1 Resonance assignment problem

The chemical shifts measured using a protein sample and NMR instrument generally are in the form of peaks in a 2D plane with shift values (in parts per million or ppm) of correlated ^1H and ^{13}C or ^1H and ^{15}N along the axes. Assigning each peak to its corresponding CH or NH group of an amino acid in the sequence of a protein is the resonance assignment problem. The resonances are assigned using procedures that either utilize solved structures or multiple complementary NMR experiments that provide information of nuclear spins that are either connected through bonds or space. Below, I discuss two recent resonance assignment methods, 4D-CHAINS and MAUS, I had the opportunity to contribute. The 4D-CHAINS method assigns resonances of all atoms using a small set of NMR spectra and does not require a structure however it is limited by size (up to 30kDa) whereas MAUS assigns resonances of methyl groups in residues (such as M, A, I, L, V, and T) in larger proteins (sizes that can be studies by NMR) and requires NMR spectra together with a crystal structure.

Terminologies necessary to understand 4D-CHAINS and MAUS

HSQC - Heteronuclear Single Quantum Coherence (HSQC) is a heteronuclear 2D experiment that captures correlation between two different nuclei such as a ^1H and a coupled nucleus, either ^{13}C or ^{15}N. Since the correlated nuclei can be plotted on a 2D plane with x-axis indicating ^1H spectrum and y-axis indicating ^{13}C spectrum, these plots are used to decipher the residue types of both ^1H and ^{13}C values.

HMQC - Heteronuclear Multiple Quantum Coherence (HMQC) is a heteronuclear 2D experiment and also captures correlation between coupled ^1H and ^{13}C nuclei or ^1H and ^{15}N nuclei. The primary difference between HSQC and HMQC is that in HMQC, magnetization of both ^1H and ^{13}C nuclei are allowed to evolve, whereas in HSQC, the magnetization of only ^{13}C nucleus is allowed to evolve.

NOESY - Nuclear Overhauser Effect Spectroscopy (NOESY) allows us to determine signals from ^1H nuclei that are closer in space.

TOCSY - Total Correlation Spectroscopy (TOCSY) allows us to determine signals from coupling nuclei that are connected through bonds.

4D-CHAINS

Nuclear magnetic resonance (NMR) structure determination relies on recording a network of nuclear Overhauser enhancement (NOE) restraints from multidimensional spectra [128]. Obtaining near-unambiguous assignments of long-range NOEs is challenging due to substantial overlap in the spectra which becomes more pronounced for larger proteins. This is typically addressed through first establishing the chemical shift assignments of backbone and sidechain atoms using multiple (6–10) triple-resonance spectra [41, 47], which are then used as anchors to guide the assignment of NOEs during iterative structure refinement [33]. State-of-the-art tools such as FLYA [104], PINE [8] and UNIO [32] can automate the resonance assignment and structure determination process. In principle, recording a smaller number of higher dimensional spectra can provide a complementary approach to increase signal dispersion and resolve ambiguities [46]. With the emergence of non-uniform sampling and reconstruction methods, such datasets can be recorded in reasonable time [48]. Recent approaches for automated resonance assignments based on three- and four-dimensional (3D and 4D) NOE data make use of a known structure to guide the assignment process [116, 92]. However, for de novo structure determination, further development is needed to perform resonance assignments at the high levels of completeness and correctness that are required for NOE data-driven structure determination.

Towards developing 4D-CHAINS, we recorded for four different protein targets of size from 15.5 to 27.3 kDa, a 4D HC(CC-TOCSY(CO))NH, and a 4D ^{13}C,^{15}N edited HMQC-NOESY-HSQC (HCNH) experiment. The largest protein target of size 27.3 kDa was chosen based on its apparent correlation time of 15 ns that still allows for TOCSY transfer to occur. We also recorded a 4D ^{13}C, ^{13}C edited HMQC-NOESY-HSQC (HCCH) experiment to further assist in structure determination. To address the assignment problem, 4D-CHAINS uses 2D probability density maps of correlated ^{13}C–^{1}H chemical shifts to effectively identify possible spin systems. In particular, 4D-CHAINS combines sequential information present in the 4D-HCNH TOCSY and intraresidue information present in 4D-HCNH NOESY ^{13}C–^{1}H planes, respectively, by clustering TOCSY or NOESY peaks to Amino Acid Index Groups (AAIGs) via their common ^{15}N–^{1}H frequency (Figure 3.1A). 4D-CHAINS computes probability scores at several steps (amino acid-type prediction, sequential AAIG relations based on TOCSY–NOESY connectivities, alignment of peptides to the protein sequence) to yield a confidence score for a given AAIG being as-

signed to a specific protein residue. Finally, 4D-CHAINS uses an Overlap Layout Consensus (OLC) assembly approach adopted from genome assembly to match continuous AAIG segments along the protein sequence. The final assignment solutions are consistent with both the joined probability score and the OLC model.

A uniform 4D-CHAINS protocol was applied to four targets (Figure 3.1B). The algorithm mapped correctly all AAIGs to the respective protein sequences with >95% completeness. Next, we evaluated the performance of assignments obtained using 4D-CHAINS in driving Rosetta structure determination of these targets. Here, we used AutoNOE-Rosetta (described in the following section titled ”Rosetta Software Suite”), a highly parallelizable iterative algorithm run on a computer cluster to perform assignment of long-range NOEs alongside the structure determination process.

The solution NMR structures determined using 4D-CHAINS/AutoNOE-Rosetta shown in Figure 3.1B have been deposited in PDB under codes 5WOT, 5WOX, 5WOY, and 5WOZ. The full details of this work is available at [25].

Methyl assignments using Satisfiability

The use of methyl probes has opened new avenues for the application of nuclear magnetic resonance (NMR) methods to study large molecular machines [98]. The signal enhancement offered by methyl-transverse relaxation optimized spectroscopy (TROSY) techniques [119] and a suite of experiments for quantitative characterization of protein dynamics occurring over a broad range of timescales [103] have rendered methyl-based NMR a formidable tool for detailed mechanistic studies of important biological systems [109]. The main bottleneck in applications of methyl-based NMR is obtaining confident resonance assignments. In the conventional approach, backbone assignments are first established using triple-resonance experiments [47]. Then, methyl resonances are connected to the backbone using either methyl out-and-back experiments [120] or, more commonly, using 15N and 13C edited amide-to-methyl nuclear Overhauser effect (NOE) measurements [99]. However, in the absence of previously established backbone assignments, deriving confident assignments for methyl resonances remains a challenge. Although site-directed mutagenesis of individual methyl-bearing residues [68] provides unambiguous assignments, the laborious, costly, and time-consuming nature of this approach limits applications to study larger, multi-domain proteins.

In this work, we describe an automated system (MAUS: Methyl Assignments Using Satisfiability), which first formulates a set of rules, and then provides a compact description of all assignment possibilities that are consistent with these rules (Figure 3.2). Specifically, MAUS generates a structure graph, G, representing all methyl NOE connectivities present in an input PDB structure or structural model of a protein of interest, and multiple independent data graphs, H, containing all possible NOE networks, which can be derived from a list of raw 3D or 4D NOESY peaks. The NOE network is supplemented with residue type, stereospecificity, and geminal methyl connectivity constraints. Then, MAUS leverages an efficient algorithm to determine all valid ways of mapping every H into G (termed subgraph isomorphism), which respects all the experimental inputs. We test our method on a benchmark set of protein targets in the 10–45 kDa size range and show that MAUS maintains a robust performance, providing 100% accurate assignments at high levels of completeness, while offering a significant performance advantage relative to existing methods using the same inputs. Using MAUS, the methyl resonances of large, multi-domain proteins can be assigned accurately in a matter of days, completely bypassing the need for more laborious backbone-based NMR spectroscopy approaches.

MAUS models the NOE data as a sparse sample of all possible connectivities present in the input structure. MAUS uses the NOE network together with additional experimental inputs, such as peak residue type information and geminal methyl resonance connectivities, to build a system of hard constraints. The constraints outline a subgraph isomorphism problem of fitting a sparse data graph H, into the original structure graph G (Figure 3.2A).

We tested MAUS using representative data sets recorded for a benchmark set of four protein targets spanning a range of sizes, folds and domain complexities: human β2-microglobulin (Hβ2m; 12 kDa, all-β-fold, single domain), maltose-binding protein (MBP; 41 kDa, all-α-fold, two-domain), and two major histocompatibility complex class-I (MHC-I) molecules of divergent heavy-chain sequences (HLA-A01; 45.5 kDa and HLA-A02; 44.8 kDa, mixed α/β-fold, three-domain) (Figure. 3.2B). The X-ray structures and ground truth assignments for these targets were obtained from the PDB and BMRB, respectively. To obtain a consistent set of experimental data for all targets, we prepared ^{13}C/^{1}H (MA)ILV-methyl-labeled samples and acquired one 2D reference ^{1}H-^{13}C heteronuclear multiple quantum coherence (HMQC) and two 3D $^{13}C_M$-$^{13}C_M$ $^{1}H_M$ NOESY-HMQC spectra, recorded with short (50 ms) and long (300 ms) mixing times [73].

Towards reducing spectral overlap in the 2D reference spectra of larger (>20 kDa) targets, we prepared proS-labeled samples, stereospecifically defining the resonances of Leu/Val methyls [73]. Using this information as input for MAUS, we find that among all possible clustering/symmetrization alternatives resulting from peak overlap (2^{40}-2^{80}), only a tractable number (2^{10}-2^{20}) can lead to valid assignment solutions and are further explored in exhaustive subgraph isomorphism (H into G) enumerations using the SATisfiability algorithm. Given the sparsity of experimental NOE data sets and our definition of a valid solution, each subgraph isomorphism instance has up to 2^{100} valid solutions for a typical 100-methyl protein. However, the solution space is not uniformly distributed among methyl peaks; remarkably, results for all targets in our set show that the NOE network, residue type and stereospecificity constraints are sufficient to provide unambiguous assignments for a large fraction (64–89%) and low-ambiguity (two to three) options for the majority (11–30%) of remaining methyl resonances [73]. Together with benchmark targets, we also assigned resonances of IL-2, and blind targets, domains of CAS9 enzyme, HNH, REC2 and REC3, with 60-90% unique and 11-29% low-ambiguity assignments. The methodology and complete results are available at [73].

3.2 Rosetta Software Suite

Rosetta is a molecular modeling suite that provides several algorithms to model, dock, design and analyze protein structures [59]. The driving force behind Rosetta is its energy function, that is based on the Anfinsen's hypothesis that native-like protein conformations are stable and found at the lowest-energy minimum of the energy landscape [5]. Rosetta energy function is a combination of statistical potentials describing residue environments derived from PDB [11] structures and their interactions. It is worth noting that there are a lot of local minima on this energy landscape and heuristic approaches generally get trapped in these minima and our goal is to find the global minimum where the native resides and this problem is challenging.

Together with physically realistic energy function, Rosetta is equipped to utilize restraints from experimental data obtained through X-ray crystallography and/or NMR spectroscopy to guide the search along the energy landscape for the native structure during modeling [59].

3.2.1 Chemical Shift-Rosetta

Chemical Shift-Rosetta (CS-Rosetta) is an automated method that employs NMR chemical shifts to model protein structures de novo. CS-Rosetta contains two algorithms, Abrelax (ab initio relax) [106] and RASREC (resolution-adapted structural recombination) [55], that each utilize torsion angle, and short- and long-range distance restraints derived from NMR chemical shifts, residual dipolar couplings (RDCs) [93] and nuclear Overhauser enhancements (NOEs) to model protein structures up to 40 kDa [57]. Compared to Abrelax, RASREC makes use of additional optimization strategies together with the support for advanced parallelization scheme toward efficient and accurate modeling of larger protein structures. CS-Rosetta also enables rapid and automated assignment of NOESY data through the AutoNOE (Automatic NOESY assignment) module [56].

RASREC-Rosetta is a Monte Carlo-based fragment (structural fragments of amino acid lengths 3 and 9) assembly approach that utilizes NMR chemical shifts to guide the conformation search for natives [55]. Alongside NMR data, RASREC-Rosetta uses optimized algorithms across six stages of resampling to achieve high structural convergence. During the initial stages of the protocol, various β-sheet topologies are sampled. In the subsequent stages, fragments derived from (i) high-resolution X-ray structures, and (ii) preliminary low-resolution conformations (from the initial sampling stages), are applied to intensify and finalize the folds of a target protein. In the final stages of the protocol, the low-resolution models generated during the initial and intermediate stages are refined in the Rosetta force field to produce high-resolution structures.

Alongside de novo structure prediction, CS-Rosetta offers an algorithm, AutoNOE, for automated assignment of 2D, 3D, or 4D NOESY cross-peaks [56]. The NOESY cross-peaks result from coupling of two proton nuclei with different chemical shift values during magnetization transfer [49]. The process of labeling these cross-peaks in the NOESY spectra is known as NOE assignment. Whereas information about atom labels can be obtained from a known structure or an assigned resonance list, AutoNOE makes use of assigned chemical shift list to carry out the NOE assignment task. These assigned NOESY cross-peaks give rise to interproton distance restraints that are employed by structure determination methods to guide the conformational search. In CS-Rosetta, the NOE assignment (or NOESY) module is inspired largely

by alternative automated tools for NOE assignment including ARIA [60, 80, 81], AutoStruct [40], and CANDID [36]. AutoNOE takes as input the assigned NMR resonances together with unassigned NOESY cross-peak lists and generates a set of NOESY assignments from which it derives NOE distance restraints for structure calculation within RASREC.

The objective of the NOESY module is to assign proton atom labels from the input chemical shift list to the NOE cross-peaks such that the values of frequencies in both these lists match within the user specified tolerance levels. To achieve this, the method starts by mapping input chemical shift frequencies to the unassigned NOESY cross-peaks. The initial mapping results in a set of ambiguously assigned NOEs due to potential overlap in the spectra [80, 81]. These ambiguous NOEs are filtered based on a series of criteria. At first, the diagonal peaks (these are the peaks, which have same frequency for the proton and carbon dimensions) are eliminated. Second, a set of scores are computed to determine the confidence in the assignments. Third, NOE cross-peak assignments are filtered using the thresholds determined based on the confidence scores. Finally, NOE cross-peak intensities are calibrated to generate distance restraints. A complete guide to CS-Rosetta is available at [75].

3.3 Solution NMR structures solved using Rosetta

We applied Rosetta together with NMR data obtained from complementary experiments to model various structures with distinct folds and complexities (Figure 3.3). In particular, we utilized Symmetric Fold and Dock [21, 22], or CS-Rosetta's [75] RASREC [55] and AutoNOE [56] protocols to model protein structures. As pointed out in previous sections, these protocols generally employ NMR chemical shifts (probes that report on the local magnetic environment of nuclei, allowing for insights into backbone secondary structure, side chain conformations, dynamics, solvation, and hydrogen bonding) [74], NOE measurements (that reveals about pairs of residues closer in space and gives rise to distance constraints within a protein structure) and residual dipolar coupling (RDC) data (provides information about the orientation of the molecules containing multiple domains with respect to a global alignment frame).

With collaboration from several labs, we modeled structures of seven important target proteins, (i) BH_10 (10^{th} β-helix, designed synthetic protein) with 77 amino acids is a first all non-local β-strand designed protein [67], (ii) NRD-TAD (Negative Regulatory domain and Transac-

tivation domain) complex with 79 amino acids that acts as a critical activator of mitotic gene expression [66], (iii) RTT (RNA processing and transcription termination factor) with 133 amino acids that mediates termination of RNA transcription [25], (iv) ms6282 (KanY protein) with 145 amino acids [25], (v) aLP (α-lytic protease) with 197 amino acids which cleaves a tetrapeptide component in bacterial cell walls [25], (vi) nEIt (Enzyme 1 that has two domains) with 248 amino acids which regulates bacterial metabolism and is a potential drug target [25], and (vii) XAA, a 288 amino acid designed homotrimer that adopts two divergent folds [126] (Figure 3.3). The structures of RTT, ms6282, aLP and nEIT were computed as part of 4D-CHAINS/AutoNOE-Rosetta combined protocol (Figure 3.1B). The coordinates of all these models are deposited in PDB under codes 5WOX, 5WOZ, 5WOY, 5WOT, 6E5C, 6OSW and 6O0I.

3.3.1 Structure of NRAS Q61K/HLA-A*01:01 complex

[The NMR structure reported in this subsection was solved by Andrew C. McShan and David Flores-Solis. The manuscript for this work is in preparation.]

The NRAS gene is an oncogene that produces a signalling protein which is important for cell proliferation and differentiation. It has been shown that mutations in NRAS gene leads to permanent activation of the NRAS protein which causes cells to grow and divide continuously leading to variety of cancers. Moreover, the NRAS Q61K mutation was found in 7% of melanoma patients that had mutations in the NRAS gene [4]. Therefore, NRAS Q61K is used as an important biomarker in antibody-based treatments in patients. Recently, a solution NMR structure of MHC-I (or specifically HLA-A*01:01) complex with NRAS Q61K mutation leading to a peptide ILDTAGKEEY of length 10 was determined using Rosetta's FlexPepDock [61] and NOE distance restraints (Figure 3.4A; manuscript in preparation). The NOE restraints captured across the entire peptide and the MHC groove help guide FlexPepDock calculations towards the native as shown by the dotted lines in the Figure 3.4A. Using this solved structure, we can understand the mechanism of NRAS Q61K mutation peptide interaction with TCRs and design targeted therapies that can kill cells displaying the mutated peptide.

Whereas it is useful to obtain structures of high-valued pMHC targets for the design of efficient therapeutics, the process of obtaining experimental restraints limits the number of struc-

tures that can be determined. To obtain the NOE restraints for HLA-A*01:01/ILDTAGKEEY, the authors of this work expressed the protein in E. coli, refolded FYAILV labelled MHC heavy chain and unlabelled β2m domain and purified. They performed a suite of complementary NMR experiments on this protein followed by the processing and assignment of the NMR data. The timeline for the whole process is typically 1 to 1.5 months where 1 week is used to obtain refolded protein, between 2 weeks to a month to process and assign NMR data and a few days to calculate structures.

To reduce the time taken to solve pMHC structures, we decided to model them with Rosetta. Due to the availability of homologous pMHC structures displaying peptides of length 10 in the PDB, we performed homology modeling of HLA-A*01:01/ILDTAGKEEY structure (Figure 3.4B). We utilized the only available HLA-A*01:01 structure in the PDB that is bound to a peptide of length 10 (PDB ID: 6AT9, neuroblastoma epitope, AQDIYRASYY, displayed by HLA-A*01:01, solved previously in our lab) as template for homology modeling and found that the homology modeled structure when superimposed with the solution NMR structure was within 0.75 Å backbone heavy-atom RMSD with respect to the peptide (Figure 3.4B and C). In contrast to the time taken to solve solution NMR structure, our modeling method took about a minute. From this example, we established that the homology modeling method can determine sub-angstrom models for important therapeutic targets and in a high-throughput manner thereby reducing the efforts and cost needed to perform NMR experiments. In the next section, we discuss our homology modeling, RosettaMHC, workflow.

3.4 Conclusions

NMR is an indispensable tool to study proteins as demonstrated in this section. Experimental restraints obtained from NMR can guide molecular modeling suites such as Rosetta to solve structures of protein targets of clinical relevance. While Rosetta together with NOEs can help obtain structures of pMHC complexes, the process of obtaining NOEs is typically tedious, time-consuming and not amenable to all the targets. We can instead employ solved structures in PDB to guide us towards obtaining native-like pMHC molecular models as shown here for HLA-A*01:01/NRAS Q61K mutation peptide. These results suggest that modeling methods have the capacity to sample accurate peptide backbones efficiently for a given pMHC molecule.

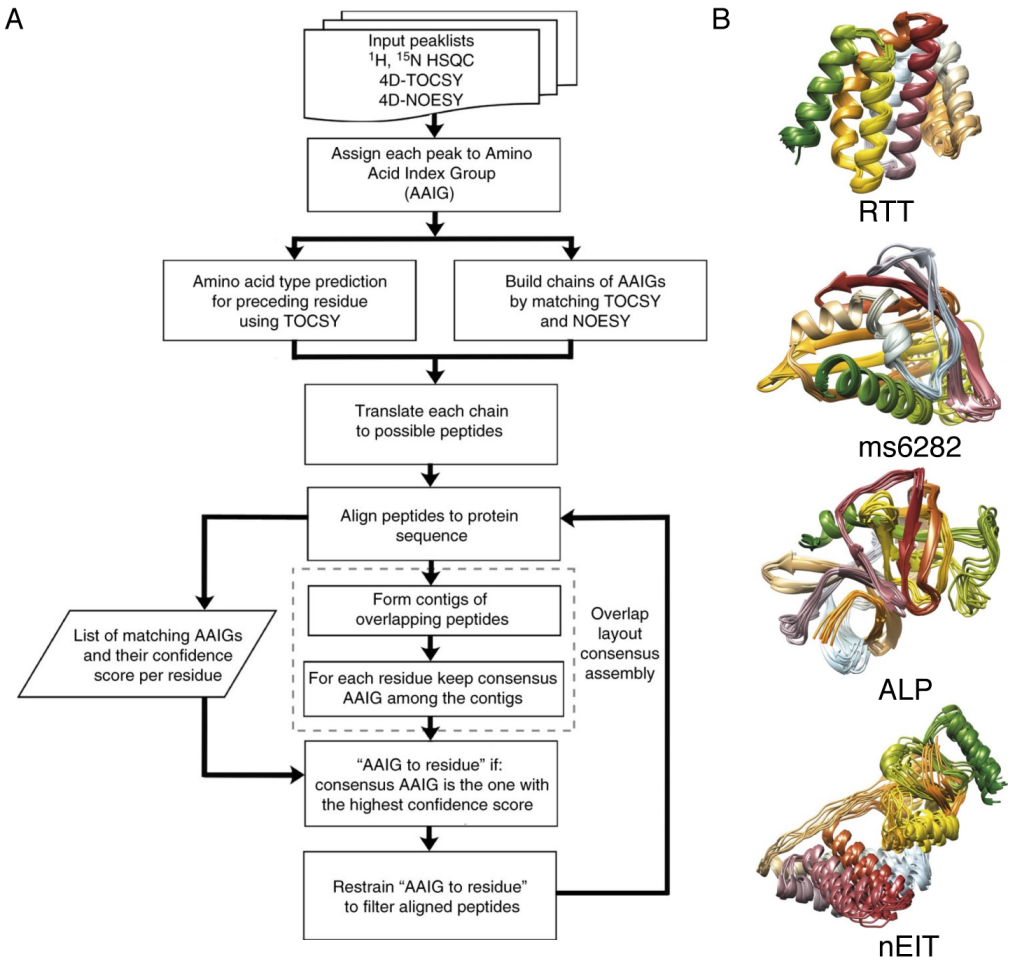


Figure 3.1: (A) Flowchart of the 4D-CHAINS algorithm for automated NMR resonance assignment from two 4D spectra (TOCSY and NOESY) (B) Structural ensemble calculated from supervised resonance assignments of Rtt103 (RTT, 133 aa), KanY (ms6282, 145 aa), α-lytic protease (aLP, 198 aa) and Enzyme I (nEIt, 248 aa) and automated NOE assignment and structure determination protocol. Panels of the figure are adopted from [25].

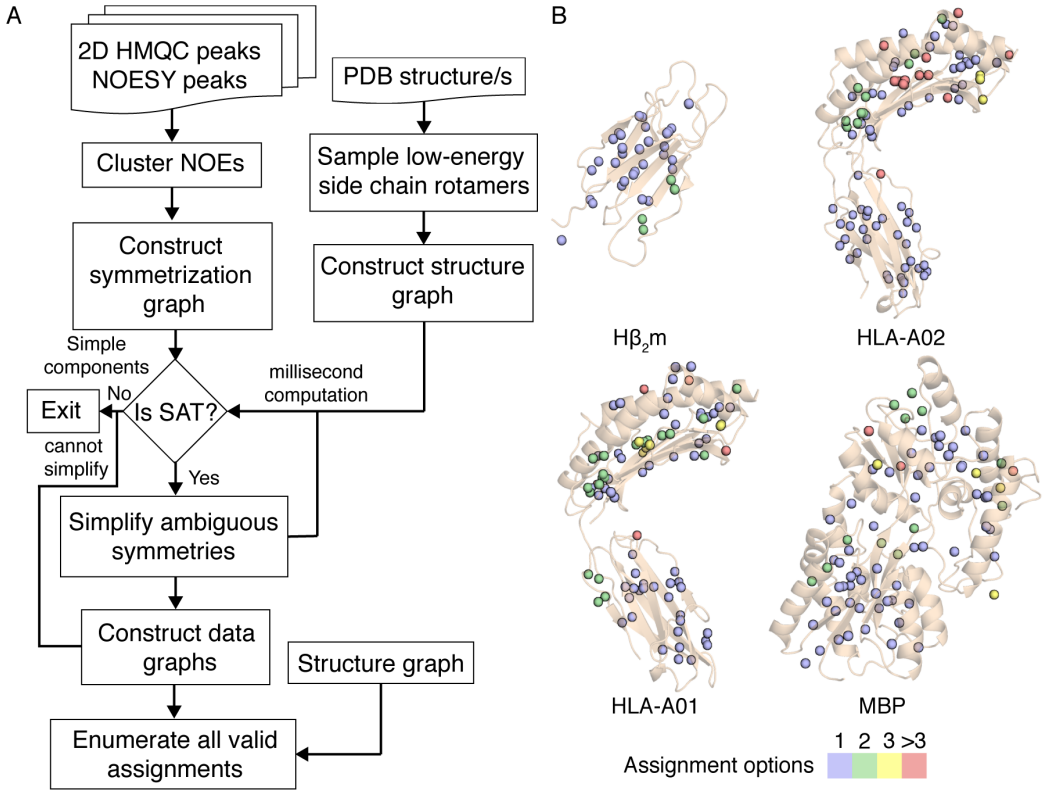


Figure 3.2: (A) Flowchart showing the working of the MAUS system. (B) Number of valid assignment options produced by MAUS for each methyl-bearing residue, are shown for protein targets of different folds (Hβ2m, HLA-A02, HLA-A01 and MBP). The colored spheres represent valid resonance assignment options, violet, green, yellow and red for 1, 2, 3 and >3 option/s, respectively. Assignment accuracy is 100% for targets. Panels of the figure are adopted from [73].

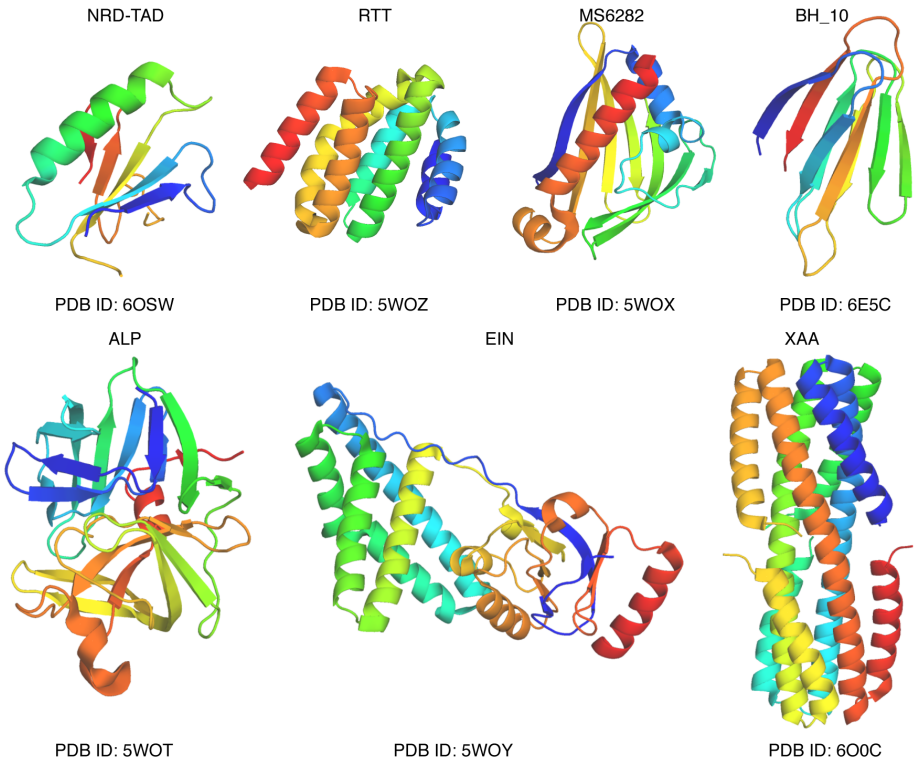


Figure 3.3: Protein structures modeled using Symmetric Fold and Dock, AutoNOE- and RASREC-Rosetta. RTT (PDB ID: 5WOZ), MS6282 (PDB ID: 5WOX), ALP (PDB ID: 5WOT) and BH_10 (PDB ID: 6E5C) were modeled using AutoNOE-Rosetta with automatically assigned NOEs from 4D NOESY data and chemical shift lists. nEIt (PDB ID: 5WOY), a dimer, was modeled using AutoNOE-Rosetta with automatically assigned NOEs from 4D NOESY data and chemical shift list together with RDC measurements to capture relative motion of the 2 domains. NRD-TAD was modeled using RASREC-Rosetta by applying NOEs derived from 3D NOESY experiments, chemical shift list and RDC data. The homotrimer (XAA) was modeled using Symmetric fold-and-dock [21] by utilizing inter- and intra-residue NOEs between domains and RDC measurements.

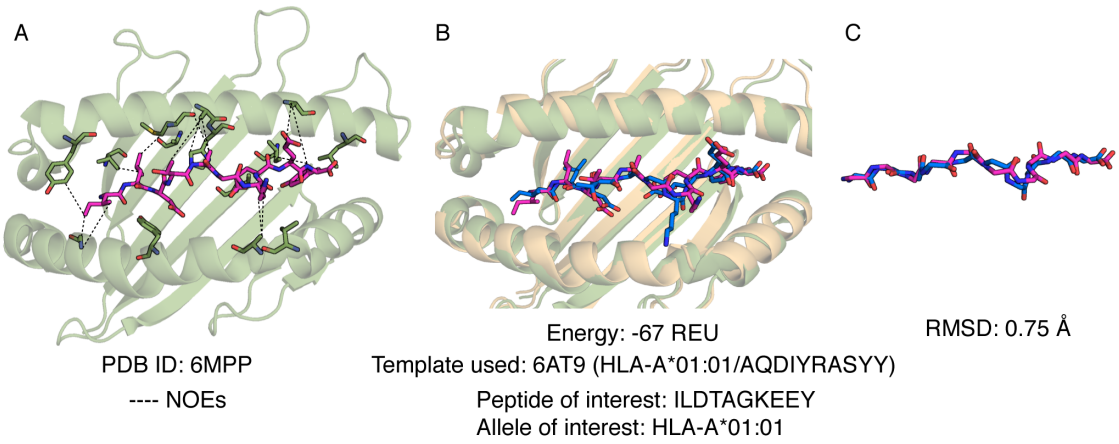


Figure 3.4: (A) Solution NMR structure (PDB ID: 6MPP) of HLA-A*01:01/NRAS Q61K (ILDTAGKEEY) modeled using Rosetta's FlexPepDock protocol and NOE restraints. The MHC heavy chain is shown in olive green, peptide is shown as stick representation in magenta and the dotted black lines showcase the NOE restraints used in modeling. (B) Superimposed structural models obtained using solution NMR (olive green and magenta, same as panel A) and our homology modeling method, RosettaMHC (wheat and blue). The structural template used for homology modeling is HLA-A*01:01 allele displaying ALK neoepitope, AQDIYRASYY (PDB ID: 6AT9, indicated below). The binding energy of the model is -67 Rosetta Energy Units or REU. (C) The superimposed peptide backbones of the solution NMR structure and the RosettaMHC model. The RMSD between the two peptide backbone conformations is 0.75 Å.

Chapter 4

RosettaMHC

[Some of the text and figures in this chapter have been published with the following citations:

- Jugmohit S. Toor, Arjun A. Rao, Andrew C. McShan, Mark Yarmarkovich, **Santrupti Nerli**, Karissa Yamaguchi, Ada A. Madejska, Son Nguyen, Sarvind Tripathi, John M. Maris, Sofie R. Salama, David Haussler, and Nikolaos G. Sgourakis. A Recurrent Mutationin Anaplastic Lymphoma Kinase with Distinct Neoepitope Conformations. *Frontiers in Immunology*, 9:99, 2018.
- **Santrupti Nerli** and Nikolaos G. Sgourakis. Structure-Based Modeling of SARS-CoV-2 Peptide/HLA-A02 Antigens. *Frontiers in Medical Technology*, 2020.
- Jason D. Fernandes, Angie S. Hinrichs, Hiram Clawson, Jairo Navarro Gonzalez, Brian T.Lee, Luis R. Nassar, Brian J. Raney, Kate R. Rosenbloom, **Santrupti Nerli**, Arjun A. Rao,Daniel Schmelter, Alastair Fyfe, Nathan Maulding, Ann S. Zweig, Todd M. Lowe, ManuelAres, Russ Corbet-Detig, W. James Kent, David Haussler, and Maximilian Haeussler. The UCSC SARS-CoV-2 Genome Browser. *Nature Genetics*, pages 1–8, September2020.
- **Santrupti Nerli**, Sagar Gupta, Andrew McShan, and Nikolaos G. Sgourakis. RosettaMHC: Consistent modeling of peptide/MHC structures for common HLA allotypes. *Manuscript in preparation*.

]

RosettaMHC is a method for comparative (or homology)-based modeling of class I MHC complexes. The user may input the names of the HLA alleles together with target peptide sequences and template PDBs to model the target peptide/MHC complex on to the template structure. As a first step, a template structure is optimized in the Rosetta force field followed by the alignment of the target and template sequences. The alignment together with refined template are used to thread the target sequence. For the purpose of throughput, only the peptide sequence and the residues in the MHC groove that are within 3.5 Å from the peptide in the threaded structure are refined using Rosetta's optimization protocol, Relax [121]. The binding energy is extracted from the refined threaded model. It is worth noting that we only homology model the α1 and α2 domains (typically 1-180 residues) of the MHC molecule since this region gives a snapshot of the pMHC surface interacting with TCRs (Figure 4.1).

RosettaMHC makes use of the protocols available in Rosetta [58] via the wrappers written in PyRosetta [16]. Specifically, it utilizes, Idealize, Relax [121], Partial_thread and InterfaceAnalyzer protocols in its pipeline.

- Idealize: This protocol is used to idealize the bond lengths and angles in the input template structure.
- Relax: We used FastRelax (henceforth referred to as Relax) protocol with default parameters to carry out all-atom refinement of the peptide and the MHC binding groove (residues in the MHC molecule that are within a distance of 3.5 Å from the peptide). Relax combines side chain packing and gradient based minimization of the backbone degrees of freedom. The default parameters of Relax allows it to perform four rounds of side chain repacking followed by minimization of the backbone torsional angles. The repulsive energy term is tuned up to 2%, 25%, 55% and 100% in each of the four rounds.
- Partial_thread: To thread a target sequence on to the template structure, we use partial_thread protocol.
- InterfaceAnalyzer: The threaded structures are relaxed and the binding energies between peptide and the MHC molecules are extracted using the InterfaceAnalyzer protocol. Here, the dG_separated score term is reported as the binding energy. This term is computed by

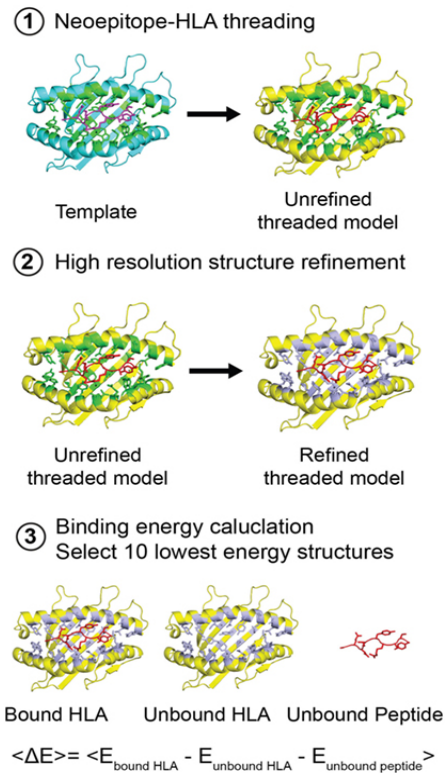


Figure 4.1: Structure-based modeling of pMHC complexes. Step 1: A template (blue) peptide/MHC complex (X-ray structure) is provided to generate a threaded model with the peptide and HLA allele of interest (yellow). MHC residues in the groove within 3.5 Å of the peptide are colored green. Step 2: Models are refined by energy minimization and side chain repacking of groove and peptide residues (gray). Step 3: The average peptide-binding energy is determined by subtracting the energy of the unbound MHC and unbound peptide from the energy of the peptide bound MHC. $\langle E \rangle$ represents the average binding energy. Figure adopted from [115].

measuring the change in the energy when the peptide and the MHC chains are separated compared to their corresponding bound state structure.

RosettaMHC (version 1.0 described here) software is available for download at github. This tool requires Python3, PyRosetta4, Biopython [18], and Clustal Omega [108]. The links to each of these tools is available on the README page of github.

We want to emphasize that alternative software suites that implement homology modeling methods, Prime (Schrödinger, LLC), SWISS-MODEL [125], MODELLER [24], Rosetta [58], and I-TASSER [100], show comparable performance in terms of modeling accuracy when the sequence identity of the target with template is greater than 30% [23]. However, Rosetta's optimization algorithm (Relax) is efficient which explains our choice of Rosetta to carry out homology

modeling [86].

To demonstrate the accuracy of RosettaMHC, we performed benchmark calculations using a non-redundant set of 90 9-mer peptide/HLA-A*02 complex structures from PDB. Each epitope was modeled from the closest template with identical anchor residues (but at least two or more mutations away in the non-anchor residue positions) present in the benchmark set, and homologous peptide sequences were excluded from template selection. From these results, we find that (i) the energies of RosettaMHC models fall within the distribution of the PDB template energies (Figure 4.2A), and (ii) the models generated for 75% and 98% of peptides fall within 1.5 Å and 2 Å backbone heavy-atom RMSD from their native X-ray structures, respectively (Figure 4.2B). These results suggest that RosettaMHC can provide accurate models of peptide/HLA-A*02 complexes, for a range of peptide sequences.

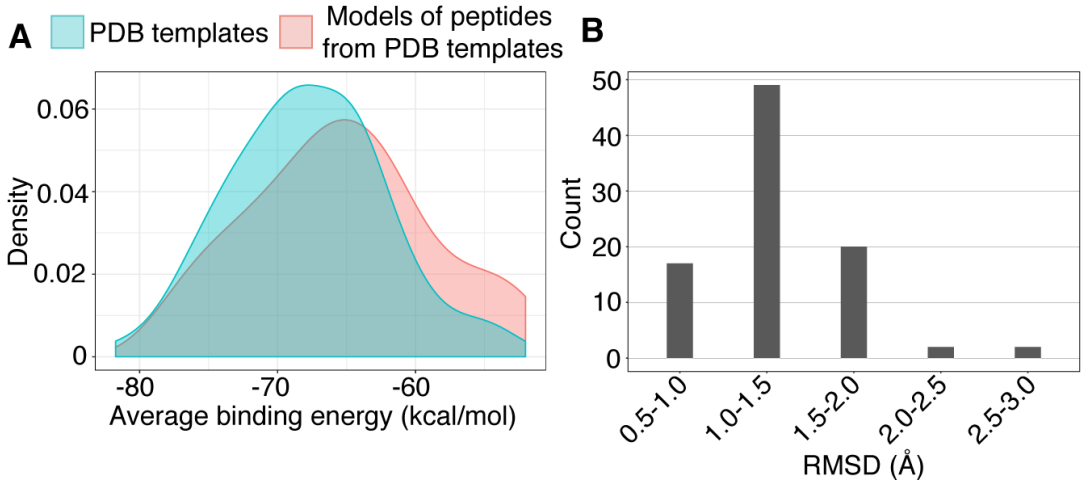


Figure 4.2: Performance of RosettaMHC. (A) Benchmark calculations of 90 9-mer peptide sequences derived from HLA-A*02 structures in PDB. The density plots showing the distribution of average binding energies of PDB templates (cyan) and the models of peptides derived from PDB templates (crimson). (B) Backbone heavy atom (N, C, CA, and O) RMSD distributions of modeled peptides relative to their native X-ray structures in PDB. This figure is adopted from [76].

4.1 Evaluation of HLA-binding repertoire of two neoepitopes derived from Anaplastic Lymphoma Kinase

The identification of recurrent HLA neoepitopes driving T cell responses against tumors poses a significant bottleneck in the development of approaches for precision cancer therapeutics. In this work, we employ a bioinformatics method, Prediction of T Cell Epitopes for Cancer Therapy (ProTECT) [95], to analyze sequencing data from neuroblastoma patients and identify a recurrent anaplastic lymphoma kinase (ALK) mutation (ALK R1275Q) that leads to two high affinity neoepitopes when expressed in complex with common HLA alleles [115]. Analysis of the X-ray structures of the two peptides bound to HLA-B*15:01 reveals drastically different conformations with measurable changes in the stability of the protein complexes, while the self-epitope is excluded from binding due to steric hindrance in the MHC groove. To evaluate the range of HLA alleles that could display the ALK neoepitopes, we modeled ALK epitopes/HLA structures using RosettaMHC. We found that RosettaMHC accurately predicts several additional high affinity interactions by comparing our results with commonly used prediction tools. Subsequent determination of the X-ray structure of HLA-A*01:01 bound neoepitope validates atomic features seen in our RosettaMHC models with respect to key residues relevant for MHC stability and T cell receptor recognition [115].

4.1.1 Identification and characterization of HLA-alleles that can putatively bind ALK neoepitopes

A patient's HLA haplotype plays a major role in determining the outcome of targeted cancer immunotherapies. Therefore, toward expanding the range of individuals that could mount a T cell response to ALK R1275Q neoepitopes, we evaluated the potential of other HLAs to display the two peptides (AQDIYRASY and AQDIYRASYY) identified by ProTECT in silico. First, we selected a non-redundant set of 2,904 HLA alleles (885 HLA-A, 1,405 HLA-B, and 614 HLA-C unique sequences) from the EMBL-EBI database [62]. We then carried out RosettaMHC calculations for each allele, using the experimentally determined HLA-B*15:01 structures for the nonamer and decamer ALK peptides as templates (Figure 4.1). In contrast to previous structure-based peptide/MHC modeling methods which use a flexible peptide docking approach [6], we

used RosettaMHC, a fixed-peptide backbone threading approach followed by energy minimization of the interacting peptide and HLA residues to drastically confine the docking degrees of freedom. Using this strategy, we extracted highly reproducible binding energies for both the nonamer and decamer peptides, which are maintained in extended and 3_{10} helical conformations, respectively, in the resulting models (Figure 4.3). As expected, the HLA-B*15 alleles rank systematically among the top binders, indicating a high degree of groove complementarity to both peptides (Figures 4.4 and 4.5, purple). Among those, the HLA-B*15:84 allele shows the lowest binding energy for the decamer (Figure 4.5, black circle), whereas the HLA-B*15:107 allele shows the lowest binding energy for the nonamer (Figure 4.4, black circle). A total of 116 HLA alleles from all A, B, and C types exhibit lower binding energies for both the nonamer and decamer peptides than our initial HLA-B*15:01 structural templates (Figures 4.4 and 4.5, red square), suggesting the potential for a broader HLA display repertoire.

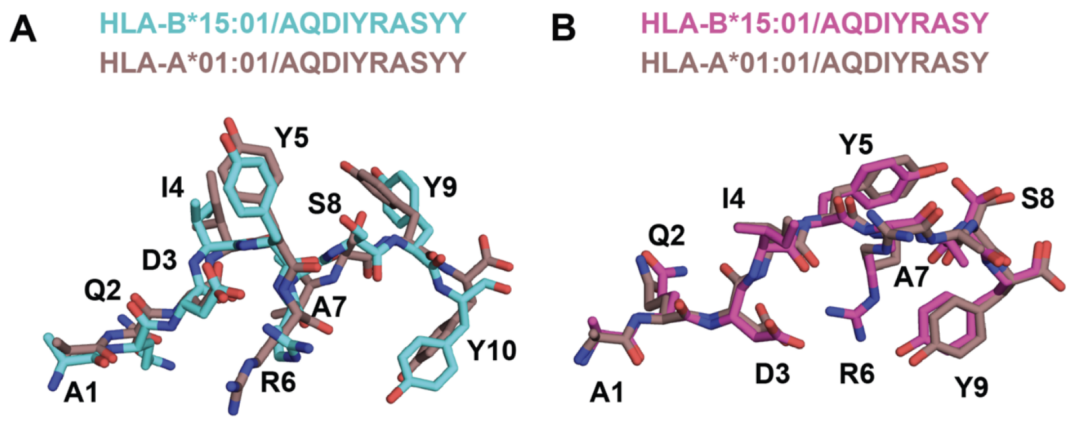


Figure 4.3: Conformation of decamer and nonamer peptides displayed by HLA-A*01:01 from structural modeling. (A) 3_{10} helical conformation of the decamer peptide observed when bound to HLA-B*15:01 (cyan) in our X-ray structure or to HLA-A*01:01 (brown) in homology modeling simulations. (B) Extended conformation of the nonamer peptide observed when bound to HLA-B*15:01 (pink) in our X-ray structure or to HLA-A*01:01 (brown) in homology modeling simulations. Figure adopted from [115].

To elucidate a sequence bias for specific residues in the HLA-binding groove that consistently yield more favorable interactions with the two peptides, we analyzed the average binding energy as a function of sequence identity score [35], calculated relative to the best binding allele for each peptide (Figure 4.1). As a negative control, we computed the binding energy for a mock HLA allele in which all residues in the MHC-binding groove are mutated to alanines

(polyAla). As expected, the mock polyAla HLA exhibits a low binding affinity (i.e., high-binding energy) to the peptide and is distant from the best binding allele (Figures 4.4 and 4.5, green triangle, top left). We observe an evident correlation between the computed binding energies and sequence similarity to the top binder. Our approach additionally allows us to decompose residue specific contributions to overall binding energy for each peptide/HLA combination. We find a clear trend for both the nonamer and decamer peptides with a set of HLA alleles where a bulk of the binding energy is provided by the ”anchor” positions (Figure 4.6). By contrast, the mock polyAla HLA exhibits considerably higher binding energy across the entire peptide length (Figure 4.6). To elucidate key sequence features that allow the peptides to be accommodated in the MHC groove, we derived a sequence profile among good binders for the two neoepitopes. Such features are highlighted in the Kullback–Leibler sequence logo, which reveals preferred residues in the HLA peptide-binding groove (Figures 4.4B and 4.5B). According to this metric, highly invariant residues in the MHC-binding groove should play an essential role in mediating pMHC interactions, as they are consistently observed in HLA alleles that exhibit high affinity binding. A close inspection of our structural models for the nonamer and decamer bound to a common allele in our data set, HLA-A*01:01, reveals similar polar contacts, primarily in the A-, B-, and F-pockets, that correlate well with the positions of invariant MHC residues (Figures 4.4C, D and 4.5C, D). Specifically, both the nonamer and decamer C-terminal anchors employ a similar interaction pattern in the F-pocket with conserved Thr, Lys, Trp, and Tyr residues of the MHC (Figures 4.4B, D and 4.5B, D). High-ranking HLA alleles according to Rosetta’s binding energy consistently demonstrate a low percentile rank using the epitope prediction method recommended by IEDB [123], which further suggests a high probability of forming a tight complex with the neoepitopes (Figure 4.7).

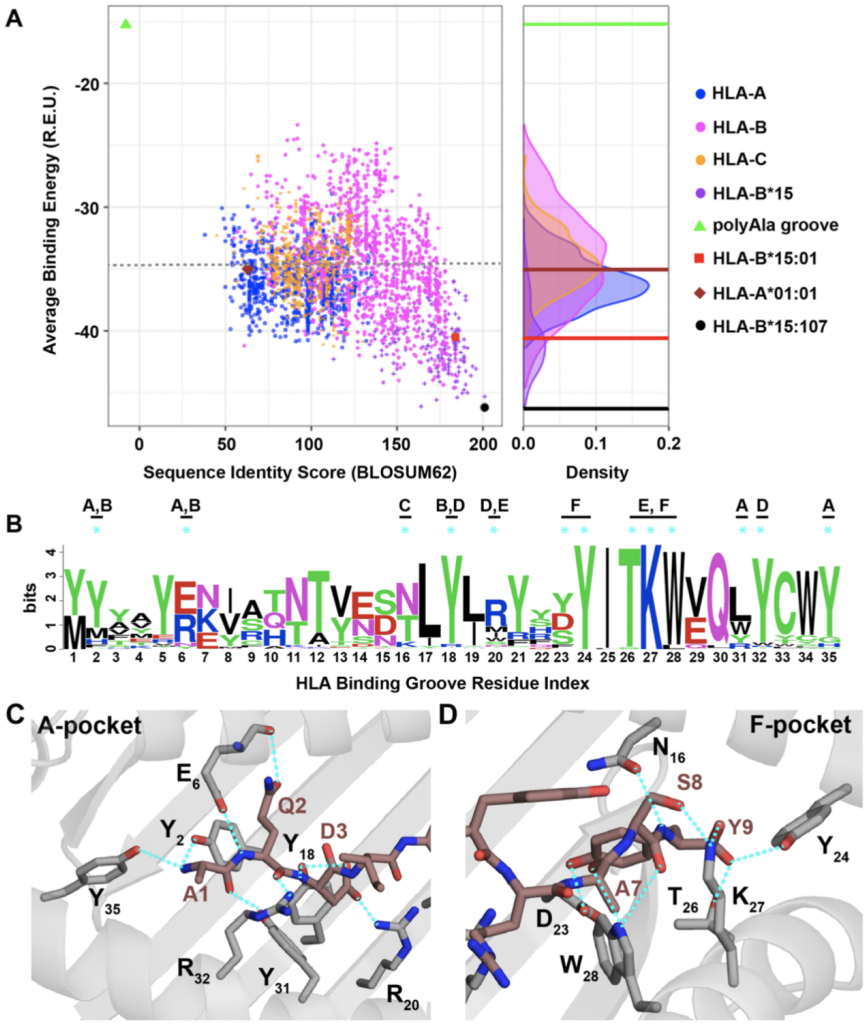

A
Average Binding Energy (R.E.U.)
-20
-30
-40
0
50
100
150
200
Sequence Identity Score (BLOSUM62)
0.0
0.1
0.2
Density
HLA-A
HLA-B
HLA-C
HLA-B*15
polyAla groove
HLA-B*15:01
HLA-A*01:01
HLA-B*15:107
B
A,B
A,B
C
B,D
D,E
F
E, F
A
D
A
bits
HLA Binding Groove Residue Index
C
A-pocket
E6
Q2
Y2
Y18
D3
A1
Y35
R32
Y31
R20
D
F-pocket
N16
S8
Y9
Y24
A7
T26
K27
D23
W28

Figure 4.4

Figure 4.4 *(previous page)*: Evaluating the HLA binding repertoire of ALK neoepitope nonamer AQDIYRASY using RosettaMHC. (A) Rosetta binding energies calculated from structure modeling of 2,904 unique HLA alleles from the IPD- IMGT/HLA [97] database, for the ALK neoepitope nonamer (AQDIYRASY) plotted as a function of sequence similarity to the top binding allele, HLA-B*15:107 (black circle). The binding energy of nonamer in our HLA-B*15:01 X-ray structure is shown as a reference (red square). A negative control was performed with a mock HLA allele where all residues in the binding groove were replaced with Ala (polyAla groove, green triangle), which shows high binding energy. The corresponding distribution of the HLA alleles on the binding energy landscape is captured in the density plot shown on the right. Sequence identity scores were calculated using the BLOSUM62 [35] matrix. REU: Rosetta Energy Units (B) Kullback-Leibler sequence logo [111] derived from multiple sequence alignment using Clustal Omega [108] of peptide binding groove residues from all the HLA alleles that exhibit better binding energies than HLA-A*01:01 (brown diamond), indicated with a gray dotted line in (A). MHC residues with polar contacts to the peptide are denoted with a cyan asterisk with corresponding MHC pocket noted. (C) and (D) Threaded structural model of HLA-A*01:01 displaying nonamer peptide. Polar contacts between the MHC groove (gray sticks) and peptide (brown sticks) are shown with cyan dotted lines in the HLA- A*01:01 A-, B-, and D-pockets (C) or C-, E-, and F-pockets (D). The residue index for each interacting MHC residue is denoted with the corresponding number from (B) using subscripts. Peptide residues (non-indexed) are labeled without subscripts. Figure adopted from [115].

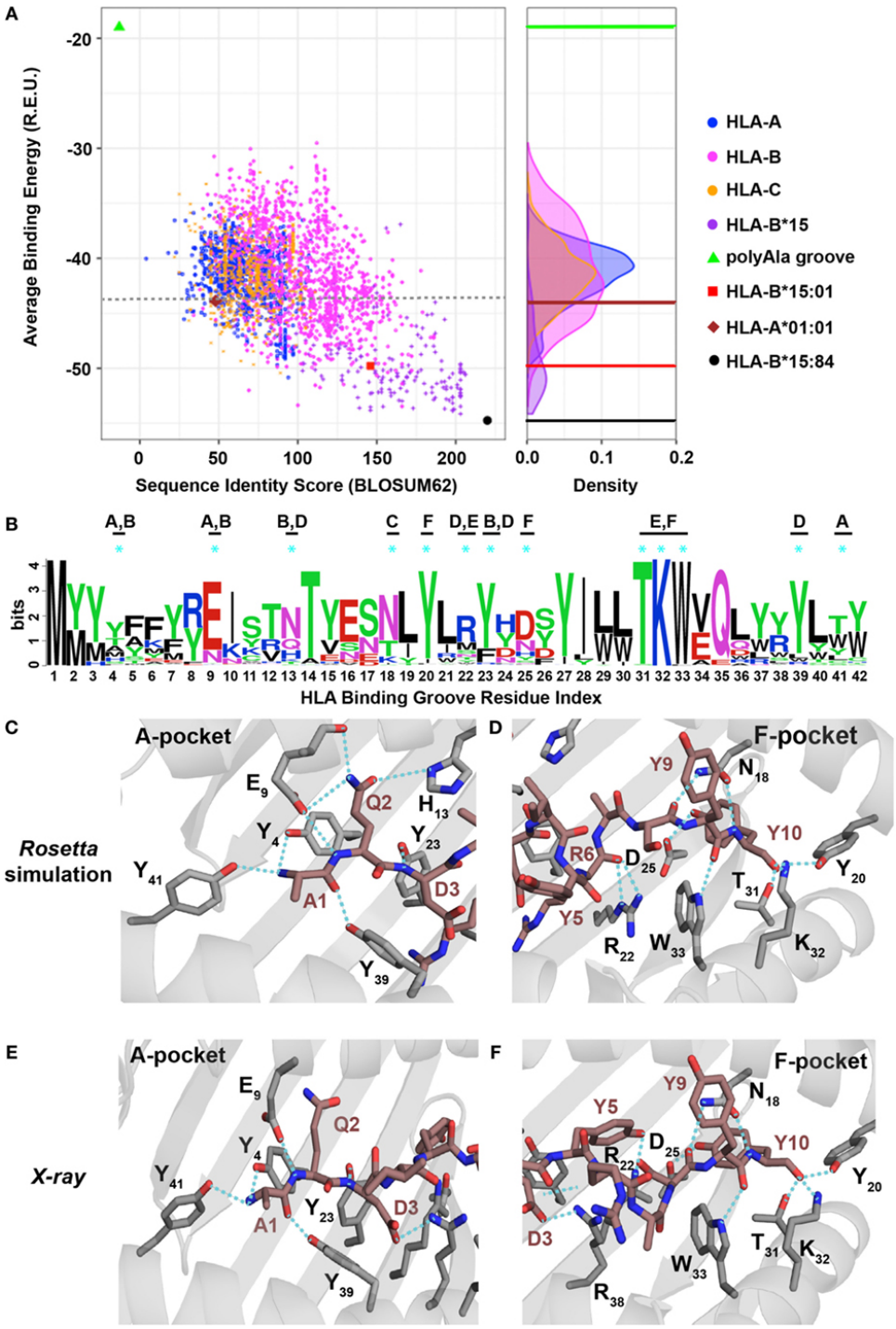

A
Average Binding Energy (R.E.U.)
-20
-30
-40
-50
0
50
100
150
200
Sequence Identity Score (BLOSUM62)
0.0
0.1
0.2
Density
HLA-A
HLA-B
HLA-C
HLA-B*15
polyAla groove
HLA-B*15:01
HLA-A*01:01
HLA-B*15:84
B
A,B
A,B
B,D
C
F
D,E
B,D
F
E,F
D
A
bits
HLA Binding Groove Residue Index
C
Rosetta simulation
A-pocket
E9
Q2
H13
Y4
Y23
Y41
A1
D3
Y39
D
F-pocket
Y9
N18
R6
D25
Y10
Y20
T31
Y5
R22
W33
K32
E
X-ray
A-pocket
E9
Q2
Y4
Y41
A1
Y23
D3
Y39
F
F-pocket
Y9
N18
Y5
D25
R22
Y10
Y20
D3
T31
K32
W33
R38

Figure 4.5 *(previous page)*: Evaluating the human leukocyte antigen (HLA)-binding repertoire of ALK decamer AQDIYRASYY using RosettaMHC. (A) Rosetta-binding energies calculated from modeling of 2,904 unique HLA alleles from the IPD-IMGT/HLA [97] database, for the ALK neoepitope decamer (AQDIYRASYY) plotted as a function of sequence similarity to the top binding allele, HLA-B*15:84 (black circle). The binding energy of decamer in our HLA-B*15:01 X-ray structure is shown as a reference (red square). A negative control was performed with a mock HLA allele where all residues in the binding groove were replaced with Ala (polyAla groove, green triangle), which shows high-binding energy. The corresponding distribution of the HLA alleles on the binding energy landscape is captured in the density plot shown on the right. Sequence identity scores were calculated using the BLOSUM62 [35] matrix. Abbreviation: REU, Rosetta energy units. (B) Kullback–Leibler sequence logo derived from multiple sequence alignment using Clustal Omega of peptide-binding groove residues from all the HLA alleles that exhibit better binding energies than HLA-A*01:01 (brown diamond), indicated with a gray dotted line in panel (A). MHC residues with polar contacts to the peptide are denoted with a cyan asterisk with corresponding MHC pocket noted. (C,D) Threaded structural model of HLA-A*01:01 displaying decamer peptide. Polar contacts between the MHC groove (gray sticks) and peptide (brown sticks) are shown with cyan dotted lines in the A-, B-, and D-pockets (C) or C-, E-, and F-pockets (D). The residue index for each interacting MHC residue is denoted with the corresponding number from panel (B) using subscripts. Peptide residues (non-indexed) are labeled without subscripts. Panels (E,F) show polar contacts observed in the A-pocket (E) and F-pocket in the X-ray structure of HLA-A*01:01/AQDIYRASYY (PDB ID 6AT9) between the peptide (brown sticks) and residues in the MHC groove (gray sticks). The residue index for each interacting MHC residue is denoted with the corresponding number from panel (B) using subscripts. Peptide residues (non-indexed) are labeled without subscripts. Figure adopted from [115].

4.1.2 High-resolution features in X-ray structure are recapitulated by the Rosetta model

To test the validity of our structure-based simulations, we performed in vitro refolding of the ALK-derived nonamer and decamer peptides with HLA-A*01:01. This allele was chosen because it is a high-frequency allele in multiple populations worldwide and has been previously shown to form stable recombinant pMHC complexes for structural characterization [63]. Finally, to conclusively test the atomic features predicted by our simulations, we determined the X-ray structure of decamer complex HLA-A*01:01/β2m/AQDIYRASYY (PDB ID 6AT9). The peptide conformation in the X-ray structure shows excellent agreement with our Rosetta model (1.1 Å backbone heavy atom RMSD), with several high-resolution features predicted by the model are confirmed by the X-ray, including polar contacts within both the A- and F-pockets of the MHC groove (Figures 4.5C–F and 4.8). Specifically, the side-chain hydroxyl group of the pep-

tide Tyr10 is in contact with the same Tyr, Lys, and Trp side-chain atoms from the F-pocket (Figures 4.5D,F). Finally, in comparison with the X-ray structure of the same peptide bound to HLA-B*15:01, the side chain of Arg6 is flipped outwards from the groove when bound to HLA-A*01:01 altering the peptide surface displayed to TCRs. Thus, our independent X-ray structure corroborates the trend observed in our structure-based binding energy simulations and further supports the potential for other HLA molecules to display the recurrent ALK neoepitopes with unique TCR interaction properties. The detailed description of the workflow together with additional experimental data are presented in the article [115].

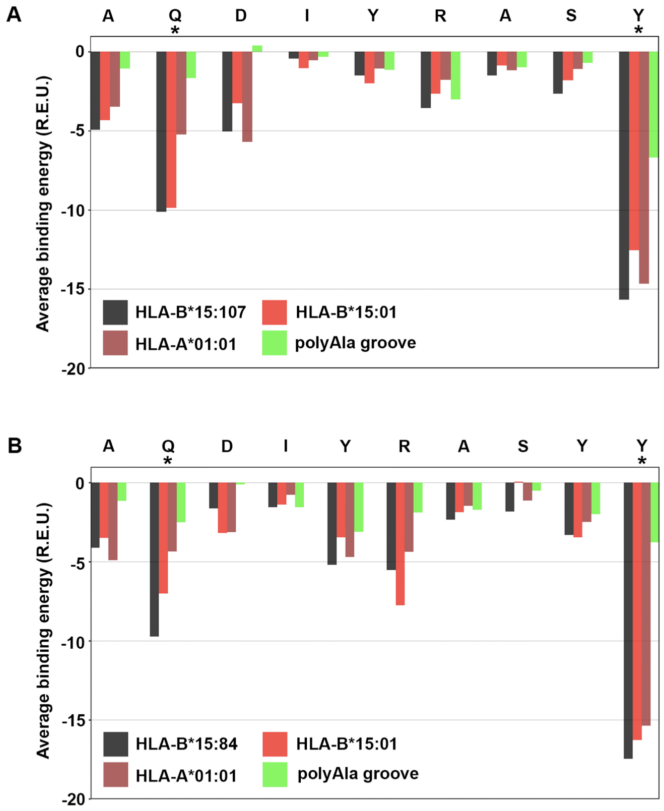


Figure 4.6: Residue specific binding energy contributions for ALK nonamer AQDIYRASY and decamer AQDIYRASY peptides against different HLA grooves. Rosetta calculated average binding energies contributed by ALK (A) nonamer and (B) decamer residues to (i) top binder HLA allele, HLA-B*15:107 or HLA-B*15:84 (black), (ii) HLA-B*15:01 (red), (iii) HLA-A*01:01 (brown), and (iv) control HLA allele (polyAla groove, green) where all the residues in the binding groove were replaced with an alanine. Y-axis shows REU: Rosetta Energy Units. X-axis represents each peptide residue. Anchor residues are indicated by an asterisk. Figure adopted from [115].

4.2 SARS-CoV-2 peptide/HLA-A*02 antigen structural models

In this subsection, we reproduce our recently published work that showcases another application of RosettaMHC [76]. Here, we utilized RosettaMHC to model all HLA-A*02:01 epitopes predicted directly from the 30 kbp SARS-CoV-2 genome. The SARS-CoV-2 protein sequences (NC_045512.2) were obtained from NCBI and used to generate all possible peptides of lengths 8, 9 and 10 (9631 8mer, 9,621 9mer and 9,611 10mer peptides). We used NetMHCpan-4.0 [45] to derive binding scores to HLA-A*02:01, and retained only peptides classified as strong or weak binders (selected using the default percentile (%) rank cut-off values). The binding classification was performed using eluted ligand likelihood predictions. Our full workflow for structure modeling is outlined in Figure 4.9A, with a flowchart shown in Figure 4.9B.

4.2.1 Selection of PDB templates

To model SARS-CoV-2/HLA-A*02:01 antigens, we identified 3D structures from the PDB that can be used as templates for homology modeling. First, we selected all HLA-A*02 X-ray structures that are below 3.5 Å resolution and retained only those that have 100% identity to the HLA-A*02:01 heavy chain sequence (residues 1-180). We found 236 redundant template structures bound to epitopes of lengths from 8 to 15 residues (of which 1 is an 8mer, 165 are 9mers and 61 are 10mers). For each SARS-CoV-2 target peptide of (i) length 8, we selected a set of candidate templates of lengths 8-9 by matching the target peptide anchor positions (P1 and P8 in the 8mer, P2 and P9 in the 9mer templates), and (ii) lengths 9 and 10, we selected candidate templates of the same peptide length, by matching the target peptide anchor positions (P2 and P9/P10) to each peptide in the template structures. Then, we used the BLOSUM62 substitution matrix to score all remaining positions in the pairwise alignment of the target/template sequences, and the peptide sequence with the top score was selected as a template for modeling. For target peptides where we found no templates which matched both peptide anchors, we scored all positions in the pairwise alignment and selected the top scoring template for modeling.

4.2.2 Structural modeling results and discussion

To predict all possible peptides expressed by SARS-CoV-2 that can bind to HLA-A*02:01, we used a recently annotated version of all open reading frames (ORFs) in the viral genome from NCBI [127], made available through the UCSC genome browser [50]. We used 8-, 9- and 10-residue sliding windows to scan all protein sequences, since these are the optimum peptide lengths for binding to the HLA-A*02:01 groove [117]. The limited availability of templates for peptides of lengths greater than 10 (9 total in the PDB) suggests that such peptides are likely to represent a small fraction of the displayed peptide repertoire, and were not considered here. While spliced peptide epitopes [71] were not considered in the current study, this set can be added to our results at a later stage. NetMHCpan-4.0 [45] predicted 54 8mer, 439 9mer and 256 10mer epitopes that can bind to HLA-A*02:01(classified as both weak and strong binders), with a majority of them originating from nsp3 protein encoded by gene orf1ab (NCBI Reference YP_009724389.1) (Figure 4.10).

To validate the SARS-CoV-2 peptide set predicted by NetMHCpan-4.0 and to derive plausible 3D models of the peptide/HLA-A*02:01 complexes, we used RosettaMHC, which leverages a database of 236 HLA-A*02:01 X-ray structures, to find the closest match to each target epitope predicted from the SARS-CoV-2 proteome (Figure 4.11A). We consider the range of available structures in the PDB as a natural sampling of different possible backbone conformations within the highly restrictive environment of the peptide-binding groove, as shown in a structural alignment of all 9mer templates (Figure 4.11B). To identify the best template for structure modeling of each target peptide, we use sequence matching criteria which first consider the peptide anchors (positions P1/P2 and P8/P9/P10 for 8mer/9mer/10mer epitopes), followed by a sequence similarity metric calculated from the full alignment (or residues from P2 through P9 for 9mer template and 8mer target peptide) between the template and target peptide sequences.

The template assignment statistics for the six different classes of SARS-CoV-2 epitopes in our set are shown in Figure 4.11C. We find that we can cover the entire set of 749 predicted 8-10 residue binders using a subset of 123 HLA-A*02:01 templates in our annotated database of PDB-derived structures (Figure 4.11D). Each target peptide sequence is then threaded onto the backbone of its best identified template, followed by all-atom refinement of the side chain and backbone degrees of freedom using Rosetta's Ref2015 energy function [5], and binding

energy calculation.

4.2.3 RosettaMHC models recapitulate features of high-resolution X-ray structures

The sequence logos derived from 9mer and 10mer peptides with good structural complementarity to the HLA-A*02:01 groove according to Rosetta's binding energy adhere to the canonical motif, with a preference for hydrophobic, methyl-bearing side chains at the peptide anchor residues P2 and P9/P10 (Figures 4.12A and 4.13A). In addition, the sequences of high-affinity binders, show preferences for specific amino acids at positions P1, P3, P6/P7, P7/P8 for 9mers and 10mer peptides, respectively (Figures 4.12A and 4.13A). These preferences are recapitulated in representative 9mer and 10mer models of the two top binders in our set as ranked by Rosetta's energy (Figures 4.12B and 4.13B), corresponding to epitopes TMADLVYAL and FLFVAAIFYL derived from the RNA polymerase and nsp4 proteins, respectively, which are both encoded by orf1ab. In accordance with features seen in high-resolution structures of HLA-A*02:01-restricted epitopes, the peptides adopt an extended, bulged backbone conformation. The free N-terminus of both peptides is stabilized by a network of polar contacts with Tyr 7, Tyr 159, Tyr 171 and Glu 63 in the A- and B- pockets of the HLA-A*02:01 groove. The Met (9mer) or Leu (10mer) side chain of P2 is buried in a B-pocket hydrophobic cleft formed by Met 45 and Val 67. Equivalently, the C-terminus is coordinated through polar contacts with Asp 77 and Lys 145 from opposite sides of the groove, with the Leu P9/P10 anchor nestled in the F-pocket defined by the side chains of Leu 81, Tyr 116, Tyr 123 and Trp 147. Residues P3-P8 form a series of backbone and side chain contacts with pockets C, D and E, while most backbone amide and carbonyl groups form hydrogen bonds with the side chains of residues lining the MHC-I groove. These high-resolution structural features are consistent across low-energy models of unrelated target peptides in our input set, suggesting that, when provided with a large set of input templates, a combined threading and side chain optimization protocol can derive accurate models (within 2 Å RMSD), as suggested by our benchmark calculations (Figure 4.2).

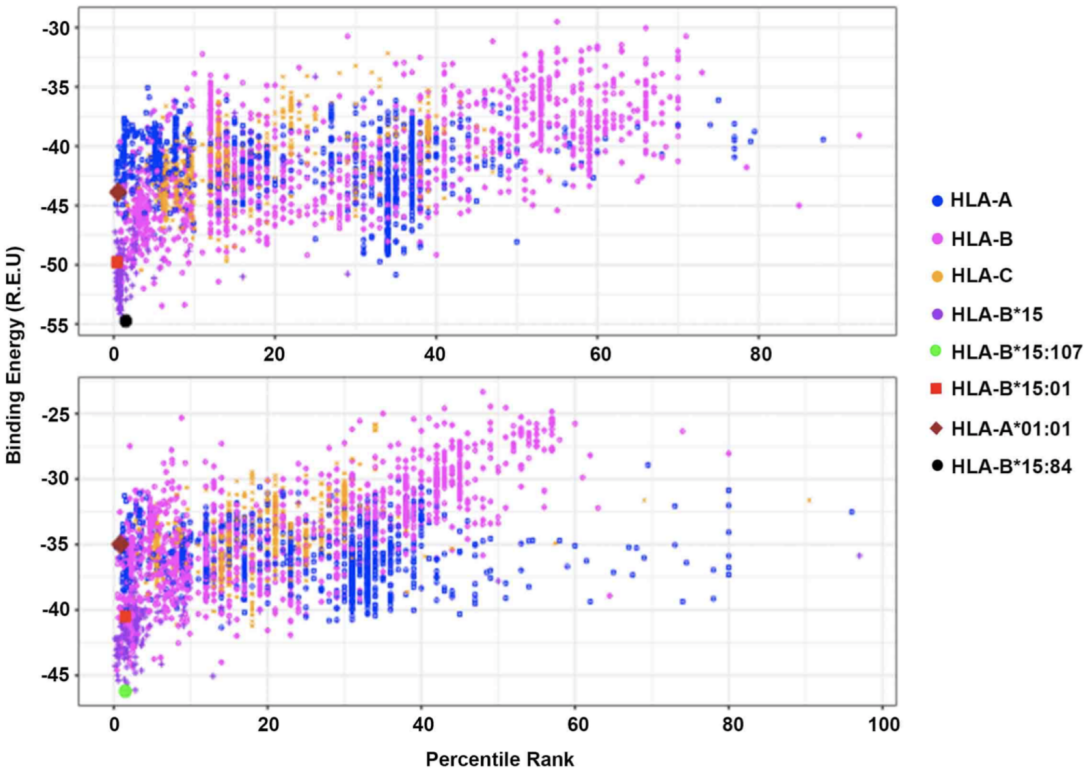


Figure 4.7: Structure-based binding energies versus IEDB epitope predictions. Paired Rosetta binding energies and IEDB [123] percentile ranks calculated for 2,904 unique HLA alleles from the IPD-IMGT/HLA Database [97]. Results are shown for the ALK decamer (AQDIYRASYY) (top panel) and nonamer (AQDIYRASY) (bottom panel) neoepitopes. Percentile ranks were computed using the recommended prediction method by IEDB, i.e. either a consensus score derived from a series of prediction methods such as artificial neural networks (ANN), stabilized matrix method (SMM) and scoring matrices derived from combinatorial peptide libraries (CombLib) [107], or a single score from NetMHCpan [45]. A low percentile rank indicates that the corresponding HLA allele is a good binder to the ALK neoepitope. REU: Rosetta Energy Units. Figure adopted from [115].

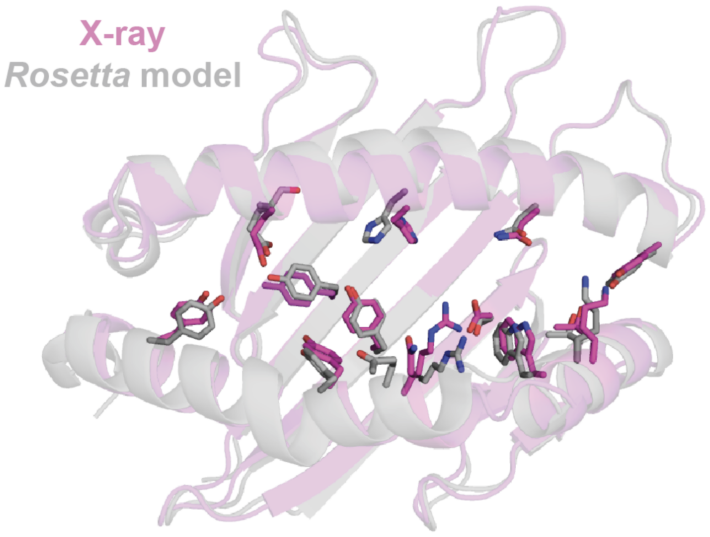


Figure 4.8: Overlay of the Rosetta modeled and X-ray determined HLA- A*01:01 groove when bound to decamer peptide. An overlay of the Rosetta model (gray) and X-ray structure (pink, PDB ID 6AT9) is shown for HLA-A*01:01 when bound to AQDIYRASYY (not shown). Side-chains are represented as sticks for representative residues in contact with the peptide (see Figure 4.5. Structures were aligned in PyMol using backbone heavy atoms (CA, C, N and O). Figure adopted from [115].

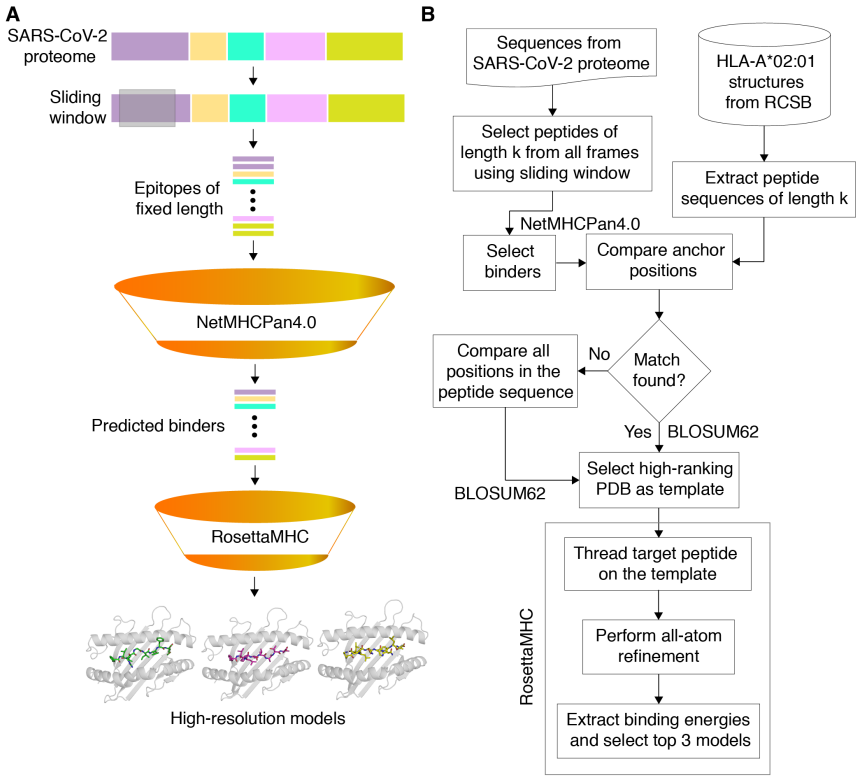


Figure 4.9: Structure-guided modeling of T cell epitopes in the SARS-CoV-2 proteome. (A) General workflow of our pipeline for structure-guided epitope ranking. (B) Protein sequences from the annotated SARS-CoV-2 proteome are used to generate peptide epitopes with a sliding window covering all frames of a fixed length (9,631 8mer, 9,621 9mer and 9,611 10mer possible peptides). Candidate peptides are first filtered by NetMHCpan-4.0 [45] to identify all predicted strong and weak binders (54 8mer, 439 9mer and 256 10mer epitopes). For rapid template matching and structure modeling, we use a local database of 236 HLA-A*02:01 X-ray structures with resolution below 3.5 Å from the PDB. Each candidate peptide is scanned against all peptide sequences of the same length in the database, and the top-scoring template is used to guide the RosettaMHC comparative modeling protocol and to compute a binding energy. Figure adopted from [76].

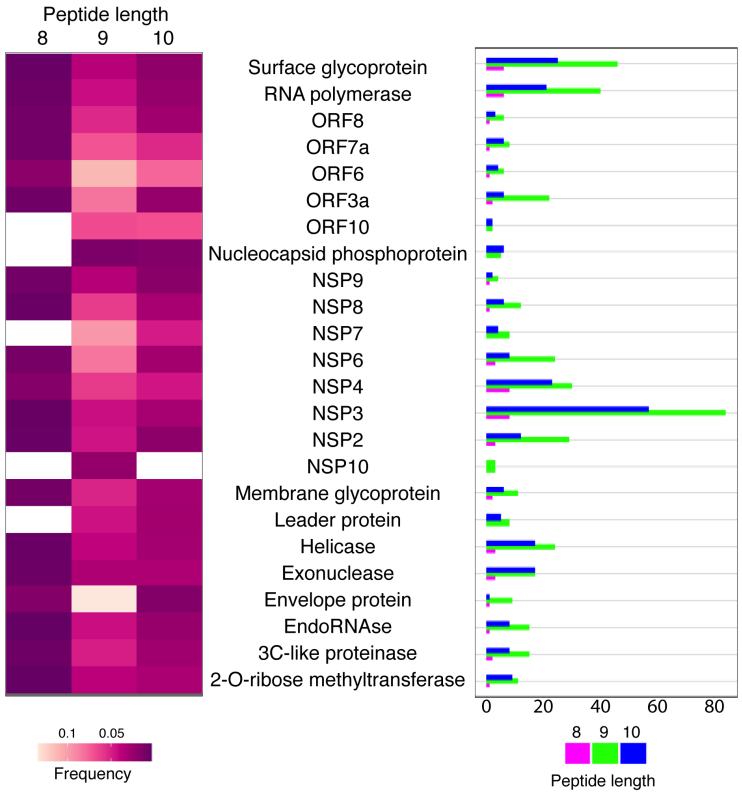


Figure 4.10: Origin of SARS-CoV-2 epitopes predicted by NetMHCpan-4.0. (Left) Heat map showing the enrichment (Frequency) of NetMHCpan-4.0 [45] filtered epitopes of lengths 8, 9 and 10, in each of the SARS-CoV-2 protein. The color scale for the frequency is shown below the heat map. (Right) Bar plot showing a total number of NetMHCpan-4.0 filtered epitopes of lengths 8, 9 and 10 (magenta, green and blue, respectively) obtained from the SARS-CoV-2 proteins. Figure adopted from [76].

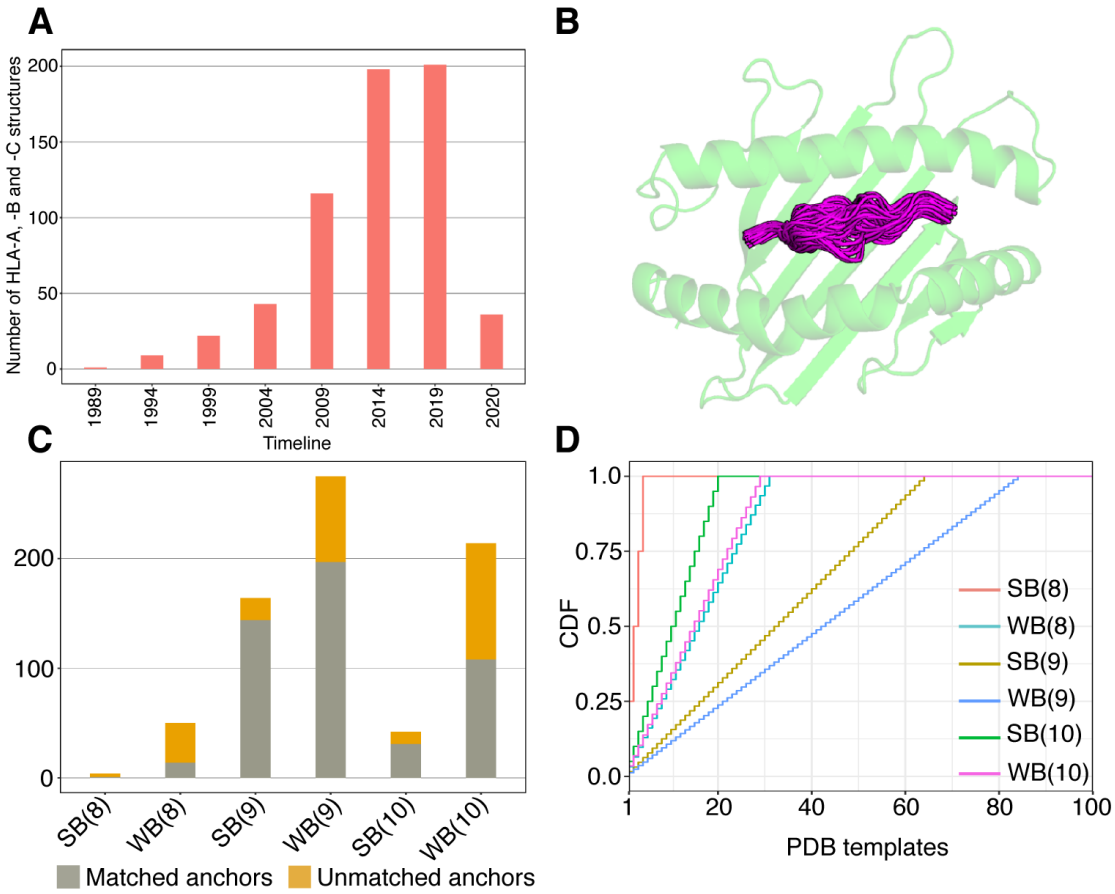


Figure 4.11: Coverage of predicted HLA-A02 epitopes by structural templates in the PDB. (A) Number of HLA structures solved over the span of 30 years. (B) Structural overlay of HLA-A*02 PDB templates displaying 9-residue peptides, with the different peptide backbone conformations shown in magenta. (C) Matching statistics for all predicted SARS-CoV-2 strong (SB) and weak binder (WB) peptides of lengths 8, 9 and 10, against an annotated database of 236 HLA-A02 X-ray structural templates derived from the PDB. (D) Plot showing cumulative distributions (CDF) for strong and weak binder peptides of lengths 8, 9 and 10, as a function of the total number of matching templates used. Figure adopted from [76].

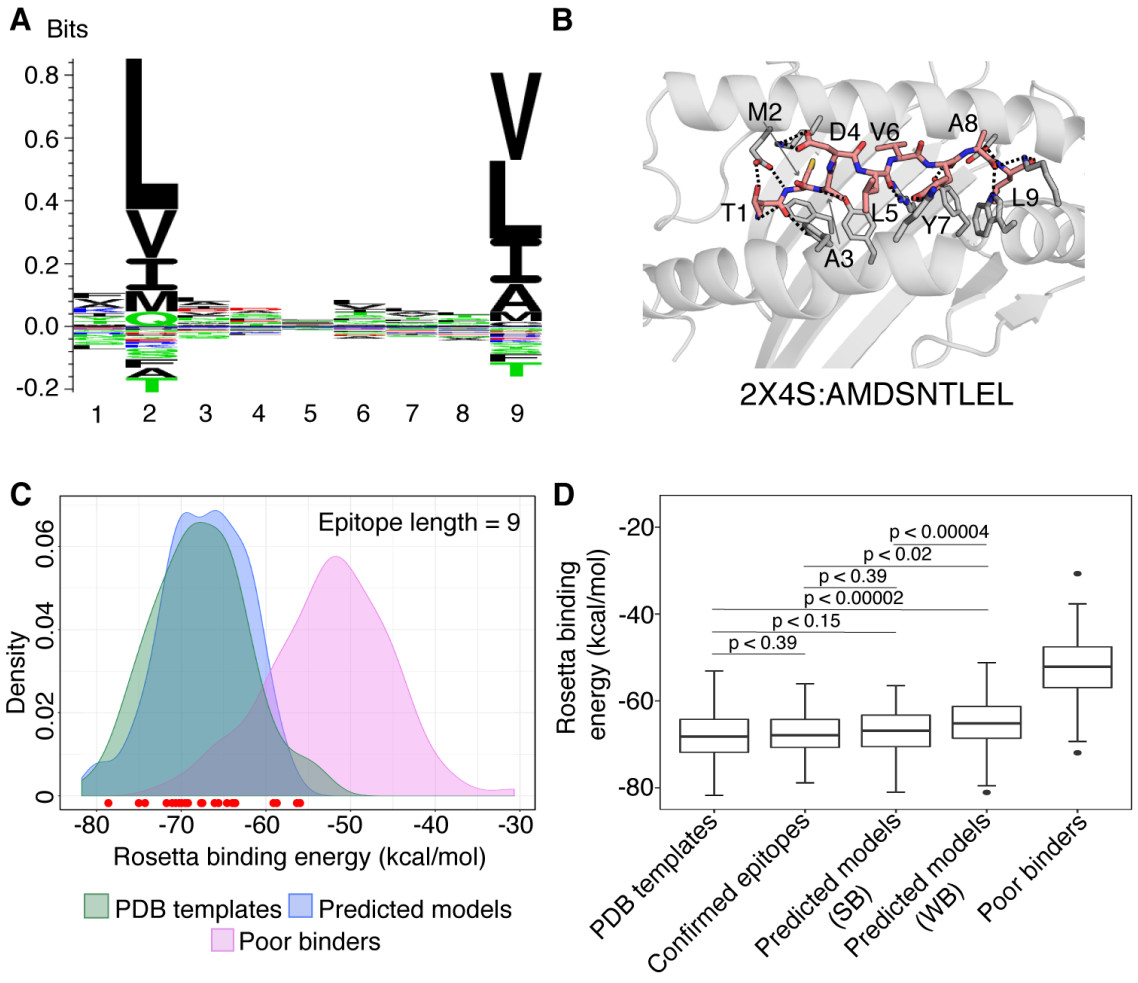

A
Bits
0.8
0.6
0.4
0.2
0.0
-0.2
1 2 3 4 5 6 7 8 9
B
M2
D4
V6
A8
T1
L5
Y7
L9
A3
2X4S:AMDSNTLEL
C
Epitope length = 9
Density
0.06
0.04
0.02
0
-80 -70 -60 -50 -40 -30
Rosetta binding energy (kcal/mol)
PDB templates
Predicted models
Poor binders
D
Rosetta binding energy (kcal/mol)
-20
-40
-60
-80
p < 0.00004
p < 0.02
p < 0.39
p < 0.00002
p < 0.15
p < 0.39
PDB templates
Confirmed epitopes
Predicted models (SB)
Predicted models (WB)
Poor binders

Figure 4.12

Figure 4.12 *(previous page)*: Summary of RosettaMHC modeling results for SARS-CoV-2 peptide epitopes. (A) Sequence logo from the 164 top ranking epitopes in the SARS-CoV-2 genome, predicted by NetMHCpan-4.0. (B) The top 9mer epitope in our refined set, TMADLVYAL, from RNA polymerase. Dotted lines indicate polar contacts between peptide and heavy chain residues, with peptide residues labelled. The template PDB ID and original peptide used for modeling the target peptide is indicated below the model. (C) Density plot showing the distribution of average Rosetta binding energies (kcal/mol) for all epitopes of length 9. Distributions reflect 93 PDB templates (green), 164 strong binder epitopes (according to NetMHCpan-4.0) (blue), and 100 SARS-CoV-2 peptides classified as poor binders by NetMHCpan-4.0 modeled using the PDB templates and used as a reference set for sub-optimal binders (Poor binders; pink). The binding energies of models generated for 28 confirmed SARS T cell epitopes from the IEDB and ViPR [30, 88, 42] are indicated by circles at the bottom of the plot. Red circles indicate epitopes that lie within the distribution of refined PDB templates. (D) Box plots showing distribution of average binding energies for 93 PDB templates, 100 sub-optimal SARS-CoV-2 peptides, 28 confirmed epitopes [30, 88, 42] and RosettaMHC models for 164 strong (SB) and 275 weak (WB) binder 9mer epitopes predicted from the SARS-CoV-2 proteome using NetMHCpan-4.0. An unpaired Mann-Whitney U test was performed for relevant pairs of distributions and their statistical significance described by the p-values (where, $p < 0.1$ is considered statistically significant) are (i) PDB templates and strong binders: $p < 0.15$ (ii) PDB templates and confirmed binders: $p < 0.39$ (iii) PDB templates and weak binders: $p < 0.00002$ (iv) confirmed epitopes and strong binders: $p < 0.39$, and (v) confirmed epitopes and weak binders: $p < 0.02$, and (vi) strong and weak binders: $p < 0.00004$, are shown inside the plot. The sequence logo was generated using Seq2Logo [112]. Figure adopted from [76].

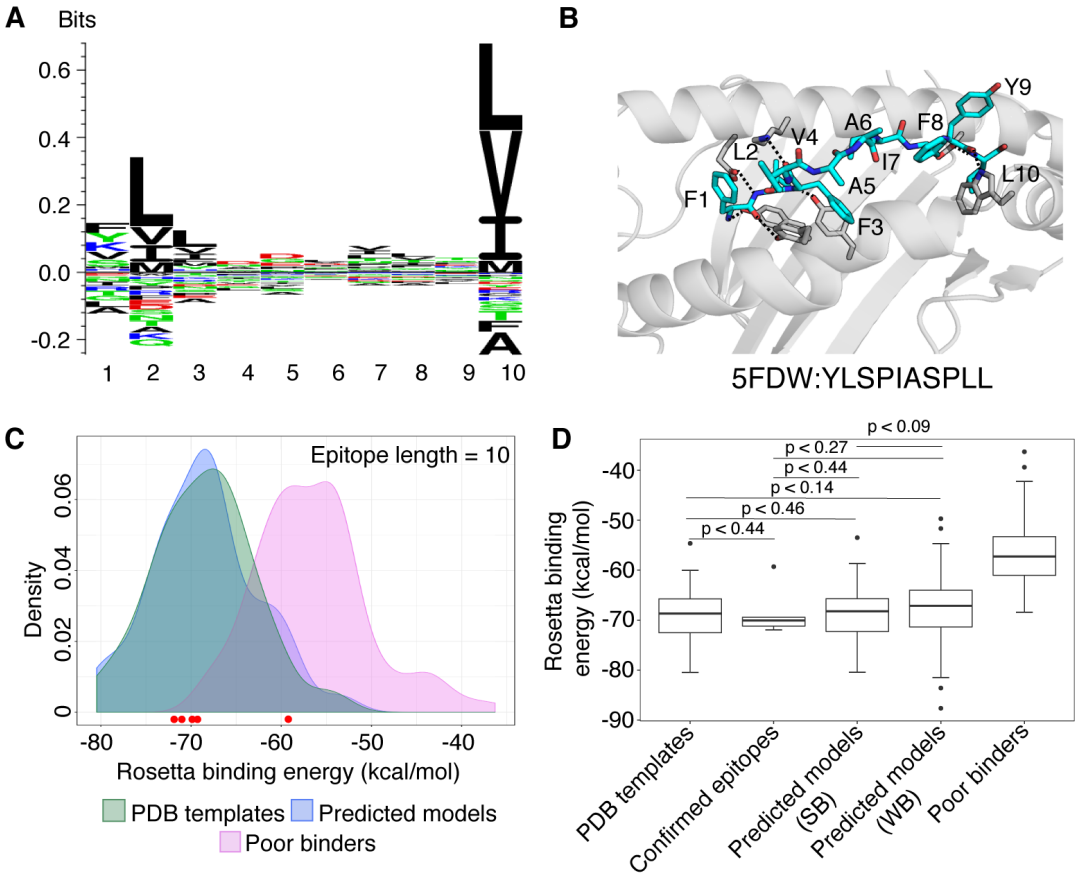

A
Bits
0.6
0.4
0.2
0.0
-0.2
1 2 3 4 5 6 7 8 9 10
B
Y9
A6
F8
V4
L2
I7
L10
A5
F1
F3
5FDW:YLSPIASPLL
C
Epitope length = 10
Density
0.06
0.04
0.02
0
-80 -70 -60 -50 -40
Rosetta binding energy (kcal/mol)
PDB templates
Predicted models
Poor binders
D
p < 0.09
p < 0.27
p < 0.44
p < 0.14
p < 0.46
p < 0.44
Rosetta binding energy (kcal/mol)
-40
-50
-60
-70
-80
-90
PDB templates
Confirmed epitopes
Predicted models (SB)
Predicted models (WB)
Poor binders

Figure 4.13

Figure 4.13 *(previous page)*: RosettaMHC modeling results for SARS-CoV-2 epitopes of length 10 (A) Sequence logo from the 42 top ranking epitopes in the SARS-CoV-2 genome, predicted by NetMHCpan-4.0. (B) The top 10mer epitope in our refined set, FLFVAAIFYL, from nsp4. Dotted lines indicate polar contacts between peptide and heavy chain residues, with peptide residues labelled. The template PDB ID and original peptide used for modeling the target peptide is indicated below the model. (C) Density plots showing distribution of average Rosetta binding energies (kcal/mol) for all epitopes of length 10. Distributions reflect 31 PDB templates (green), 42 strong binder epitopes (according to NetMHCpan-4.0) (blue), and 100 SARS-CoV-2 peptides classified as poor binders by NetMHCpan-4.0 modeled using the PDB templates and used as a reference set for sub-optimal binders (Poor binders; pink). The binding energies of models generated for 5 confirmed SARS T cell epitopes from the IEDB and ViPR [30, 88, 42] are indicated by circles at the bottom of the plot. Red circles indicate epitopes that lie within the distribution of refined PDB templates. (D) Box plots showing distribution of average binding energies for 31 PDB templates, 100 sub-optimal SARS-CoV-2 peptides, 5 confirmed epitopes [30, 88, 42] and RosettaMHC models for 42 strong (SB) and 214 weak (WB) binder 10mer epitopes predicted from the SARS-CoV-2 proteome using NetMHCpan-4.0. An unpaired Mann-Whitney U test was performed for relevant pairs of distributions and their statistical significance described by the p-values (where, $p < 0.1$ is considered statistically significant) are (i) PDB templates and strong binders: $p < 0.46$ (ii) PDB templates and confirmed binders: $p < 0.44$ (iii) PDB templates and weak binders: $p < 0.14$ (iv) confirmed epitopes and strong binders: $p < 0.44$, and (v) confirmed epitopes and weak binders: $p < 0.27$, and (vi) strong and weak binders: $p < 0.09$ are shown inside the plot. The sequence logos were generated using Seq2Logo [112]. Figure adopted from [76].

4.2.4 The Rosetta energy function generally distinguishes native-like models

To evaluate the accuracy of our models and fitness of each peptide within the HLA-A*02:01 binding groove, we computed Rosetta all-atom binding energies across all complexes modeled for different peptide sets. High binding energies can be used as an additional metric to filter low-affinity peptides in the NetMHCpan-4.0 predictions, with the caveat that high energies can be also due to incomplete optimization of the Rosetta energy function as a result of significant deviations between the target and template backbone conformations, not captured by our protocol. We performed 10 independent calculations for each peptide which may allow Rosetta's optimization protocol to sample slight changes in peptide backbone (up to 1 Å from starting structure), and the 3 lower-energy models were selected as the final ensemble and used to compute an average binding energy. The results for all 9mer and 10mer peptides are summarized in Figures 4.12C, D, and 4.13C and D while additional results for 8mers are provided through our web-interface (SARS-CoV-2 models). As a positive reference, we used

the binding energies of the idealized and relaxed PDB templates, which are at a local minimum of the Rosetta scoring function. As a reference set for sub-optimal binders, we modeled structures of peptides from SARS-CoV-2 proteome that are classified as poor binders according to NetMHCpan-4.0 (highest %rank values).

We observe a significant, favorable (approx. -15 kcal/mol) energy gap between the average binding energies computed from the refined PDB templates relative to models obtained for poor binder peptides. The binding energies for all predicted 9mers and 10mer binders show a significant overlap with the refined PDB template energies (Figures 4.12C and 4.13C). Comparison of energy distributions of epitopes that are classified as strong versus weak binders by NetMHCpan-4.0 shows a moderate bias towards lower binding energies for the strong binders and a larger spread in energies for weak binders, likely due to suboptimal residues at the P2 and P9/P10 anchor positions (Figures 4.12D and 4.13D, with a significance level, $p < 0.1$ between strong and weak binders for both 9-mers and 10-mers). As an intendent positive set, we also modeled 28 9mer and 5 10mer peptides that are homologous to peptides in the SARS viral genome and have been previously reported to bind HLA-A*02:01 in the IEDB and ViPR [30, 88, 42] databases. Inspection of Rosetta binding energies derived from models in this set shows a similar distribution to the epitopes predicted by NetMHCpan-4.0, with the energies of all the peptides falling well within the distribution of the refined PDB templates (red dots in Figures 4.12C and 4.13C).

4.2.5 Applications

Electrostatic surface analysis of SARS-CoV-2 epitopes relative to homologous peptides derived from common cold coronavirus strains

T-cell responses to megapools of viral peptides have been observed in individuals not exposed to SARS-CoV-2, thus providing evidence for cross-reactivity [132, 12] of T cells with similar epitopes expressed by homologous coronavirus strains [14, 31]. Therefore, we analyzed the electrostatic surfaces of our models relative to the models of homologous peptide/HLA-A*02:01 complexes derived from four strains of human common cold coronavirus (229E, HKU1, NL63, OC43) and identified putative SARS-CoV-2 specific and cross-reactive epitopes.

To perform a structure-based classification of SARS-CoV-2 peptide/HLA-A*02:01 complexes

and compare their surfaces to homologous peptides from common cold coronavirus strains, we (i) aligned respective protein sequences (specifically, orf1ab, membrane, spike, envelope and nucleocapsid proteins) from all the strains using Clustal Omega [108], (ii) extracted 395 (out of 439) epitopes of length 9 from common cold coronavirus strains based on sequence homology with SARS-CoV-2 binders predicted by NetMHCpan-4.0 using default %rank cut-off values (44/439 SARS-CoV-2 epitope sequences are from proteins not considered), (iii) filtered out 141 epitopes containing insertions and deletions in the sequence alignment and those that do not have homologous peptides in one or many strains of common cold coronavirus, (iv) modeled structures of the remaining 254 common cold coronavirus peptide/HLA-A*02:01 complexes from each strain using RosettaMHC, and (v) performed a comparison of surface electrostatic potentials between each SARS-CoV-2 epitope and its corresponding common cold coronavirus peptide homologs using multipipsa4.0.2 [114]. The multipipsa4.0.2 software applies the Adaptive Poisson-Boltzmann Solver (or APBS) [9] method to first compute electrostatic potentials, and then compares the potentials using the Protein Interaction Property Similarity Analysis (or PIPSA) protocol [124]. The side chains of the modeled complexes are protonated using PROPKA [84] followed by the assignment of atom charges and radii using the Amber force field [20] at a pH of 7.2. The electrostatic potentials of the structures are calculated by solving a linear Poisson-Boltzmann equation for 129 points on a cubic grid using 150 mM ionic strength at 298.15 Kelvin with protein dielectric of 1.0, and solvent dielectric of 78 using a probe radius of 1.4 Å. Next, the PIPSA protocol compares the electrostatic potentials quantitatively by using grid points on the superimposed regions (regions are at a distance of σ from the van der Waals surface and are of thickness δ) around the complexes. The similarity between any two electrostatic surfaces is captured by the Hodgkin similarity index (HSI, ranges from -1 to 1, where -1, 0, and 1 indicate electrostatic anticorrelation, no correlation and electrostatic identity respectively) [38], which is converted into a distance measure, D ($D = \sqrt{(2 - 2HSI)}$), that assigns values between 0 and 2 (0: identity, 1: no correlation and 2: anticorrelation). For our study, we have used 4 thickness (δ) and a distance of 3 from the molecular surface (σ) [64].

From 395 predicted strong binders from our set, 141 peptides are exclusive to SARS-CoV-2, since there are no homologous sequences present in the four common coronavirus strains considered here, or the corresponding sequences in the other four genomes have insertions or deletions. To identify cross-reactivity according to structural and surface features, we first used

RosettaMHC to model peptide/HLA-A*02:01 complex structures for all homologous peptides, in addition to our previously described models for SARS-CoV-2 (Figure 4.14A). As mentioned previously, we computed surface electrostatic potentials for each model using APBS [9], followed by a pairwise comparison of potentials computed for the four homologous structures relative to each SARS-CoV-2 peptide using PIPSA [124], which is summarized in four distance scores for each peptide (Figure 4.14B). From the examination of similarity scores of models, we found that (i) peptide SLAIDAYPL from orf1ab has highly conserved sequence and surface features across all coronavirus strains (distance score = 0), and therefore T cells specific for this epitope should be highly cross-reactive across different strains (ii) epitopes AIMTRCLAV, YLGGMSYYC, FVDGVPFVV, RIIPARARV, RILGAGCFV, RLANECAQV, SVFNICQAV, IFVDGVPFV, and GVAPGTAVL from orf1ab are conserved with one or more common strains, and are putatively cross-reactive (distance score $\leq$ 0.3) (Figure 4.14B), and (iii) there is no apparent correlation between SARS-CoV-2 and common cold coronavirus pMHC surface features for ALLSDLQDL (orf1ab), QLNRALTGI (spike), MLAKALRKV (orf1ab) and KIYSKHTPI (spike) epitopes (distance score $>$ 0.8) (Figure 4.14C). Six epitopes, known to induce CD8 T cell responses in COVID-19 patients and healthy donors have distance scores ranging from 0.5 to 0.9 (Table 4.1) [17, 72] suggesting that the SARS-CoV-2 epitopes are possibly cross-reactive.

Electrostatic potentials calculated from our models also allow us to compare distinct surfaces for TCR recognition between different high-affinity epitopes, as demonstrated for the four top-scoring models by Rosetta binding energy (Figure 4.15A). Here, PIPSA analyses of electrostatic potentials of these models allowed us to cluster them into two groups, (i) TMADLVYAL and NLIDSYFVV, and (ii) KLWAQVCQL and FLAFVVFLL, where the surface exposed residues at P2-P8 positions of the (i) and (ii) groups exhibit moderately negative and positive charges, respectively (Figure 4.15). Full classification and ranking of all binders in our set on the basis of their molecular surface features would further enable the selection of the most diverse panel of peptides for high-throughput pMHC-tetramer library generation which can be used to identify immunodominant epitopes, an avenue which is currently actively explored in our lab [85]. Moreover, tetramer screening of T cells from COVID-19 patients, recovered individuals and healthy donors can be used to identify critical gaps in the T cell repertoire of high-risk groups, and to design epitope DNA strings for vaccine development.

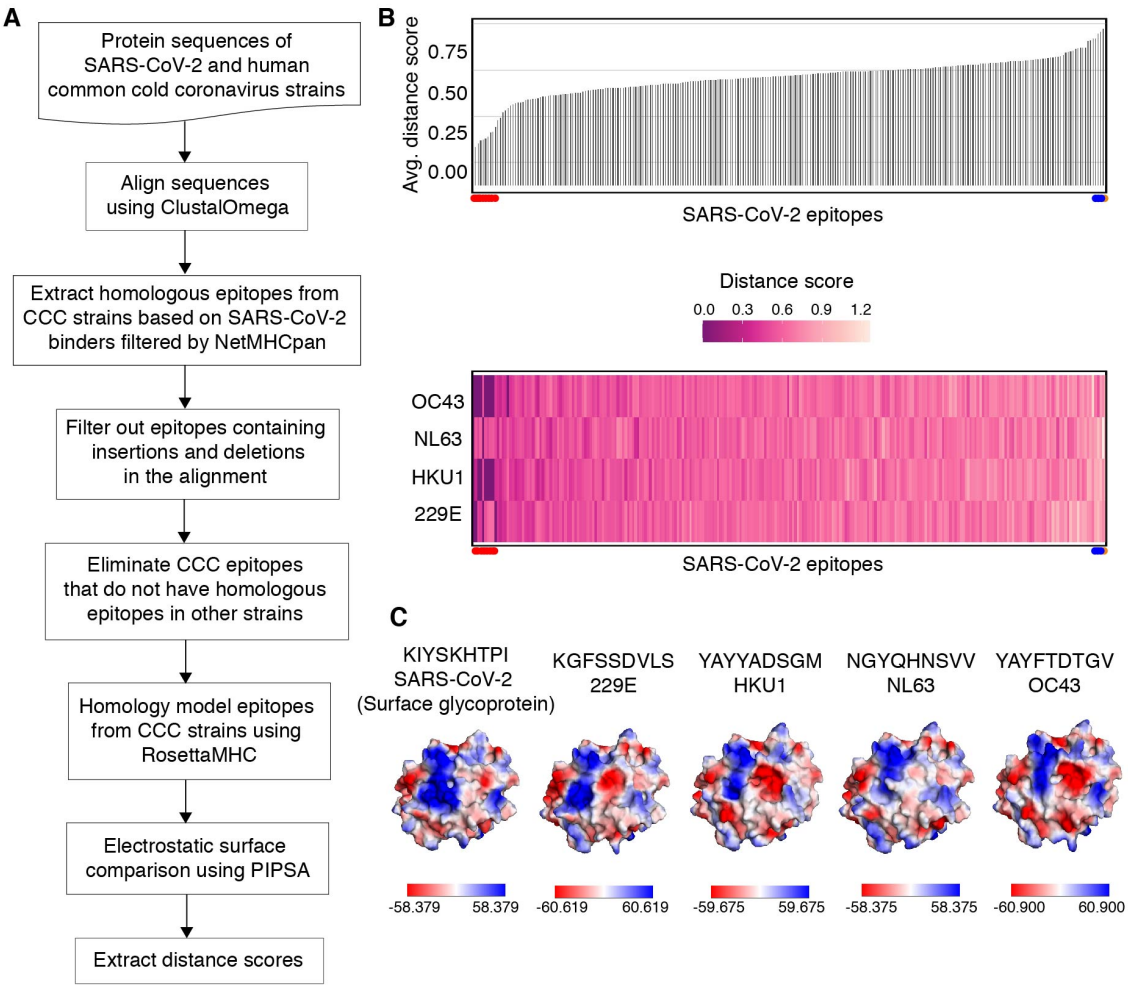

A
Protein sequences of SARS-CoV-2 and human common cold coronavirus strains
Align sequences using ClustalOmega
Extract homologous epitopes from CCC strains based on SARS-CoV-2 binders filtered by NetMHCpan
Filter out epitopes containing insertions and deletions in the alignment
Eliminate CCC epitopes that do not have homologous epitopes in other strains
Homology model epitopes from CCC strains using RosettaMHC
Electrostatic surface comparison using PIPSA
Extract distance scores
B
Avg. distance score
0.75
0.50
0.25
0.00
SARS-CoV-2 epitopes
Distance score
0.0 0.3 0.6 0.9 1.2
OC43
NL63
HKU1
229E
SARS-CoV-2 epitopes
C
KIYSKHTPI
SARS-CoV-2
(Surface glycoprotein)
KGFSSDVLS
229E
YAYYADSGM
HKU1
NGYQHNSVV
NL63
YAYFTDTGV
OC43
-58.379 58.379
-60.619 60.619
-59.675 59.675
-58.375 58.375
-60.900 60.900

Figure 4.14 *(previous page)*: Electrostatic surface similarity scores of SARS-CoV-2 epitopes and peptides derived from four human common cold coronavirus strains. (A) Flow diagram describing the calculation of scores to measure similarity between pMHC models of SARS-CoV-2 epitopes, relative to their homologous epitopes in four human Common Cold Coronavirus (CCC) strains based on electrostatic surface potentials. (B) A bar plot (top) showing the average electrostatic similarity score obtained using PIPSA [124] (Y-axis) measured between SARS-CoV-2 epitopes and homologous common cold coronavirus peptides. An electrostatic distance score ranging between 0 and 2, where 0, 1 and 2 indicate electrostatic identity, no correlation and anticorrelation. A heat map (bottom) shows the distance scores between SARS-CoV-2 epitopes (X-axis) and homologous peptides from each strain of common cold coronavirus (Y-axis). The distance scores are indicated by the colored scale (from 0 to 1.2). The SARS-CoV-2 epitopes SLAIDAYPL, AIMTRCLAV, YLGGMSYYC, FVDGVPFVV, RIIPARARV, RILGAGCFV, RLANECAQV, SVFNICQAV, IFVDGVPFV, GVAPGTAVL that share similar electrostatic surfaces with homologous peptides from common cold coronaviruses are highlighted using red dots on the left in top and bottom plots. Similarly, the epitopes ALLSDLQDL, QLNRALTGI, MLAKALRKV (blue dots) and KIYSKHTPI (orange dot) that exhibit no apparent correlation or are electrostatically dissimilar are shown on the right in top and bottom plots. (C) A SARS-CoV-2 epitope, KIYSKHTPI from surface/spike glycoprotein has electrostatic surface similarity score of 0.794, 0.964, 0.784, 0.914 with homologous peptides derived from 229E, HKU1, NL63 and OC43 strains of common cold coronavirus respectively. The peptide sequences along with strains are indicated on top of each electrostatic surface. Solvent-accessible surface representation with electrostatic potential in the indicated ranges (down to −61 kcal/(mole) in red and up to +61 kcal/(mole) in blue) were calculated using the APBS solver [44] in made available through multipipsa4.0.2 software [114]. Pymol [1] is used to generate the electrostatic surfaces for visualization. All calculations were performed at 150 mM ionic strength, 298.15 Kelvin, pH 7.2, protein dielectric 1.0, and solvent dielectric 78 with a probe radius of 1.4 . Figure adopted from [76].

4.2.6 Accessibility of SARS-CoV-2 peptide/HLA-A*02:01 models through the UCSC genome browser

Immunology tracks (annotations) contain SARS-CoV-2 protein epitopes reported in the literature [26]. Included are epitopes that have been predicted and/or validated to be immunogenic. These data can be overlaid with structural and variation information to track mutations that overlap with potential therapeutic targets. These tracks (IEDB predictions, Poran HLA I and Poran HLA II) also display epitopes recognized by CD8+ or CD4+ T cells when presented by human leukocyte antigen (HLA) molecules on host cells [30, 91]. When possible, the latter are organized according to the HLA allele of the host used in their presentation. For the track CD8 RosettaMHC, interactive 3D models from Rosetta are available through clicking on the annotated epitope [76]. These tracks will be updated as validation and identification of epitopes

Table 4.1: SARS-CoV-2 CD8+ cross-reactive T-cell epitopes known to induce immune responses in COVID-19 patients and healthy donors. Table adopted from [76].

Epitope	Homologous peptide sequences from common cold coronavirus strains	BLOSUM62 sequence similarity score	Electrostatic surface distance score
RLNEVAKNL	KLQTLIDNI (229E) LIQESIKSL (HKU1) ELQGLIDQI (NL63) RLQEAIKVL (OC43)	12 12 3 19	0.772 0.573 0.801 0.608
YLQPRTFLL	ALASYADVL (229E) PLSKRQYLL (HKU1) AFATFVDVL (NL63) PLTSRQYLL (OC43)	0 15 -5 14	0.648 0.463 0.703 0.506
FLHVTYVPA	FLHTVLLPT (229E) FMHFSYKPI (HKU1) FLHTVLLPT (NL63) FIHFSYVPT (OC43)	25 27 25 34	0.564 0.540 0.564 0.592
RLDKVEAEV	RLDTIQADQ (229E) RLDALEAQV (HKU1) RLDSIQADQ (NL63) RLDALEAEA (OC43)	23 30 24 29	0.523 0.500 0.567 0.522
LLFNKVTLA	ILFSKLVTS (229E) LLFDKVKLS (HKU1) LLFSKVVTS (NL63) LLFDKVKLS (OC43)	19 28 24 28	0.510 0.542 0.566 0.542
KIYSKHTPI	KGFSSDVLS (229E) YAYYADSGM (HKU1) NGYQHNSVV (NL63) YAYFTDTGV (OC43)	2 0 5 6	0.794 0.964 0.784 0.914

continues.

4.3 Limitations of homology modeling using single template strategy

RosettaMHC relies on the availability of accurate structural templates for modeling a pMHC of interest. Selection of incorrect templates may result in steric hindrance of residues in the peptides with the residues in the MHC groove leading to sub-optimal models and binding energies. For instance, several RosettaMHC models containing bulky/charged arginine residues in their peptide sequence had their corresponding side chains buried within the MHC groove instead of orienting towards the TCR as highlighted by five HLA-A*02 structures in the PDB (Figure 4.16).

This limitation can be overcome by sampling diverse peptide backbone conformations. Here, we can employ FlexPepDock or fragment-based approaches, however using these methods could impact throughput and accuracy.

In addition, we have a representative number of crystal structures in the PDB for the frequent alleles such as HLA-A*02:01 and no structural templates for underrepresented alleles (lack of generality). Therefore, we need a mechanism to derive structural templates from existing PDB structures so that RosettaMHC can model pan-allelic complexes.

4.3.1 Analysis of pMHC structures in the PDB

To overcome the limitations listed in the previous sections, we aimed to select optimal templates from the PDB. From superposition of pMHC structures in the PDB, we found that peptide backbones should be strictly less than 1 Å to preserve correct backbone and side chains across the peptide (Figure 4.17A). To model sub-angstrom structures, we examined the distribution of RMSD values of the peptide backbone pairs and found that a randomly selected peptide backbone for a query peptide sequence in the PDB is above 1 Å for approx. 60% of the cases (Figure 4.17B). However, we found alternative sub-angstrom peptide backbones for every peptide of a pMHC complex in the PDB (Figure 4.17C). Moreover, the peptide backbones that are closer to each other may be displayed by different alleles which makes it possible to model any peptide bound to any allele as highlighted in panel D of Figure 4.17.

Through these preliminary analyses of peptide backbones in the PDB, we reasoned that we have sufficient backbone conformations naturally sampled to perform homology modeling accurately. While we can model a peptide of interest using multiple-templates, particularly with all available peptide conformations in the PDB, we believe that this strategy drastically reduces throughput. Therefore, we devised a new strategy based on the hypothesis that similar MHC grooves exhibit similar peptide repertoires and discuss this approach in detail in the subsequent sections.

4.4 Multi-template homology modeling of peptide/MHC-I complexes

To sample diverse set of peptide backbone conformations, we select multiple PDB templates based on high sequence identity to the MHC groove (MHC groove is formed by the residues within a distance of 3.5 Å from the residues in the peptide) in preference to the whole MHC heavy chain sequence consisting of $\alpha 1$ and $\alpha 2$ domains, also referred to as the groove-based approach. Specifically, for each allele, we select templates from 318 MHC class I structures in PDB (17 8mers, 232 9mers and 69 10mers) as candidates if the MHC groove residues match with at least 70% identity in the sequence alignment of MHC heavy chains and display peptides with identical lengths (Figure 4.18). For instance, to select 2GTZ as a template to model HLA-A*02:01 with a peptide of interest, we align the MHC heavy chain sequences of 2GTZ and HLA-A*02:01 obtained from IMGT/HLA database [97]. Further, we select the residues (referred to as groove residues) in the MHC heavy chain of 2GTZ that are within 3.5 from all the residues in the peptide displayed by 2GTZ. We compare the groove residues of 2GTZ with residues in HLA-A*02:01 that agree in the alignment (Figure 4.18A). Finally, we select 2GTZ as a template for modeling if the groove residues match with at least 70% sequence identity. Using this criteria, we obtain a significantly higher number and diverse set of templates with naturally sampled backbones for the most frequently found alleles in the Caucasian (CAU), Hispanic (HIS), African American (AFA) and Asia Pacific Islander (API) populations (Table 4.2 and Figure 4.18B) [29]. We restrict multi-template homology modeling to peptides of length 9.

4.4.1 Evaluation of RosettaMHC with templates selected using MHC groove sequence identity

We evaluated RosettaMHC by modeling (i) a blind target, PHOX2B peptide/HLA-A*24:02 complex [130], (ii) peptides of length 9 in the PDB displayed by all the HLA-A, -B, -C, -E and -G alleles, and (iii) five SARS-CoV-2 derived peptides bound to HLA-A*02:01 whose structures were deposited in the PDB recently [110]. To perform modeling, we first select templates for each allele using the groove-based approach (Figure 4.18A), homology model each pMHC of interest on all the templates that have similar grooves, refine the conformations in the Rosetta

Table 4.2: Most frequent alleles in CAU, HIS, AFA, and API population [29].

Super type	Total	Alleles
HLA-A	21	A*02:01, A*23:01, A*03:01, A*30:01, A*30:02, A*68:02, A*74:01, A*33:03, A*01:01, A*02:02, A*02:07, A*02:03, A*26:01, A*02:06, A*11:01, A*24:02, A*32:01, A*68:01, A*29:02, A*31:01, A*25:01
HLA-B	21	B*53:01, B*07:02, B*35:01, B*15:03, B*42:01, B*45:01, B*44:03, B*58:02, B*58:01, B*08:01, B*40:01, B*51:01, B*46:01, B*40:06, B*15:02, B*52:01, B*15:01, B*44:02, B*18:01, B*14:02, B*40:02
HLA-C	17	C*04:01, C*07:01, C*16:01, C*02:02, C*06:02, C*07:02, C*17:01, C*03:04, C*08:02, C*05:01, C*01:02, C*08:01, C*15:02, C*03:02, C*12:02, C*12:03, C*03:03

force field and extract binding energies.

PHOX2B peptide/HLA-A*24:02 structural models

PHOX2B is an important target for childhood cancers [130]. Here, we model the structure of PHOX2B peptide (QYNPIRTTF) displayed by HLA-A*24:02 [130]. To model PHOX2B peptide/HLA-A*24:02 complex we utilized all alleles as templates selected using the groove-based approach. We filtered top 5 best energy producing models and found that the top scoring model is optimal and has a peptide backbone heavy-atom RMSD value of 0.69 Å (Figure 4.19). The energy gap between the top scoring and the second best scoring models is significant with $\approx$-10

energy units suggesting that Rosetta energy function has the capacity to distinguish native-like models (Figure 4.19).

Peptides derived from pMHC complexes in the PDB

To evaluate our method further, we homology modeled 225 peptide sequences of length 9 derived from 43 alleles across HLA-A, -B, -C, -E and -G super type structures in the PDB using all the templates (except the natives) from the MHC groove-based template library and examined if the Rosetta energy function could distinguish native from non-native templates consistently (Figure 4.20). From our analyses, we found that approx. 63% of the models contain peptide backbones within 1 Å RMSD from their natives (Figure 4.20A). Moreover, the number of structures for specific super types such as HLA-A*02 are over represented in the PDB, despite that we obtain accurate models for non-HLA-A*02 super types (Figure 4.20B); peptides displayed by all the alleles modeled with sub-angstrom accuracy and are top scoring by Rosetta energy). However, it is worth noting that we can recover sub-angstrom backbones for the remaining 27% of the peptides within top 5 structures scored by energy, 6% within top 10 and remaining among all the models generated using multi-template homology modeling (Figure 4.20C). A general trend we observe is that the best models (or lowest RMSD models) have higher binding energies compared to the top scoring models identified by Rosetta even though energies from both models overlap significantly with the energies of optimized native structures (Figure 4.20D). These results suggest that the energy function is low-resolution and cannot fully discriminate between accurate and inaccurate peptide backbones (for at least 37% of the peptides).

4.4.2 Multi-template modeling samples accurate conformations relative to single-template modeling for SARS-CoV-2 derived peptides

Recently, crystal structures of five SARS-CoV-2 peptide/HLA-A*02:01 complexes were published (PDB IDs: 7KGO, 7KGP, 7KGQ, 7KGR, and 7KGS), we treated these peptides as blind targets and modeled them using RosettaMHC with groove-based template selection approach [110]. Specifically, each peptide was modeled with 132 templates (that matched with HLA-A*02:01 with at least 70% sequence identity among the MHC groove residues). We compared the peptide backbones of the models with the X-ray structures and found that four out of five

structures were modeled with sub-angstrom accuracy (Figure 4.21; pale green models). In comparison with structures computed using single template modeling based on peptide sequence, multi-template modeling helps sample diverse and more accurate backbones as demonstrated in Figure 4.21 (crimson and pale green models). While multi-template modeling holds promise in selecting sub-angstrom pMHC structures for 63% of the peptides (based on our benchmark dataset), the energy function is not high-resolution. Moreover, we have much closer peptide backbones that have higher energies compared to those indicated as lowest energy models as shown in the energy vs RMSD plot (Figure 4.21; right panels). In an ideal case, we expect models closer to the native (or lower RMSD models) to have lower energy, therefore, the energy vs RMSD plots guide us towards the bottom of the energy funnel where the native-like models are found. In the case of SARS-CoV-2 peptides, we do not have a clear signal between closer and farther peptides suggesting that we need to either improve the energy function or utilize a complementary approach that can help us identify accurate peptide backbones or eliminate incorrect templates (Figure 4.21; energy funnels).

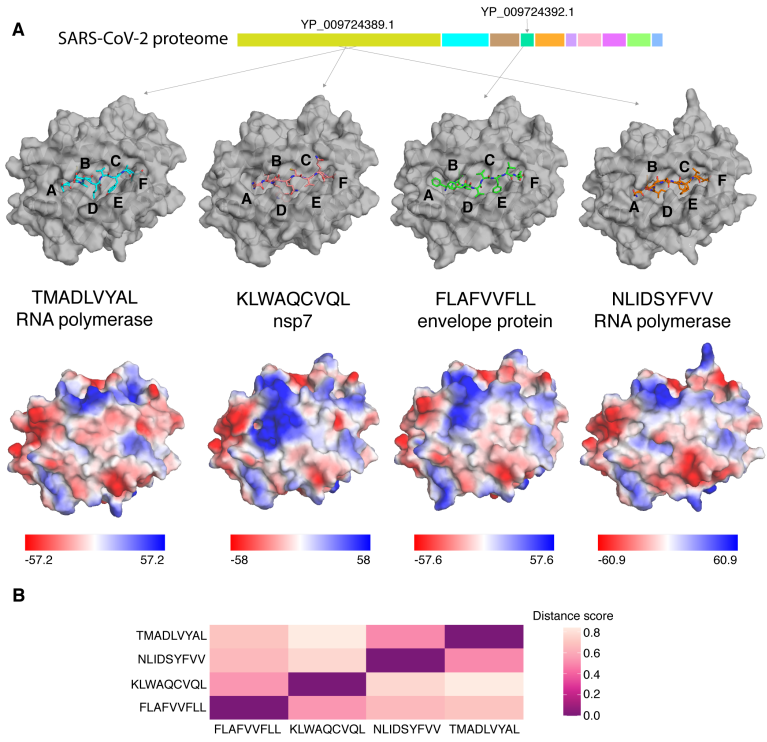


Figure 4.15: Variability in TCR recognition surfaces of HLA-A*02 with different high-affinity peptides. (A) Molecular surfaces of SARS-CoV-2/HLA-A*02:01 RosettaMHC models are shown for four top-scoring epitopes (ranked by Rosetta binding energy from left to right) captured in the A, B, C, D, E and F pockets of the MHC-I groove (top panel). The origins of the peptide epitopes in the 30 kbp SARS-CoV-2 genome are noted. Electrostatic surfaces computed for the same models are shown in the bottom panel [9]. Solvent-accessible surface representation with electrostatic potential in the indicated ranges (down to −61 kcal/(mole) in red and up to +61 kcal/(mole) in blue) were calculated using the APBS solver [9] in the mulitpipsa4.0.2 [114] software. The visual representations of the electrostatic surfaces for the models are obtained from Pymol [1]. All calculations were performed at 150 mM ionic strength, 298.15 Kelvin, pH 7.2, protein dielectric 1.0, and solvent dielectric 78 with a probe radius of 1.4 . (B) A heat map showing electrostatic surface distance scores among four top-scoring epitopes listed in (A), computed using PIPSA [124, 114]. A distance score key is shown on the right, where a score of 0 indicates electrostatic identity, 1 indicates no correlation and 2 indicates electrostatic anticorrelation. Figure adopted from [76].

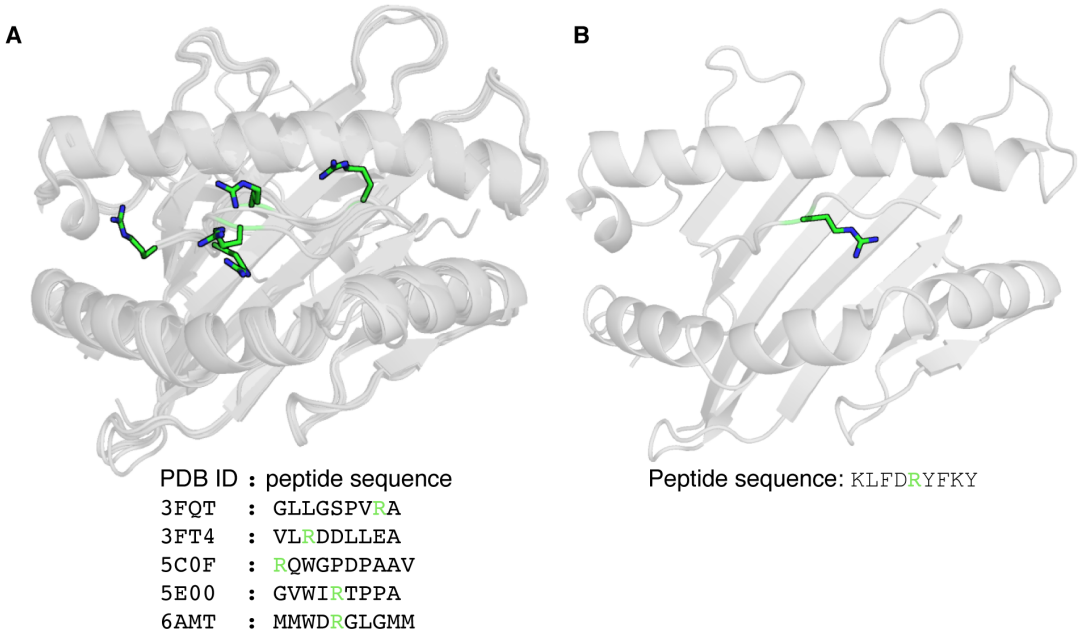


Figure 4.16: Sub-optimal placement of side chains due to the selection of incorrect peptide backbone. (A) Overlay of five HLA-A*02 PDB structures containing arginine residues (green sticks) in their peptide sequences. The PDB IDs and the peptide sequences are 3FQT (GLLGSPVRA), 3FT4 (VLRDDLLEA), 5C0F (RQWGPDPAAV), 5E00 (GVWIRTPPA), 6AMT (MMWDRGLGMM). (B) RosettaMHC model of SARS-CoV-2 epitope KLFDRYFKY/HLA-A*02:01 complex with arginine of the peptide buried in the MHC groove.

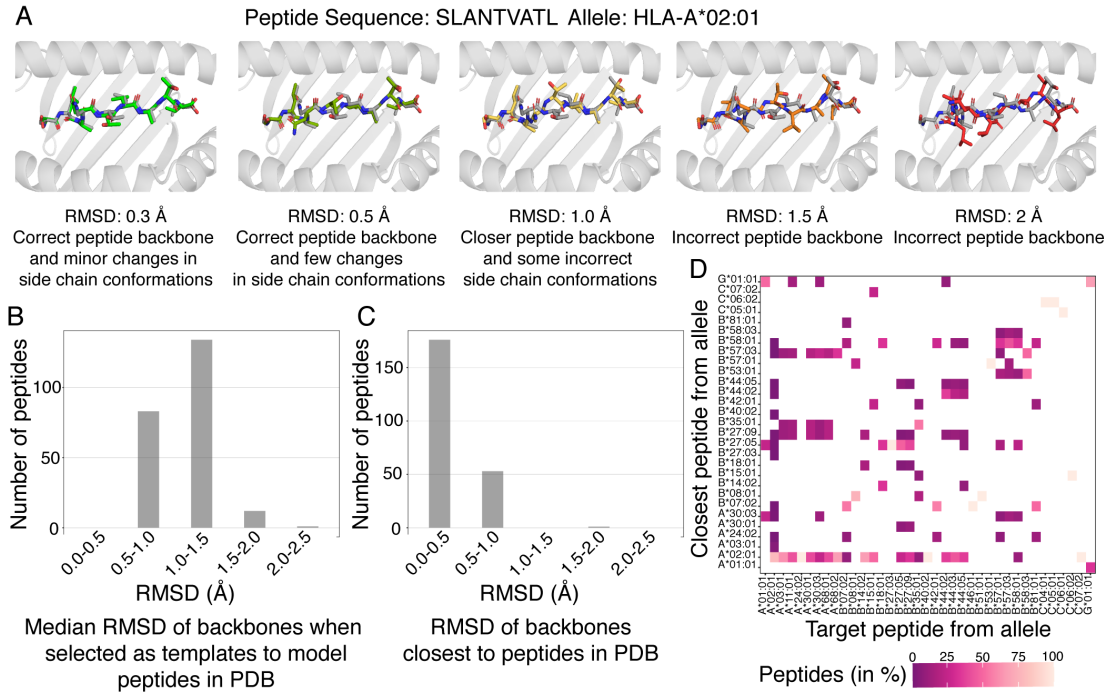


Figure 4.17: Pairwise RMSD analyses of pMHC structures in the PDB. (A) Overlay of five SLANTVATL/HLA-A*02 models with crystal structure in the PDB demonstrating the differences in the peptide backbones in terms of RMSD values (in Å). The models with RMSD value greater than 1 Å lose backbone and side-chain accuracy. (B) Distribution showing median pairwise RMSD values between peptide backbones derived from 9mer pMHC structures in the PDB. (C) Distribution showing RMSD values of peptide backbones with their closest templates in the PDB. (D) Heatmap showing the origin of closest peptide backbones in the PDB for different alleles. Figure adopted from a manuscript in preparation.

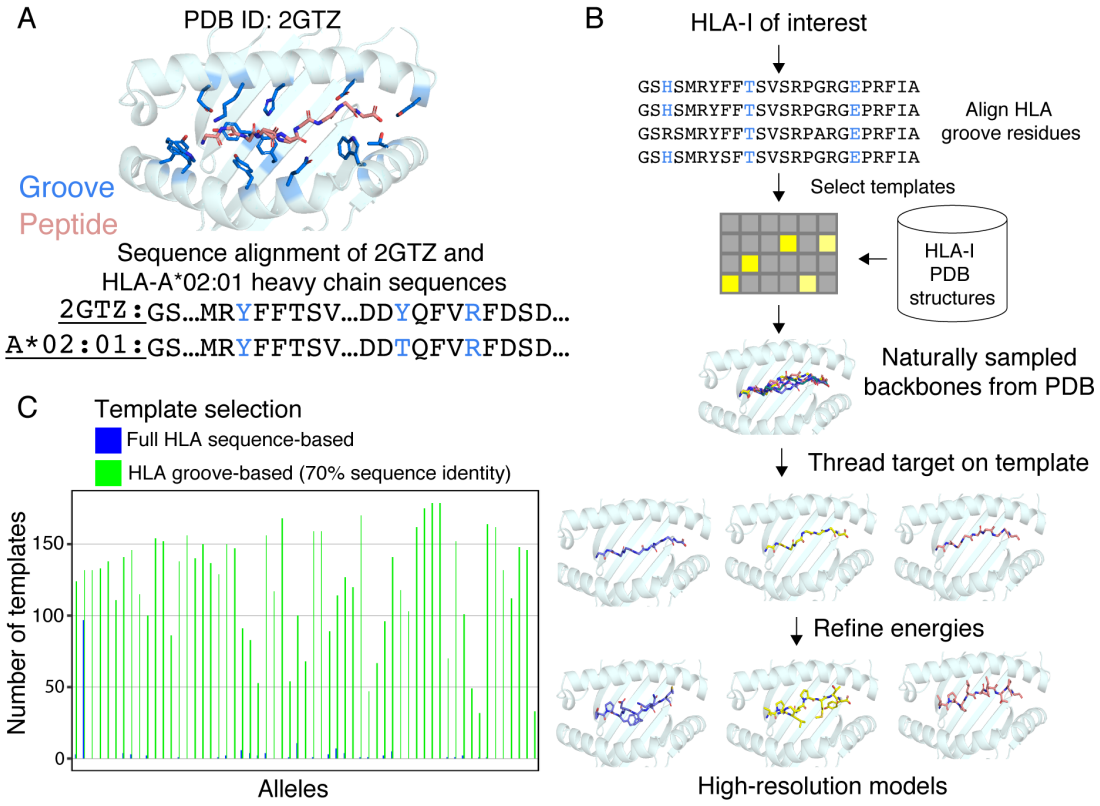


Figure 4.18: Multi-template library for homology modeling. (A) Peptide/MHC structure (PDB ID: 2GTZ) highlighting the peptide residues in salmon and the groove residues in blue. Sequence alignment between 2GTZ and HLA-A*02:01 heavy chains focusing the groove residues in blue. (B) RosettaMHC workflow showing the mechanism of PDB template selection based on high sequence identity to the MHC groove and homology modeling using multiple templates. (C) Bar plot showing the number of PDB templates displaying peptides of length 9 as a function of the most frequent alleles (Table 4.2). The number of templates selected using full sequence-based (blue) and MHC groove-based with 70% sequence identity (green) criteria are shown. Figure adopted from a manuscript in preparation.

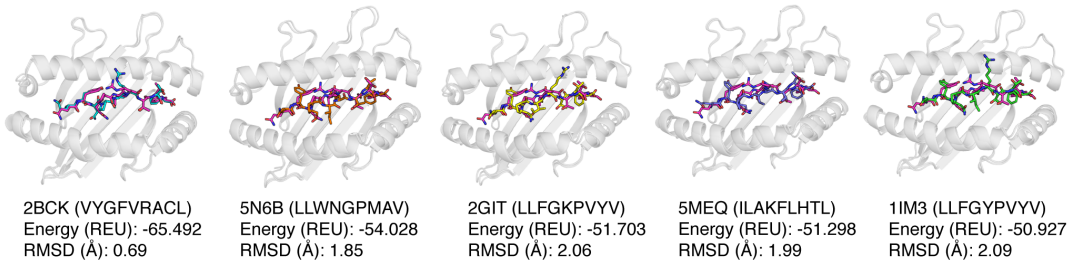


Figure 4.19: RosettaMHC identifies an optimal template for PHOX2B peptide/HLA-A*24:02 complex [130]. (A) Overlay of top 5 scoring RosettaMHC structures modeled using a new template library. HLA-A*24:02 structures are shown in gray. The peptides are shown in stick representation. PHOX2B peptide is colored pink, the models from corresponding templates are highlighted in cyan (2BCK), salmon (5N6B), yellow (2GIT), purple (5MEQ) and green (1IM3). The peptide sequences of the templates are provided. The binding energies (Rosetta Energy Units, REU) and backbone heavy-atom (N, CA, C, O) are shown below the structure diagrams.

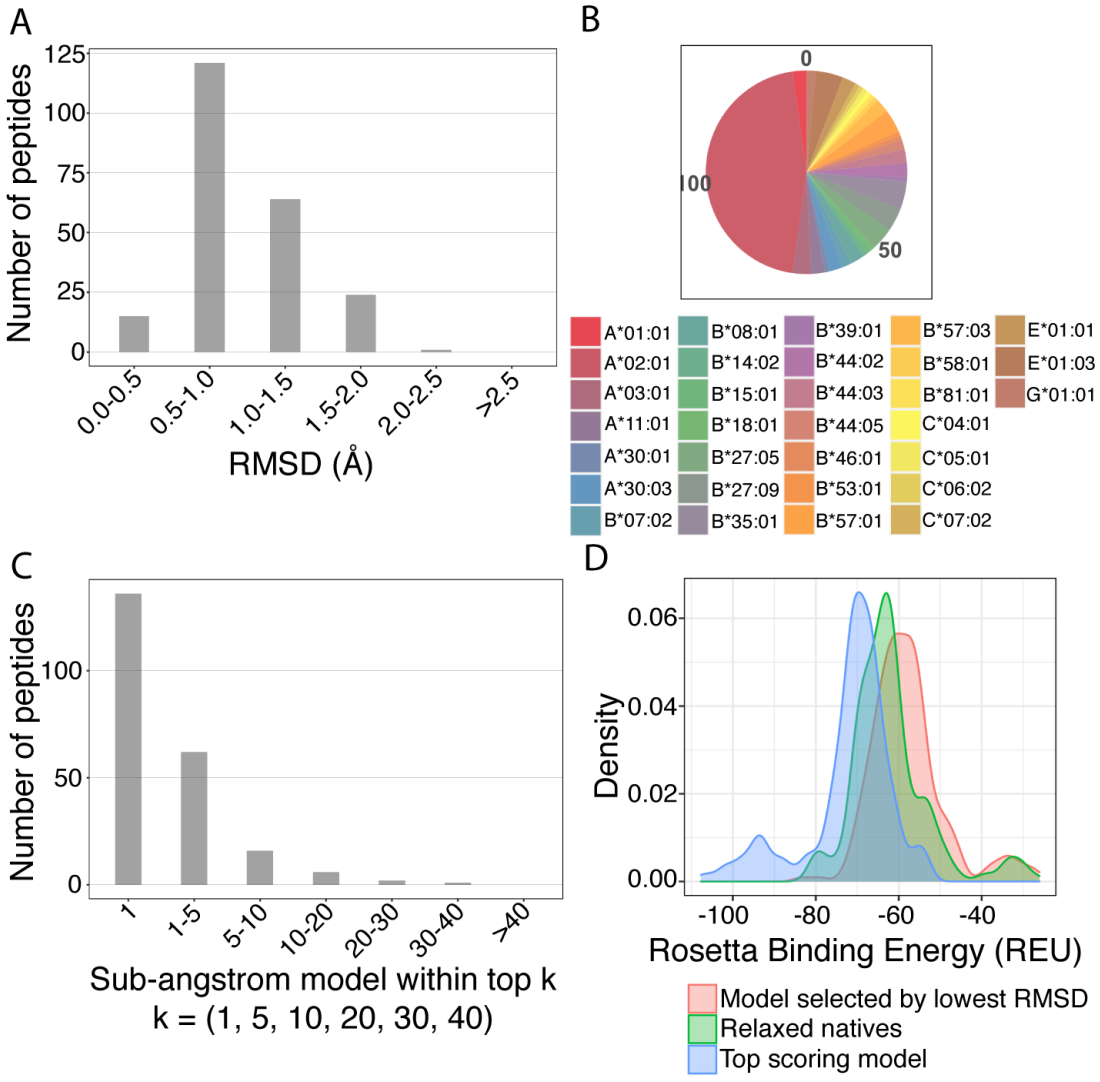


Figure 4.20: Benchmark results of the groove-based multi-template selection criteria. (A) RMSD (Å) distribution of peptides backbones derived from pMHC structures in the PDB and the same peptide/MHC sequences (also referred to as benchmark set) modeled from templates selected using the groove-based template selection. RMSD values are calculated for the heavy atoms, N, C, CA, O, of the peptide backbone. (B) Pie plot showing the alleles for which RosettaMHC successfully modeled top scoring sub-angstrom peptide backbones. (C) Distribution plot showing the number of peptides in the benchmark set that have sub-angstrom backbones and within top k when scored by binding energy. In an ideal case, the top scoring model is sub-angstrom. (D) Rosetta binding energy (REU) distributions of native X-ray structures relaxed in the Rosetta force field (green), top scoring models of the benchmark set (blue) and models selected by lowest RMSD values (crimson). Figure adopted from a manuscript in preparation.

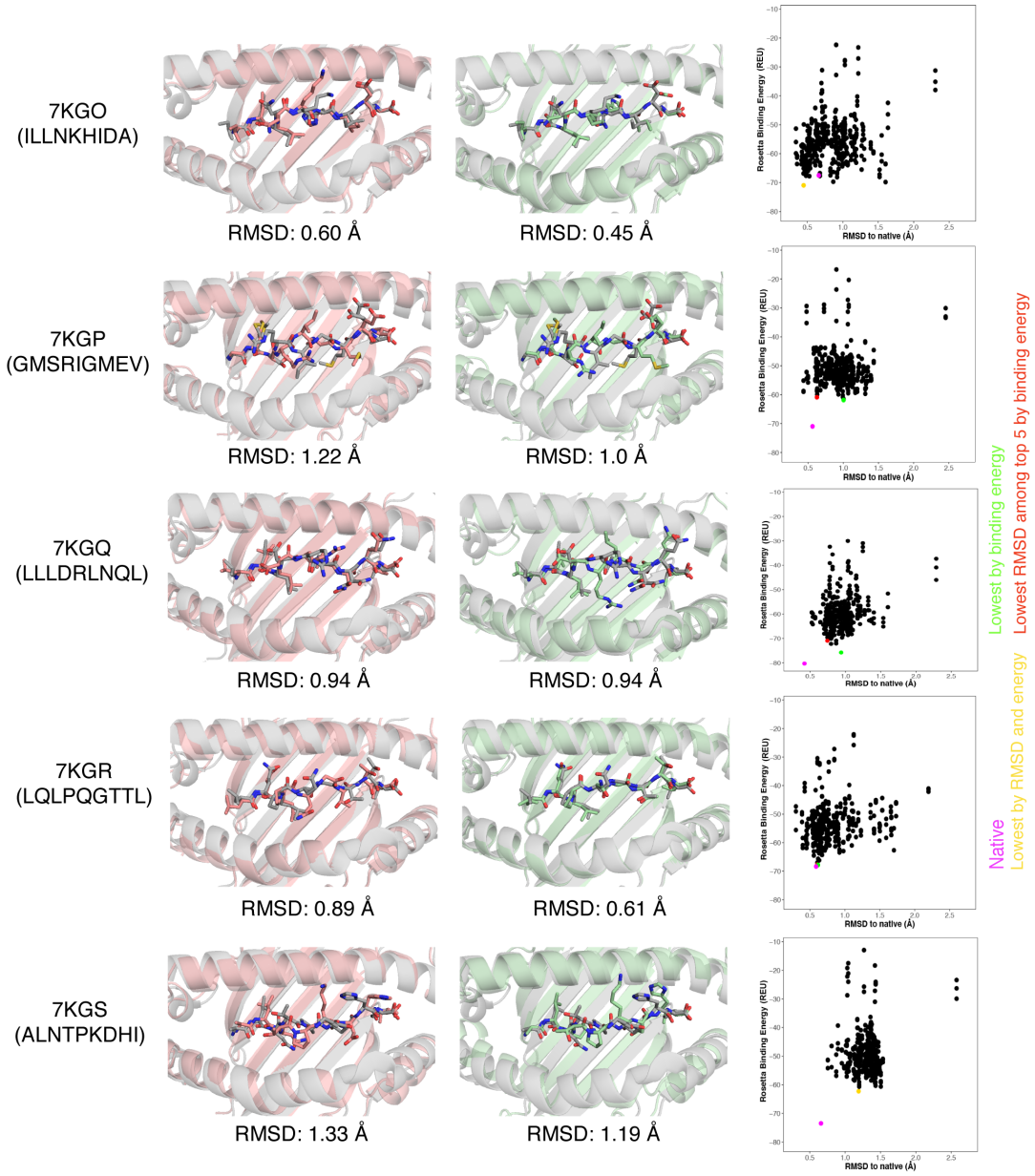

7KGO
(ILLNKHIDA)
RMSD: 0.60 Å
RMSD: 0.45 Å
7KGP
(GMSRIGMEV)
RMSD: 1.22 Å
RMSD: 1.0 Å
7KGQ
(LLLDRLNQL)
RMSD: 0.94 Å
RMSD: 0.94 Å
7KGR
(LQLPQGTTL)
RMSD: 0.89 Å
RMSD: 0.61 Å
7KGS
(ALNTPKDHI)
RMSD: 1.33 Å
RMSD: 1.19 Å
Rosetta Binding Energy (REU)
RMSD to native (Å)
Native
Lowest by RMSD and energy
Lowest by binding energy
Lowest RMSD among top 5 by binding energy

Figure 4.21

Figure 4.21 *(previous page)*: Multi-template modeling selects optimal peptide backbone conformations relative to single template modeling. (Left) Overlay of X-ray structures (gray) of five SARS-CoV-2 peptide/HLA-A*02:01 complexes on models (crimson) generated using RosettaMHC and templates selected using peptide sequence similarity. (Middle) Overlay of X-ray structures (gray) of five SARS-CoV-2 peptide/HLA-A*02:01 complexes on models (pale green) generated using RosettaMHC multi-template modeling approach. The RMSD values between X-ray structures and the models are reported below each panel. The PDB IDs of the five X-ray structures together with peptides displayed are indicated on the leftmost column next to the structure diagrams. (Right) Rosetta binding energy (REU) vs RMSD (Å) plot generated from the groove-based multi-template modeling results. Each point represents a model generated using different template. In these plots, native or the X-ray structure is highlighted by magenta, lowest energy model using green, lowest RMSD model among top five models scored by energy using red and lowest RMSD and energy model using yellow points. Figure adopted from a manuscript in preparation.

4.5 Limitations of MHC groove-based multi-template homology modeling

We demonstrated that we can accurately model sub-angstrom structures of approx. 63% and approx. 90% of peptides among top-scoring and top-5 scoring models in the PDB benchmark and SARS-CoV-2 blind data sets. Despite such high-performance, it is difficult for expert and non-expert users to select a correct structure among top-5 scoring models. Generally, when we examine Energy vs RMSD values, we expect models having lower energy to also be closer to the native, whereas in our case, we lack the signal between energy and RMSD values for a subset of the peptides which is otherwise useful for better discrimination. Therefore, we built a workflow that utilizes dihedral angles to discriminate optimal models (Figure 4.22A). To evaluate if dihedral angles are a good indicator of the peptide backbone, we computed pairwise distance scores between dihedral angles as given by the equation 4.1 [82]. Here, we found that lower the distance scores between peptide backbones, closer they are to the native structures (measured in terms of RMSD values) (Figure 4.22B).

$$Distance\ score = \sum_{i=1}^{k} 2(1 - cos(\theta^{i}_{template} - \theta^{i}_{target})) \tag{4.1}$$

In the Eq. 4.1, k $\in$ {2, 3, 4, 5, 6, 7, 8} and $\theta \in \{\phi, \psi\}$.

We hypothesized that dihedral angles can identify optimal templates for homology modeling,

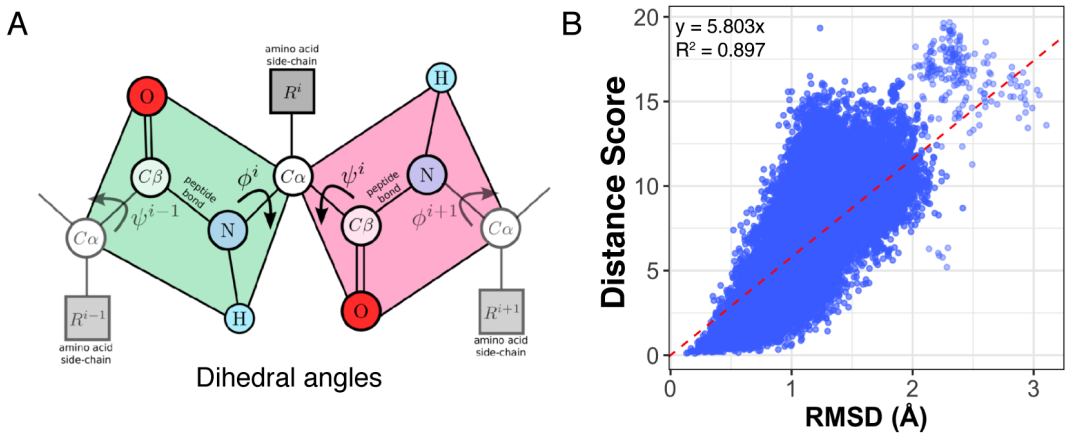


Figure 4.22: Dihedral angles can discriminate near-native templates. (A) Diagram showing the dihedral angles (ϕ vs ψ) for a section of a protein backbone. This panel is adopted from [27]. (B) Scatter plot showing a correlation between backbone dihedral angles and RMSD values across all the peptide pairs displayed by the MHC molecules in the PDB. Figure adopted from a manuscript in preparation.

therefore, we built an artificial neural network to predict dihedral angles given the sequence of a peptide. Moreover neural networks have been used to predict dihedral orientation restraints from sequence information to successfully model complicated CDR3 loops of antibodies and more generally in identifying errors in loop regions in a given structure [102, 37]. Therefore, we use artificial neural network in conjunction with groove-based approach to help us eliminate templates that are potentially inaccurate. We discuss the architecture and performance of ANNs for different tasks in detail in the next sections.

4.6 Artificial Neural Network to predict dihedral angles

A peptide sequence can adopt a backbone conformation within the MHC groove by sampling from a finite set of backbone dihedral angles that fall within the regions described by the Ramachandran plots for the amino acids (Figure 4.23A). For the purpose of demonstration, we derived ϕ and ψ dihedral angles for 150 peptides of length 9 across positions 2 to 8 and found that positions, 2, 3, 7 and 8 are conserved relative to positions 4, 5, and 6 (Figure 4.23A), which is in agreement with the previous findings by the other groups [79].

The first feed forward neural network is trained to predict dihedral angles of peptide sequences. We use as input, sequences of peptides displayed by HLA-A, -B, and -C alleles in

the PDB. Next, we derive backbone dihedral angles from the corresponding pMHC structures in PDB (Figure 4.23) for training. In particular, we are interested in the dihedral angles (ϕ and ψ) of residues 2 to 8 of the peptide that result in a total of 14 ϕ and ψ angles (see the distribution of angles in peptides derived from the pMHC structures in Figure 4.23A).

Input features

We used one-hot encoding of the sequences provided as input to the network. Therefore, we have a total of $L \times 20$, where L is the length of the amino acid sequence of the peptide. For instance, for a peptide of length 9, we have $9 \times 20 = 180$ input features for encoding the peptide sequence.

Training dataset

We utilized 214 manually curated pMHC structures in the PDB (deposited prior to 2019). These structures display peptides of length 9 and are from super types HLA-A, -B, -C, -E and -G and have resolution of less than 3.5 Å. All the peptides are bound to a total of 43 alleles shown in Appendix (Table B) We idealized all the pMHC structures in the Rosetta force field to adjust dihedral angles that were identified as Ramachandran outliers.

Validation dataset

We performed leave-one-out cross-validation of 214 peptides. In particular, we applied 213 pMHC structures in the PDB to train the ANN and predict dihedral angles for one peptide and repeated these calculations for all the peptides in the set.

Network architecture

Our neural network has an input layer that accepts one-hot encoded peptide sequences. The $L \times 20$ input features are passed to a hidden layer with 20 nodes. We tested the ANN architecture with several nodes ranging from 4 to 64 and applied 20 since we obtained lower mean absolute error values for our validation test set. Finally, we have an output layer which outputs 14 dihedral angles. We applied distance score function as a loss function (Eq. 4.2; [82]) (lower

the distance, better the prediction) with Adam optimizer and L2 regularization penalty of 0.001 with 100 epochs (Figure 4.23B).

$$Distance_i = \sum_{i=1}^{k} 2(1 - cos(\theta^i_{template} - \theta^i_{target})) \tag{4.2}$$

In the Eq. 4.2, $Distance_i$ is the distance between same type of angles at residue i in the template and the target structures and $\theta \in \{\phi, \psi\}$.

Results

We performed leave-one-out cross-validation of the ANN to evaluate the dihedral angle predictions of the peptides displayed by pMHC molecules in the PDB (Figure 4.23C). From our analyses, it is evident that the ANN can predict correctly for 11 out of 14 dihedral angles as shown by the average RMSD distance values between predicted and template dihedral angles across the peptides (Figure 4.23C). Here, we have used a lenient 40° threshold to classify if a predicted angle is correct. The ANN performs poorly for ψ_4 and ϕ_7 (Figure 4.23C) and these are the residues with high degree of variations in the peptide backbone as highlighted by Ramachandran plots in Figure 4.23A.

While our predictions are generally accurate for residues 2, 3, 6, and 8, however, these residues are not necessarily the true indicators of variability seen in the peptide backbones. We know that our dataset consists of templates with conserved backbone angles for anchor positions. Therefore, we need to weigh the predictions based on their angles and their positions in the peptide sequence. To address this problem, we constructed another neural network that accepts distances between predicted and template dihedral angles and classifies the templates as optimal or sub-optimal.

4.7 Artificial Neural Network to classify optimal templates

The second feed forward neural network is trained to classify the pMHC templates as optimal or sub-optimal by using distance scores between predicted and template dihedral angles. We use as input, predicted dihedral angles of the peptide by the first ANN and output the PDB IDs of all the templates classified as optimal. The distance scores between all pairs of peptides displayed

by the HLA-A, -B, -C, -E and -G alleles in the PDB are used for training the ANN (Figure 4.24).

Input features

We used distance scores between peptide and PDB templates computed as shown in Eq. 4.2 as input to the network. Therefore, we have a total of 14 values, seven ϕ and seven ψ values for residues 2 to 8 for every target/template peptide pair.

Training dataset

We utilized 214 manually curated pMHC structures in PDB (same set used to train the first ANN). We computed distance scores between dihedral angles for every peptide pair in the PDB with the output label as optimal or sub-optimal if their backbone heavy-atom RMSD values are <1 Å or >1 Å .

Validation dataset

We performed leave-one-out cross-validation of 214 peptides. In particular, we applied 213 pMHC structures in PDB to train the ANN, selected all the templates classified as optimal for the peptide and repeated these calculations for all the peptides in the set. We used receiver operating characteristic curve to estimate the threshold to classify if a template is optimal or sub-optimal. A threshold was selected by choosing a value that filtered optimal templates for most peptides in the benchmark dataset.

Blind test dataset

We used 7 SARS-CoV-2 pMHC X-ray structures [110, 77] as a blind test dataset. The PDB IDs of these structures are 7KGO, 7KGP, 7KGQ, 7KGR, 7KGS, 7M8U, and 7M8T and the structures have a resolution of less than 2.2 Å. A total of five structures have peptides bound to HLA-A*02:01, and two are displayed by HLA-A*11:01 and HLA-B*35:01 alleles respectively. These structures are not included in either training or validation sets.

Network architecture

Our neural network has an input layer that accepts distance scores between predicted and template dihedral angles. The 14 input features are passed to a hidden layer with 14 nodes. Finally, we have an output layer which outputs a score between 0 and 1 for the target/template pair. A threshold is applied to classify a given template are optimal or sub-optimal. We applied distance score function as a loss function (Eq. 4.2; [82]) with Adam optimizer and L2 regularization penalty of 0.001 with 100 epochs (Figure 4.24A).

Results

We performed leave-one-out cross-validation of the ANN for the peptides in the PDB to evaluate (i) the reduction in the number of templates for each peptide, and (ii) the sub-angstrom models retained post combined groove-based and ANN-based template selection (Figure 4.24). For peptides displayed by 43 alleles in the PDB, we observe an average decrease of 70% of the templates relative to the number of templates selected using only the groove-based approach (Figure 4.24B), which boosts the throughput of our method significantly. In addition, we found that 70% of the peptides are predicted to be sub-angstrom and top scoring (Figure 4.24C). Lastly, we can recover sub-angstrom backbones for 70% of peptides displayed by the alleles from all super-types in the PDB (Figure 4.24D).

To further test the accuracy and performance of the ANN and groove-based approach on the unbiased test set, we modeled the structures of seven SARS-CoV-2 peptides displayed by three different alleles [110, 77]. Six out of seven peptides were modeled with sub-angstrom accuracy (Figure 4.25). Upon observation of the two success cases for peptides across two different alleles, we found that the placement of side chains is recovered accurately in the top scoring models (Figure 4.25). Moreover, all the models in the success cases filtered using ANN are found within the 1 Å from their native. In a failure case, we found that the models selected in general are sub-optimal, so we reasoned that ANN cannot filter sub-angstrom models resulting in a peptide backbone of 1.4 Å (Figure 4.25).

We demonstrated that we can accurately predict models of pMHC complexes with sub-angstrom accuracy in a high-throughput manner for over 70% of the benchmark and the blind test sets. We may further improve the performance of the template selection process by ana-

lyzing side chain dynamics of the naturally sampled peptide backbones. The hypothesis of a possible approach is discussed in the Future Work chapter of this thesis.

4.8 Conclusions

Homology modeling of pMHC-I complexes relies on the selection of accurate templates displaying near-native peptide backbones. To select templates, we applied (i) peptide sequence similarity, and (ii) MHC groove sequence identity criteria. We found that using peptide sequence similarity, we can obtain structural models within 1.5 Å, whereas, the accuracy is higher (or RMSD between the models and the natives is < 1 Å) for a majority of the peptides displayed by different alleles in our benchmark dataset if we use MHC groove-based approach. In addition, we employed ANNs to filter out incorrect peptide backbones. Through our analyses using benchmark, validation and blind datasets we demonstrate that we can model sub-angstrom pMHC structures which are useful to study different diseases and devise better therapies.

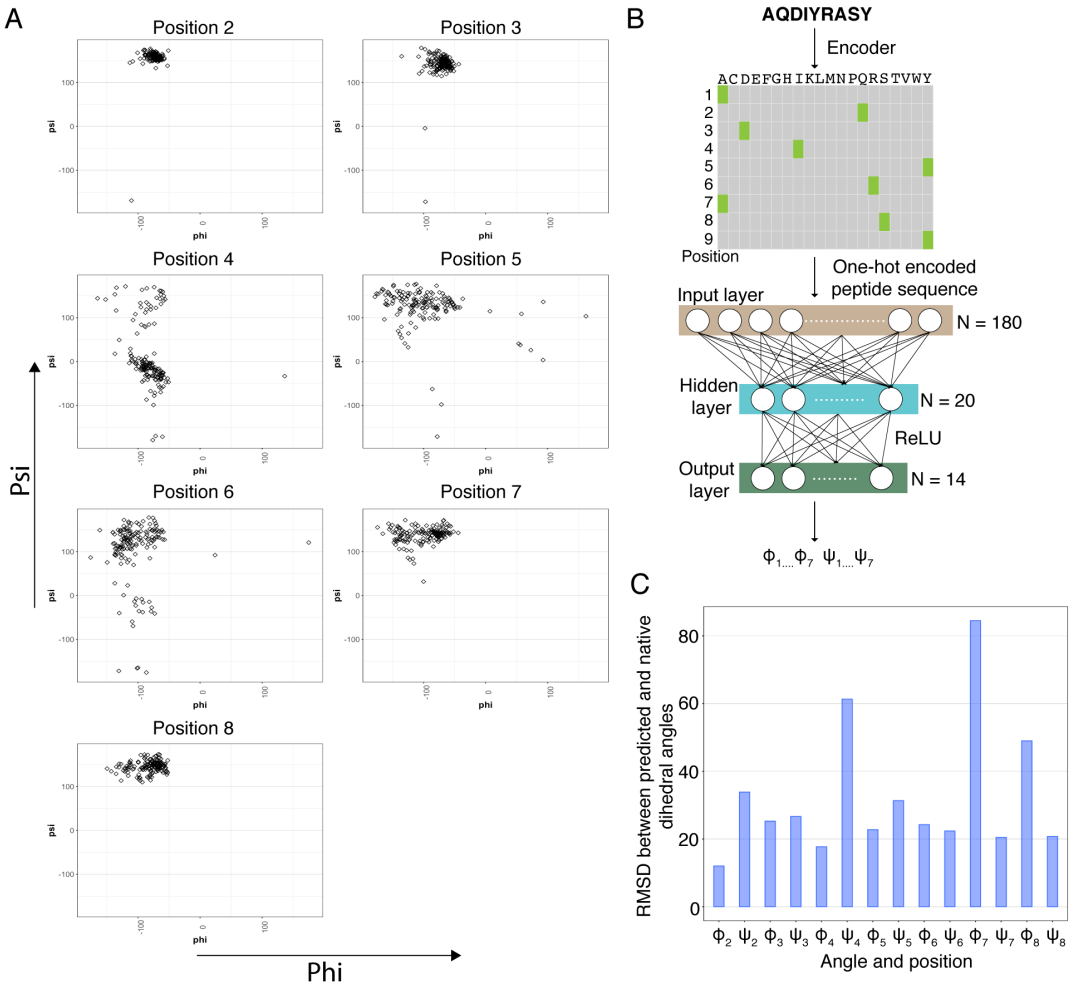


Figure 4.23: (A) Each plot shows the distribution of dihedral angles (ϕ vs ψ) for positions 2 to 8 in approx. 150 peptides derived from pMHC structures in the PDB. These distributions demonstrate that positions 2, 3, 7 and 8 are conserved relative to positions 4, 5 and 6. (B) Feed forward neural network used to predict dihedral angles given the peptide sequences. The peptide sequences together with 14 dihedral (7 ϕ and 7 ψ) angles from residues 2 to 8 are used to train a feed forward neural network. The network has an input layer that accepts one-hot encoded peptide sequence, a hidden layer that has 20 nodes and uses a relu activation function and an output layer that predicts the dihedral angles. (C) Bar plot showing the average RMS difference between predicted and template dihedral angles across all the peptides in leave-one-out cross validation. Figure adopted from a manuscript in preparation.

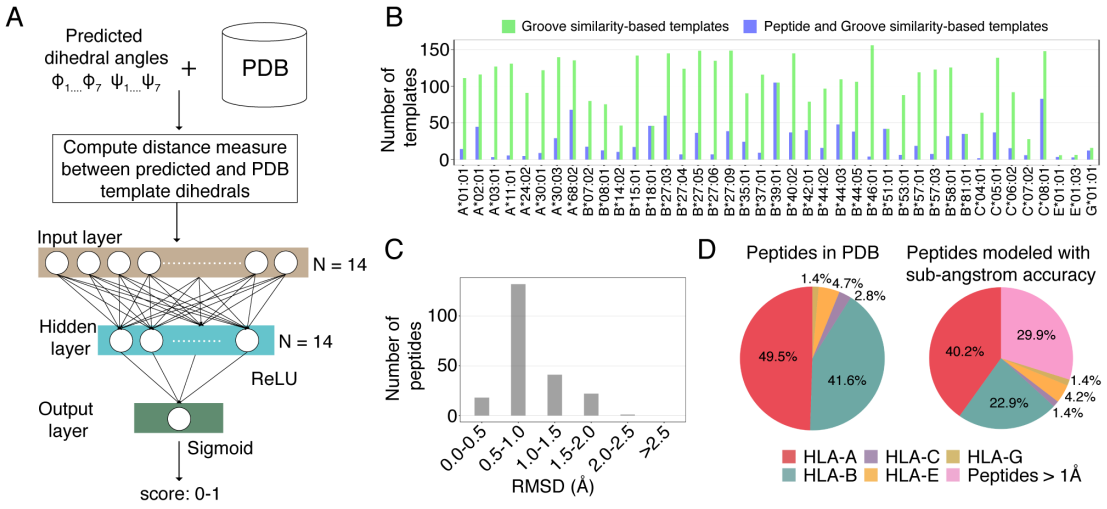


Figure 4.24: (A) Feed forward neural network used to classify pMHC templates as optimal and sub-optimal given the predicted dihedral angles of the peptide of interest. The network has an input layer that accepts distance scores of predicted and template dihedral angles, a hidden layer that has 14 nodes and uses a relu activation function and an output layer with a sigmoid activation function that returns a score between 0 and 1 for each template in the PDB. A threshold score is determined and used to classify templates as optimal. (B) Bar plot showing the average reduction in number of templates selected as optimal using groove-based and ANN approach (blue) vs only the groove-based approach (green) across all the alleles in the PDB. (C) RMSD (Å) distribution of peptides backbones derived from pMHC structures in the PDB and the same peptide/MHC sequences (also referred to as benchmark set) modeled using templates selected using the combined ANN-based optimal template selection and groove-based approach. RMSD values are calculated for heavy atoms, N, C, CA, O, of the peptide backbone. (D) (Left) Pie plot showing the distribution of peptides displayed by different alleles in the PDB. (Right) Pie plot showing the distribution of peptides displayed by different alleles modeled by RosettaMHC with sub-angstrom accuracy. The pink region of the pie plot represents the peptide backbones modeled with >1 Å accuracy. Figure adopted from a manuscript in preparation.

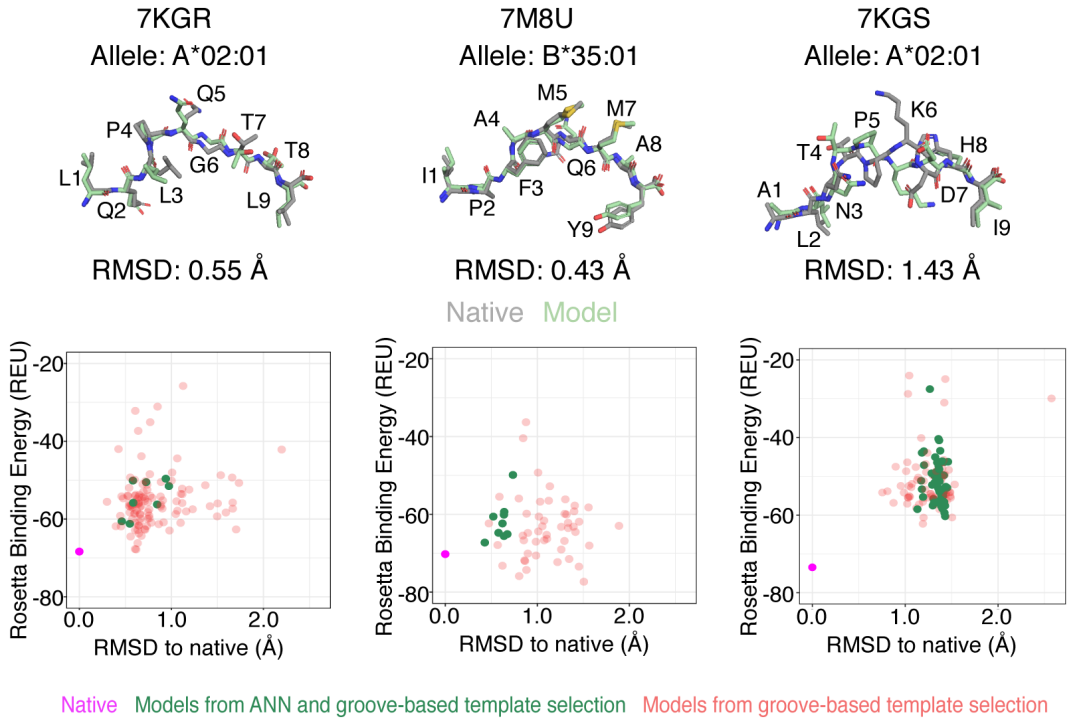


Figure 4.25: Backbone conformations sampled using ANN and groove-based template selection for SARS-CoV-2 peptides bound to HLA-A*02:01 and HLA-B*35:01. (Top) Overlay of SARS-CoV-2 peptide backbone and side chain conformations of native (grey) and best scoring structures (pale green) modeled by RosettaMHC. Alleles displaying each peptide and the native PDB IDs are listed. The RMSD values (in Å) between native and models are reported below. Amino acids of each position of the peptide are indicated on the stick diagrams. (Bottom) Each sub-panel shows Rosetta energy (in Rosetta Energy Units; REU) vs RMSD (in Å) plot for SARS-CoV-2 derived peptides bound to HLA-A*02:01 (PDB IDs: 7KGR and 7KGS), and HLA-B*35:01 (PDB ID: 7M8T). Energies of native models are highlighted using magenta, models from groove-based only template selection are highlighted by red and models from the combined ANN-based and groove-based template selection are highlighted by green. Figure adopted from a manuscript in preparation.

Chapter 5

Conclusions

5.1 Introduction

In this thesis, I present my work on three dimensional modeling of peptide/MHC-I complexes using homology modeling in Rosetta. The 3D structural models of peptide/MHC complexes can provide molecular basis to understand TCR interaction and peptide immunogenicity thereby aid in the development of efficient therapeutics to fight diseases.

5.2 Chapters

The Chapter 1 introduces MHC class I molecules and their role in adaptive immunity, and MHC-I antigen processing and presentation pathway. Further, we discuss (i) the importance of T-cells and MHC restriction, (ii) how peptide/MHC-I structures can aid in understanding TCR recognition, and (iii) the significance of structure modeling methods that eliminate the need to perform tedious experiments to determine pMHC structures. In the end, we highlight the application of building accurate modeling methods and how it can help design peptide-based vaccines for malignant diseases.

In Chapter 2, we review the existing sequence-based and structure-based methods that are used to predict peptide binding to MHC-I molecules and their immunogenicity. The sequence-based methods employ artificial neural networks trained using experimental data obtained using elution assays, binding affinity assays and mass spectrometry. In contrast, we review articles

that demonstrate that structure-based methods do not require any knowledge-base to predict pMHC binding but rather use computationally demanding algorithms limiting their applicability in modeling entire peptide repertoire of the MHC molecules. Lastly, we highlight the latest work, that combines sequence-based and structure-based methods to identify highly specific and accurate pMHC binding profiles.

In Chapter 3, we discuss the basics of nuclear magnetic resonance spectroscopy, my contribution to two methods 4D-CHAINS and MAUS that are used to assign backbone and side chain resonances, and methyl resonances respectively. In addition, we elaborate on structure modeling software suite, Rosetta and CS-Rosetta and how we utilized it together with NMR data to solve structures of important protein targets. Finally, we introduce our homology modeling method built using Rosetta to model structures of pMHC complexes and demonstrate that these models when compared to the solved solution NMR structure of HLA-A*01:01/NRAS Q16K mutation peptide, are very close and that our method has the capacity to recapitulate native peptide backbone.

The chapter 4 describes our method, RosettaMHC, to model peptide/MHC-I molecules using homology modeling. We explain multiple strategies employed to select templates for homology modeling based on peptide sequence, and MHC groove sequence similarities. We apply these strategies to model peptides derived from tumors such as ALK and PHOX2B and SARS-CoV-2 virus bound to different alleles. We evaluate these methods extensively using pMHC crystal structures in PDB and blind targets, SARS-CoV-2 peptide/MHC complexes. Due to the low-resolution of Rosetta energy function, we implemented artificial neural networks to help us filter out models that are potentially incorrect. We believe that, RosettaMHC can accurately model structures of peptide/MHC-I complexes with high accuracy for up to 70% of the cases (top scoring and approx. 90% among top-5 scoring models) across several alleles thus taking a small step towards reducing efforts to carry out structure determination process experimentally using X-ray crystallography or NMR spectroscopy.

In summary, we describe RosettaMHC, to predict three dimensional structures of peptide/MHC-I complexes. We show that the models generated by RosettaMHC can be used to (i) identify public cancer neoepitopes that bind many alleles and help in building peptide-based therapeutics to fight cancer, and (ii) understand surface features of the pMHC complexes that face TCRs and hence identify peptides that are unique vs. cross-reactive in the case of infectious diseases.

Chapter 6

Future Work

[Some of the text and figures in this chapter have been published with the following citations: Viviane S. De Paula, Kevin M. Jude, Santrupti Nerli, Caleb R. Glassman, K. Christopher Garcia, and Nikolaos G. Sgourakis. Interleukin-2 druggability is modulated by global conformational transitions controlled by a helical capping switch. Proceedings of the National Academy of Sciences, 117(13):7183–7192, March 2020.]
[I would like to acknowledge Dimitris Achlioptas for designing the SATsfiability algorithm discussed in this chapter.]

The MHC groove- and ANN-based template selection can help us obtain sub-angstrom models for many cases, however, the method may still sample inaccurate peptide backbones (for approx. 30% of peptides in the benchmark dataset if we select models using energy function). A method that can potentially overcome this limitation, must analyze peptide backbone of a model in detail and suggest if the sampled backbone is accurate or not. We believe that, if we can analyze side chain dynamics of a peptide in a particular backbone conformation, we can sample the accurate peptide backbone for any target pMHC molecule. In the following sections, we discuss a method that can be used to study side chain dynamics in proteins followed by how such methods can potentially help choose an accurate peptide backbone while improving performance.

6.1 Side chain packing using Satisfiability

A sequence of a protein is made up of a combination of 20 amino acids. Each amino acid has a main chain containing amine and carboxyl groups and distinct side chains. The physical properties of amino acids such as charge, size and polarity are determined by their side chains. The unique conformation of an amino acid side chain called rotamer (or rotational isomer) is defined by its *dihedral angles* (χ1, χ2, χ3, and $\chi 4$). The dihedral angles adopt discrete set of favorable values. While the use of discrete values (rather than continuous values) for dihedral angles reduces the space of conformations significantly, it is still large enough to make the search for optimal rotamer combination impractical. For instance, if we consider a 100 residue protein with 3 side chain rotamers at each position, there are in principle $3^{100} =\approx 10^{47}$ ways to generate a full-atom 3D structure (also referred to as side chain rotamer space). Therefore, exhaustive enumeration of all possible conformations is not feasible and hence, many popular methods apply heuristic algorithms to sample from this rotamer space [19].

The main aim of this work is to understand the side chain dynamics and estimate side chain conformational entropy-based packing score (or just packing score) in proteins. To achieve these goals, as a first step, we transform the space of valid side chain conformations to a form that allows us to understand how the structural fold affects side chain placement. In particular, we represent rotamers as boolean variables and *valid combination of rotamer conformations* which encode structural information, as hard constraints. We utilize a general-purpose machine that can take as input these constraints and perform an exhaustive global check to identify and eliminate rotamer combinations that are infeasible for a given backbone to reduce the space of rotamers. Later, we utilize the reduced rotamer space to (i) infer dynamics in Interleukin-2 (IL-2) [87], and (ii) estimate side chain conformational entropy-based packing score which can be used to filter optimal structural models. The use of Satisfiability to reduce the space of rotamers is inspired by another work that utilizes similar principles to decipher NMR data [73].

Method

Let us consider a *n* amino acids protein (Definition 1) for which we want to estimate packing score.

Definition 1 *Let $R = \{r_1, r_2, \cdots, r_n\}$ denote a set of **n** residues of a protein.*

Here, each residue has a set of rotamers (Definition 2) called support set of that residue.

Definition 2 *Let $\Phi^{r_i} = \{s_1^{r_i}, s_2^{r_i}, \cdots, s_m^{r_i}\}$ denote a set of side chains rotamers (or support set) of residue r_i, where $1 \leq i \leq n$.*

We know that the side chain rotamer space for a given protein structure is the Cartesian product of residue support set sizes (Equation 6.1).

$$\Pi_{i=1}^{n} |\Phi^{r_i}|$$

Satisfiability to probe large space of solutions

We start with a rotamer library and a protein structure whose backbone is rigid. The rotamer library contains backbone dependent side chain rotamers for each residue except for residues alanine and glycine. Next, we generate a series of hard constraints that comply with a set of rules given by the definition of valid rotamer combination (see below). The input data and constraints are given to a Satisfiability (SAT) solver. The solver takes all this information and provides a single arbitrary solution from the space of valid solutions or a mathematical proof that it cannot be solved (see Figure 6.1).

Definition of a valid solution

- *A candidate solution consists of one rotamer for each residue.*
- *Clashes with the protein backbone are not allowed.*
- *Clashes between any two pairs of side chains are not allowed.*

To convert the side chain rotamer space to SAT-space we represent each rotamer as a boolean variable that can take values 0 and 1 (see Definition 3). If a rotamer is selected for a specific residue, then the corresponding variable is turned on (or set to 1), otherwise, it is turned off (or set to 0).

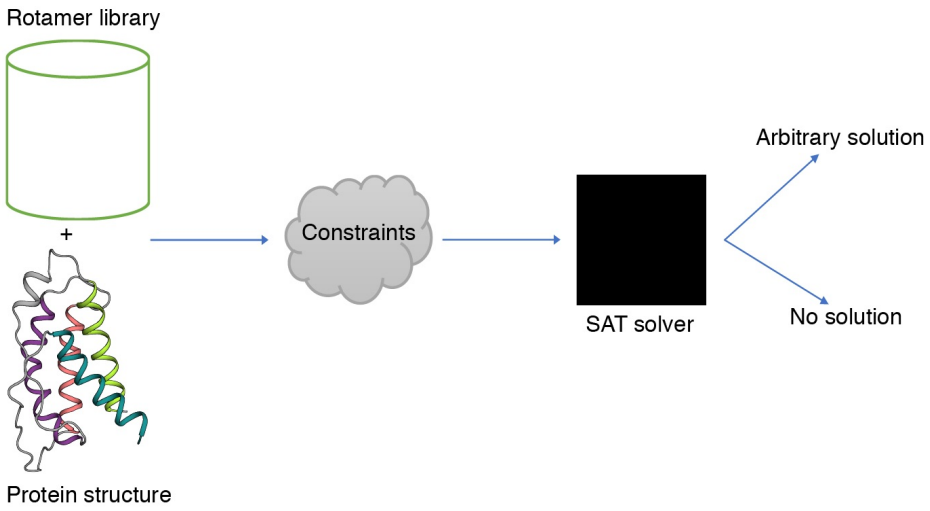


Figure 6.1: Overview of Satisfiability-based side chain packing. Backbone-dependent rotamer library is used to obtain plausible rotamers for a given protein structure backbone. Hard constraints are generated utilizing the definitions of valid rotamer combinations and input to SAT solver that either outputs a valid rotamer combination or a mathematical proof that any valid solution cannot be found.

Definition 3 *Let X_{ij} be a boolean variable representing a side chain rotamer $s_j^{r_i} \in \Phi^{r_i}$ where $r_i \in R$. Here, **i** represents a residue index in the protein sequence and **j** represents a side chain rotamer index in the support set of residue i.*

Encoding of side chain rotamers

We create a propositional formula for the SAT solver which encodes our definition of valid solution. For instance, to encode that the candidate solution can have only one rotamer for each residue, we add the exactly one clause (Appendix, Algorithm 2). To obtain exactly one clause, we need to satisfy both at most one and at least one clauses for one rotamer of each residue. Similarly, to encode clashes between two side chain rotamers, say X_{ij} and Y_{mn}, we add a clause $\{X_{ij} \wedge \neg Y_{mn}\}$ (this clause explicitly states to the SAT solver that if we consider rotamer X_{ij}, then we should not select Y_{mn} and vice versa [52].

Exhaustive enumeration of residue support sets

Due to a number of favorable rotamers for each residue after the removal of infeasible rotamers, we may still have large support set sizes for these residues. In principle, these support sets

may give rise to large number of solutions, which makes it challenging to enumerate them. For example, if we start with a 100 residue protein, where each residue has 4 rotamers, after running this target through SAT-based approach, we may have 3 rotamers for every residue. Therefore, the space of solutions certainly reduced from 4^{100} to 3^{100}, but it is not possible to explore 3^{100} solutions. To overcome this issue, we make use of millisecond time-scale SAT solver iteratively to construct reduced residue support sets without enumerating the entire space of valid solutions.

Here, we pick a residue and its rotamer that is turned on, say X_{ij} and add a temporary clause to the propositional formula, where we provide that X_{ij} can never be true. We then check if the formula returns alternative satisfying assignment upon which we examine the next rotamer of that residue, otherwise, we repeat this process for the next residues. In the process of iterating over these rotamers, we will construct corresponding residue support sets, which now contain only those rotamers that make the formula satisfiable. This procedure can run in time linear to the size of residue support sets making this strategy high-throughput.

Application: Dynamics in IL-2

We applied the SAT-based method to understand dynamics in IL-2 [87]. IL-2 regulates the activities of white blood cells which in turn affect the development of our immune system. The allosteric pathway in IL-2 controls its immune receptors recognition through an open-closed conformational switching (Figures 6.2 and 6.3). Here, we applied two complimentary NMR experiments, Carr–Purcell–Meiboom–Gill (CPMG) [34] and chemical-exchange saturation transfer (CEST) [122] to show that, in solution, IL-2 exists in two interconverting states (referred to as uncapped or open and capped or closed; termed based on the location of residue 52). To enumerate all possible side chain rotamers that can be adopted by each residue, we performed the global analysis of compatible rotamer pairs using SAT-based approach in Rosetta described previously (Figure 6.2) and mapped our results on the uncapped and capped structures (Figure 6.3). Using the backbone conformations of the ”open” and ”closed” states as inputs, our analysis highlights differences in rotamer sets that can be accessed by both states. The change of state from capped to uncapped involves a large displacement of the AB loop (Figure 6.3). In particular, we identified a large set of residues (L28, L39, M42, L48, M53, F58, F132, L133, W136, and F139) spanning the AB loop, adjacent A and D helices, and part of the hydrophobic core.

For these residues, the space of rotamers was significantly different between the closed and open states, indicating a plausible remodeling of packing interactions. Specifically, a 10-residue segment (V129 to F139) forming the hydrophobic face of the amphipathic D helix exhibits expansions and contractions in allowed rotamer sets as IL-2 transitions between the two states (Figure 6.2). This allosteric effect is captured in a heatmap (Figure 6.3), where uncapped and capped states show differences in the support set sizes of the buried residues.

6.2 Side chain conformational entropy-based score can aid in the selection of accurate peptide backbone

We want to utilize SAT-based side chain packing approach to estimate side chain conformational entropy-based score (or packing score) of the modeled pMHC structures (Packing score estimation is provided in Eq. 6.2 [89]). Due to shape complementarity between peptide and MHC, we expect the packing score of the peptide that is closer to the native to be lower (more ordered since the peptide fits perfectly) compared to alternative peptide backbones. We believe that this approach can aid in sampling the accurate template peptide backbone from the database of naturally sampled backbones in PDB (which has >200 peptide backbones).

$$\delta S = -R[\sum_{i=1}^{N} P_{folded}(i) ln P_{folded}(i) - \sum_{i=1}^{N} P_{unfolded}(i) ln P_{unfolded}(i)]$$

Here, δS is the change in packing score of folded state relative to the unfolded state for a given backbone, $P_{folded}(i)$ and $P_{unfolded}(i)$ is the probability of a residue being in rotamer i in the folded and unfolded states respectively, and R is a gas constant [89]. A similar equation is used to estimate side chain conformational entropy as described in [89].

To test the SAT-based approach in the context of pMHC modeling, we estimated the packing score using Eq. 6.2 for the 5 top scoring PHOX2B peptide/HLA-A*24:02 RosettaMHC models and found that the score can be used to filter accurate peptide backbone (Table 6.1). However, extensive benchmarking of the packing score is required to conclude if the packing score can distinguish native-like models, which is an avenue I will be pursuing in the near future.

Table 6.1: Side chain entropy-based packing score for the top 5 scoring PHOX2B peptide/HLA-A*24:02 RosettaMHC models.

Template	Packing score
X-ray	-10.64
2BCK (optimal template)	-9.57
5N6B	-5.36
2GIT	-6.51
5MEQ	-7.47
1IM3	-4.91

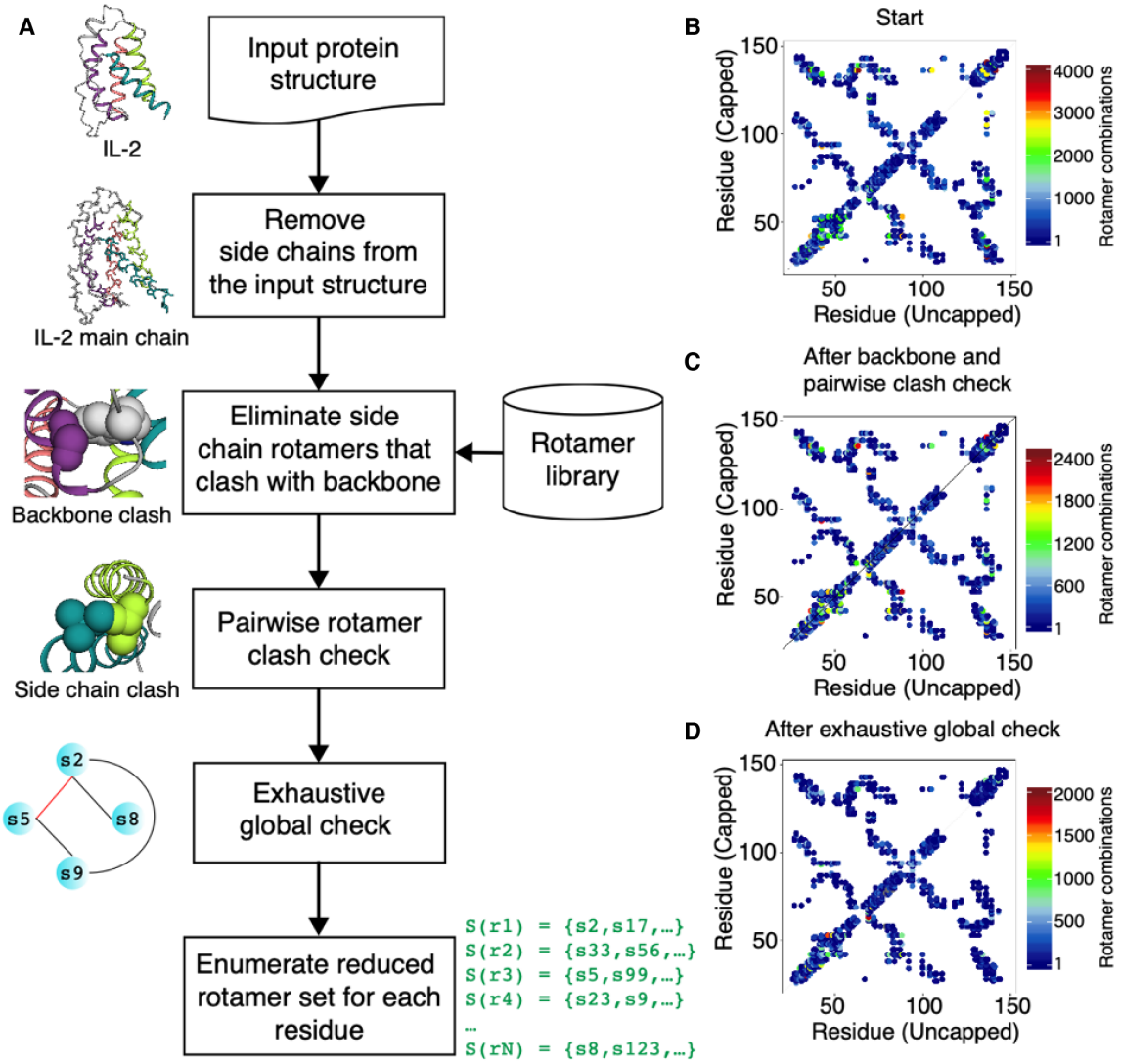


Figure 6.2: Side chain packing analysis in IL-2. (A) Workflow of the method used to perform side chain rotamer space analysis. (B) Number of rotamer combinations between neighboring residues along the sequence of IL-2. An upper bound for number of rotamer combinations after (C) backbone and pairwise clash check, and (D) exhaustive global check (after invoking SAT solver) between neighboring residues. Here, the upper triangular matrix shows rotamer combinations for a capped structure, whereas the lower triangular matrix for an uncapped structure. Exhaustive global check here refers to the execution of SAT solver. Figure was adopted with permission from [87].

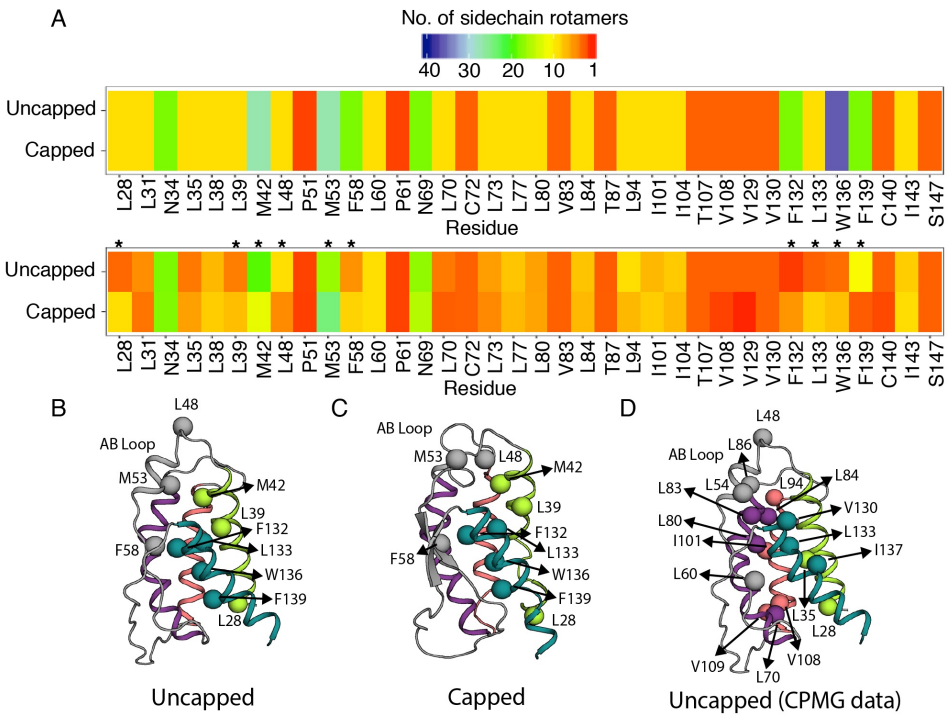


Figure 6.3: SAT-based side-chain packing approach reveals allosteric communication in IL-2. (A) Heatmap showing number of valid side chain rotamers in capped and uncapped structures before (top panel) and after (bottom panel) running through our method. Only the union of buried residues in both capped and uncapped structures is shown in the heatmap. Buried residues were selected using 10 $Å^2$ solvent accessible surface area threshold in PyMol. Residues that exhibit significant difference (difference of 3 or higher) in number of rotamers are highlighted by stars (*). (B,C) Uncapped and capped structures that exhibit significant difference in number of side chain rotamers as indicated in panel A. (D) Uncapped structure showing residues that exhibit chemical shift perturbation as revealed by CPMG experiments. In the structure diagrams, the color coded helices are, A: light green, B: purple, C: salmon and D: dark green. Figure was adopted with permission from [87].

Bibliography

[1] The PyMOL Molecular Graphics System.

[2] The Adaptive Immune System. In *Immunology*, pages 41–67. John Wiley & Sons, Ltd, 2011. Section: 3 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781119998648.ch3.

[3] 14.1: An Introduction to NMR Spectroscopy, June 2014.

[4] AACR Project GENIE Consortium. AACR Project GENIE: Powering Precision Medicine through an International Consortium. *Cancer Discovery*, 7(8):818–831, August 2017.

[5] Rebecca F. Alford, Andrew Leaver-Fay, Jeliazko R. Jeliazkov, Matthew J. O'Meara, Frank P. DiMaio, Hahnbeom Park, Maxim V. Shapovalov, P. Douglas Renfrew, Vikram K. Mulligan, Kalli Kappel, Jason W. Labonte, Michael S. Pacella, Richard Bonneau, Philip Bradley, Roland L. Dunbrack, Rhiju Das, David Baker, Brian Kuhlman, Tanja Kortemme, and Jeffrey J. Gray. The Rosetta all-atom energy function for macromolecular modeling and design. *Journal of chemical theory and computation*, 13(6):3031–3048, June 2017.

[6] Dinler A. Antunes, Jayvee R. Abella, Didier Devaurs, Maurício M. Rigo, and Lydia E. Kavraki. Structure-based Methods for Binding Mode and Binding Affinity Prediction for Peptide-MHC Complexes. *Current Topics in Medicinal Chemistry*, 18(26):2239–2255, 2018.

[7] Michelle P. Aranha, Yead S. M. Jewel, Robert A. Beckman, Louis M. Weiner, Julie C. Mitchell, Jerry M. Parks, and Jeremy C. Smith. Combining Three-Dimensional Modeling with Artificial Intelligence to Increase Specificity and Precision in Peptide–MHC Binding

Predictions. *The Journal of Immunology*, September 2020. Publisher: American Association of Immunologists Section: NOVEL IMMUNOLOGICAL METHODS.

[8] Arash Bahrami, Amir H. Assadi, John L. Markley, and Hamid R. Eghbalnia. Probabilistic Interaction Network of Evidence Algorithm and its Application to Complete Labeling of Peak Lists from Protein NMR Spectroscopy. *PLoS Computational Biology.*, 5(3), 2009.

[9] N. A. Baker, D. Sept, S. Joseph, M. J. Holst, and J. A. McCammon. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proceedings of the National Academy of Sciences of the United States of America*, 98(18):10037–10041, August 2001.

[10] Michal Bassani-Sternberg and David Gfeller. Unsupervised HLA Peptidome Deconvolution Improves Ligand Prediction Accuracy and Predicts Cooperative Effects in Peptide–HLA Interactions. *The Journal of Immunology*, 197(6):2492–2499, September 2016. Publisher: American Association of Immunologists Section: SYSTEMS IMMUNOLOGY.

[11] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, January 2000. Publisher: Oxford Academic.

[12] Oleg Y. Borbulevych, Kurt H. Piepenbrink, Brian E. Gloor, Daniel R. Scott, Ruth F. Sommese, David K. Cole, Andrew K. Sewell, and Brian M. Baker. T cell receptor cross-reactivity directed by antigen-dependent tuning of peptide-MHC molecular flexibility. *Immunity*, 31(6):885–896, December 2009.

[13] Tyler Borrman, Jennifer Cimons, Michael Cosiano, Michael Purcaro, Brian G. Pierce, Brian M. Baker, and Zhiping Weng. ATLAS: a database linking binding affinities with structures for wild-type and mutant TCR-pMHC complexes. *Proteins*, 85(5):908–916, May 2017.

[14] Julian Braun, Lucie Loyal, Marco Frentsch, Daniel Wendisch, Philipp Georg, Florian Kurth, Stefan Hippenstiel, Manuela Dingeldey, Beate Kruse, Florent Fauchere, Emre Baysal, Maike Mangold, Larissa Henze, Roland Lauster, Marcus A. Mall, Kirsten Beyer, Jobst Röhmel, Sebastian Voigt, Jürgen Schmitz, Stefan Miltenyi, Ilja Demuth, Marcel A.

Müller, Andreas Hocke, Martin Witzenrath, Norbert Suttorp, Florian Kern, Ulf Reimer, Holger Wenschuh, Christian Drosten, Victor M. Corman, Claudia Giesecke-Thiel, Leif Erik Sander, and Andreas Thiel. SARS-CoV-2-reactive T cells in healthy donors and patients with COVID-19. *Nature*, July 2020.

[15] Cynthia Xin Lei Chang, Lingyun Dai, Zhen Wei Tan, Joanna Ai Ling Choo, Antonio Bertoletti, and Gijsbert M. Grotenbreg. Sources of diversity in T cell epitope discovery. *Frontiers in Bioscience (Landmark Edition)*, 16:3014–3035, June 2011.

[16] Sidhartha Chaudhury, Sergey Lyskov, and Jeffrey J. Gray. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics*, 26(5):689–691, March 2010.

[17] William Chour, Alex M. Xu, Alphonsus H. C. Ng, Jongchan Choi, Jingyi Xie, Dan Yuan, John K. Lee, Diane C. Delucia, Rick Edmark, Lesley Jones, Thomas M. Schmitt, Mary E. Chaffee, Venkata Duvvuri, Philip D. Greenberg, Kim Murray, Julie Wallick, Heather A. Algren, William R. Berrington, D. Shane O'Mahoney, Jason D. Goldman, and James R. Heath. Shared Antigen-specific CD8+ T cell Responses Against the SARS-COV-2 Spike Protein in HLA A*02:01 COVID-19 Participants. *medRxiv*, page 2020.05.04.20085779, May 2020. Publisher: Cold Spring Harbor Laboratory Press.

[18] Peter J. A. Cock, Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel J. L. de Hoon. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, June 2009. Publisher: Oxford Academic.

[19] José Colbes, Rosario I. Corona, Christian Lezcano, David Rodríguez, and Carlos A. Brizuela. Protein side-chain packing problem: is there still room for improvement? *Briefings in Bioinformatics*, 18(6):1033–1043, November 2017.

[20] Wendy D. Cornell, Piotr Cieplak, Christopher I. Bayly, Ian R. Gould, Kenneth M. Merz, David M. Ferguson, David C. Spellmeyer, Thomas Fox, James W. Caldwell, and Peter A. Kollman. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids,

and Organic Molecules. *Journal of the American Chemical Society*, 117(19):5179–5197, May 1995. Publisher: American Chemical Society.

[21] Rhiju Das, Ingemar André, Yang Shen, Yibing Wu, Alexander Lemak, Sonal Bansal, Cheryl H. Arrowsmith, Thomas Szyperski, and David Baker. Simultaneous prediction of protein folding and docking at high resolution. *Proceedings of the National Academy of Sciences of the United States of America*, 106(45):18978–18983, November 2009.

[22] Frank DiMaio, Andrew Leaver-Fay, Phil Bradley, David Baker, and Ingemar André. Modeling Symmetric Macromolecular Structures in Rosetta3. *PLOS ONE*, 6(6):e20450, June 2011.

[23] Michael A. Dolan, James W. Noah, and Darrell Hurt. Comparison of common homology modeling algorithms: application of user-defined alignments. *Methods in Molecular Biology (Clifton, N.J.)*, 857:399–414, 2012.

[24] Narayanan Eswar, Ben Webb, Marc A. Marti-Renom, M. S. Madhusudhan, David Eramian, Min-Yi Shen, Ursula Pieper, and Andrej Sali. Comparative protein structure modeling using MODELLER. *Current Protocols in Protein Science*, Chapter 2:Unit 2.9, November 2007.

[25] Thomas Evangelidis, Santrupti Nerli, Jiří Nováček, Andrew E. Brereton, P. Andrew Karplus, Rochelle R. Dotas, Vincenzo Venditti, Nikolaos G. Sgourakis, and Konstantinos Tripsianes. Automated NMR resonance assignments and structure determination using a minimal set of 4D spectra. *Nature Communications*, 9(1):384, January 2018.

[26] Jason D. Fernandes, Angie S. Hinrichs, Hiram Clawson, Jairo Navarro Gonzalez, Brian T. Lee, Luis R. Nassar, Brian J. Raney, Kate R. Rosenbloom, Santrupti Nerli, Arjun A. Rao, Daniel Schmelter, Alastair Fyfe, Nathan Maulding, Ann S. Zweig, Todd M. Lowe, Manuel Ares, Russ Corbet-Detig, W. James Kent, David Haussler, and Maximilian Haeussler. The UCSC SARS-CoV-2 Genome Browser. *Nature Genetics*, pages 1–8, September 2020. Publisher: Nature Publishing Group.

[27] Michael Golden, Eduardo García-Portugués, Michael Sørensen, Kanti V. Mardia, Thomas Hamelryck, and Jotun Hein. A Generative Angular Model of Protein Structure Evolution. *Molecular Biology and Evolution*, 34(8):2085–2100, August 2017.

[28] Ananda W. Goldrath and Michael J. Bevan. Selecting and maintaining a diverse T-cell repertoire. *Nature*, 402(6763):6–13, December 1999. Number: 6763 Publisher: Nature Publishing Group.

[29] Faviel F. Gonzalez-Galarza, Antony McCabe, Eduardo J. Melo dos Santos, James Jones, Louise Takeshita, Nestor D. Ortega-Rivera, Glenda M. Del Cid-Pavon, Kerry Ramsbottom, Gurpreet Ghattaoraya, Ana Alfirevic, Derek Middleton, and Andrew R. Jones. Allele frequency net database (AFND) 2020 update: gold-standard data classification, open access genotype data and new query tools. *Nucleic Acids Research*, 48(D1):D783–D788, January 2020. Publisher: Oxford Academic.

[30] Alba Grifoni, John Sidney, Yun Zhang, Richard H. Scheuermann, Bjoern Peters, and Alessandro Sette. A Sequence Homology and Bioinformatic Approach Can Predict Candidate Targets for Immune Responses to SARS-CoV-2. *Cell Host & Microbe*, March 2020.

[31] Alba Grifoni, Daniela Weiskopf, Sydney I. Ramirez, Jose Mateus, Jennifer M. Dan, Carolyn Rydyznski Moderbacher, Stephen A. Rawlings, Aaron Sutherland, Lakshmanane Premkumar, Ramesh S. Jadi, Daniel Marrama, Aravinda M. de Silva, April Frazier, Aaron F. Carlin, Jason A. Greenbaum, Bjoern Peters, Florian Krammer, Davey M. Smith, Shane Crotty, and Alessandro Sette. Targets of T Cell Responses to SARS-CoV-2 Coronavirus in Humans with COVID-19 Disease and Unexposed Individuals. *Cell*, 181(7):1489–1501.e15, 2020.

[32] P Guerry, VD Duong, and T Herrmann. CASD-NMR 2: robust and accurate unsupervised analysis of raw NOESY spectra and protein structure determination with UNIO. *Journal of Biomolecular NMR*, 62(4):473–80, 2015.

[33] Peter Güntert. Automated structure determination from NMR spectra. *European Biophysics Journal*, 38(2):129–43, 2009.

[34] D. Flemming Hansen, Pramodh Vallurupalli, and Lewis E. Kay. Using relaxation dispersion NMR spectroscopy to determine structures of excited, invisible protein states. *Journal of biomolecular NMR*, 41(3):113–120, July 2008.

[35] S Henikoff and J G Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, 89(22):10915–10919, November 1992.

[36] Torsten Herrmann, Peter Güntert, and Kurt Wüthrich. *Journal of Molecular Biology*, 319(1):209–227, 2002.

[37] Naozumi Hiranuma, Hahnbeom Park, Minkyung Baek, Ivan Anishchenko, Justas Dauparas, and David Baker. Improved protein structure refinement guided by deep learning based accuracy estimation. *Nature Communications*, 12(1):1340, February 2021. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Machine learning;Protein structure predictions;Software Subject_term_id: machine-learning;protein-structure-predictions;software.

[38] Edward E. Hodgkin and W. Graham Richards. Molecular similarity based on electrostatic potential and electric field. *International Journal of Quantum Chemistry*, 32(S14):105–110, 1987. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/qua.560320814.

[39] Joseph P. Hornak. *Basics of NMR*.

[40] Yuanpeng Janet Huang, G. V. T. Swapna, P. K. Rajan, Haiping Ke, Bing Xia, Kamal Shukla, Masayori Inouye, and Gaetano T. Montelione. Solution NMR Structure of Ribosome-binding Factor A (RbfA), A Cold-shock Adaptation Protein from Escherichia coli. *Journal of Molecular Biology*, 327(2):521–536, March 2003.

[41] M Ikura, LE Kay, and A Bax. A novel approach for sequential assignment of proton, carbon-13, and nitrogen-15 spectra of larger proteins: heteronuclear triple-resonance three-dimensional NMR spectroscopy. Application to calmodulin. *Biochemistry*, 29(19):4659–67, 1990.

[42] Hsueh-Ling Janice Oh, Samuel Ken-En Gan, Antonio Bertoletti, and Yee-Joo Tan. Understanding the T cell immune response in SARS coronavirus infection. *Emerging Microbes & Infections*, 1(1):1–6, January 2019.

[43] Charles A. Janeway Jr, Paul Travers, Mark Walport, Mark J. Shlomchik, Charles A. Janeway Jr, Paul Travers, Mark Walport, and Mark J. Shlomchik. *Immunobiology*. Garland Science, 5th edition, 2001.

[44] Elizabeth Jurrus, Dave Engel, Keith Star, Kyle Monson, Juan Brandi, Lisa E. Felberg, David H. Brookes, Leighton Wilson, Jiahui Chen, Karina Liles, Minju Chun, Peter Li, David W. Gohara, Todd Dolinsky, Robert Konecny, David R. Koes, Jens Erik Nielsen, Teresa Head□Gordon, Weihua Geng, Robert Krasny, Guo-Wei Wei, Michael J. Holst, J. Andrew McCammon, and Nathan A. Baker. Improvements to the APBS biomolecular solvation software suite. *Protein Science*, 27(1):112–128, 2018. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pro.3280.

[45] Vanessa Jurtz, Sinu Paul, Massimo Andreatta, Paolo Marcatili, Bjoern Peters, and Morten Nielsen. NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *Journal of Immunology (Baltimore, Md.: 1950)*, 199(9):3360–3368, 2017.

[46] LE Kay, GM Clore, A Bax, and AM Gronenborn. Four-dimensional heteronuclear triple-resonance NMR spectroscopy of interleukin-1 beta in solution. *Science*, 249(4967):411–414, 1990.

[47] Lewis E. Kay, Mitsuhiko Ikura, Rolf Tschudin, and Ad Bax. Three-dimensional triple-resonance NMR spectroscopy of isotopically enriched proteins. *Journal of Magnetic Resonance*, 89(3):496–514, 1990.

[48] K Kazimierczuk and V Orekhov. Non-uniform sampling: post-Fourier era of NMR data collection and processing. *Magnetic Resonance Chemistry*, 53(11):921–6, 2015.

[49] James Keeler. *Understanding NMR Spectroscopy, 2nd Edition*. Wiley, 2nd edition, May 2010.

[50] W. James Kent, Charles W. Sugnet, Terrence S. Furey, Krishna M. Roskin, Tom H. Pringle, Alan M. Zahler, and and David Haussler. The Human Genome Browser at UCSC. *Genome Research*, 12(6):996–1006, June 2002. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring

Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

[51] Javed Mohammed Khan and Shoba Ranganathan. pDOCK: a new technique for rapid and accurate docking of peptide ligands to Major Histocompatibility Complexes. *Immunome Research*, 6(Suppl 1):S2, September 2010.

[52] Will Klieber and Gihwon Kwon. Efficient CNF Encoding for Selecting 1 from N Objects. page 14.

[53] Jerzy K. Kulski, Takashi Shiina, Tatsuya Anzai, Sakae Kohara, and Hidetoshi Inoko. Comparative genomic analysis of the MHC: the evolution of class I duplication blocks, diversity and complexity from shark to man. *Immunological Reviews*, 190(1):95–122, 2002. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1034/j.1600-065X.2002.19008.x.

[54] Hyun-Ho Kyeong, Yoonjoo Choi, and Hak-Sung Kim. GradDock: rapid simulation and tailored ranking functions for peptide-MHC Class I docking. *Bioinformatics*, 34(3):469–476, February 2018.

[55] OF Lange and David Baker. Resolution-adapted recombination of structural features significantly improves sampling in restraint-guided structure calculation. *Proteins*, 80(3):884–895, 2012.

[56] Oliver F. Lange. Automatic NOESY assignment in CS-RASREC-Rosetta. *Journal of Biomolecular NMR*, 59(3):147–59, 2014.

[57] Oliver F. Lange, Paolo Rossi, Nikolaos G. Sgourakis, Yifan Song, Hsiau-Wei Lee, James M. Aramini, Asli Ertekin, Rong Xiao, Thomas B. Acton, Gaetano T. Montelione, and David Baker. *Proceedings of the National Academy of Sciences of the United States of America*, 109(27):10873–10878, 2012.

[58] Andrew Leaver-Fay, Michael Tyka, Steven M. Lewis, Oliver F. Lange, James Thompson, Ron Jacak, Kristian W. Kaufman, P. Douglas Renfrew, Colin A. Smith, Will Sheffler, Ian W. Davis, Seth Cooper, Adrien Treuille, Daniel J. Mandell, Florian Richter, Yih-En Andrew Ban, Sarel J. Fleishman, Jacob E. Corn, David E. Kim, Sergey Lyskov, Monica Berrondo, Stuart Mentzer, Zoran Popović, James J. Havranek, John Karanicolas, Rhiju

Das, Jens Meiler, Tanja Kortemme, Jeffrey J. Gray, Brian Kuhlman, David Baker, and Philip Bradley. Chapter nineteen - Rosetta3: An Object-Oriented Software Suite for the Simulation and Design of Macromolecules. In Michael L. Johnson and Ludwig Brand, editors, *Methods in Enzymology*, volume 487 of *Computer Methods, Part C*, pages 545–574. Academic Press, January 2011.

[59] Julia Koehler Leman, Brian D. Weitzner, Steven M. Lewis, Jared Adolf-Bryfogle, Nawsad Alam, Rebecca F. Alford, Melanie Aprahamian, David Baker, Kyle A. Barlow, Patrick Barth, Benjamin Basanta, Brian J. Bender, Kristin Blacklock, Jaume Bonet, Scott E. Boyken, Phil Bradley, Chris Bystroff, Patrick Conway, Seth Cooper, Bruno E. Correia, Brian Coventry, Rhiju Das, René M. De Jong, Frank DiMaio, Lorna Dsilva, Roland Dunbrack, Alexander S. Ford, Brandon Frenz, Darwin Y. Fu, Caleb Geniesse, Lukasz Goldschmidt, Ragul Gowthaman, Jeffrey J. Gray, Dominik Gront, Sharon Guffy, Scott Horowitz, Po-Ssu Huang, Thomas Huber, Tim M. Jacobs, Jeliazko R. Jeliazkov, David K. Johnson, Kalli Kappel, John Karanicolas, Hamed Khakzad, Karen R. Khar, Sagar D. Khare, Firas Khatib, Alisa Khramushin, Indigo C. King, Robert Kleffner, Brian Koepnick, Tanja Kortemme, Georg Kuenze, Brian Kuhlman, Daisuke Kuroda, Jason W. Labonte, Jason K. Lai, Gideon Lapidoth, Andrew Leaver-Fay, Steffen Lindert, Thomas Linsky, Nir London, Joseph H. Lubin, Sergey Lyskov, Jack Maguire, Lars Malmström, Enrique Marcos, Orly Marcu, Nicholas A. Marze, Jens Meiler, Rocco Moretti, Vikram Khipple Mulligan, Santrupti Nerli, Christoffer Norn, Shane Ó'Conchúir, Noah Ollikainen, Sergey Ovchinnikov, Michael S. Pacella, Xingjie Pan, Hahnbeom Park, Ryan E. Pavlovicz, Manasi Pethe, Brian G. Pierce, Kala Bharath Pilla, Barak Raveh, P. Douglas Renfrew, Shourya S. Roy Burman, Aliza Rubenstein, Marion F. Sauer, Andreas Scheck, William Schief, Ora Schueler-Furman, Yuval Sedan, Alexander M. Sevy, Nikolaos G. Sgourakis, Lei Shi, Justin B. Siegel, Daniel-Adriano Silva, Shannon Smith, Yifan Song, Amelie Stein, Maria Szegedy, Frank D. Teets, Summer B. Thyme, Ray Yu-Ruei Wang, Andrew Watkins, Lior Zimmerman, and Richard Bonneau. Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nature Methods*, 17(7):665–680, July 2020. Number: 7 Publisher: Nature Publishing Group.

[60] Jens P. Linge, Michael Habeck, Wolfgang Reiping, and Michael Nilges. ARIA: automated

NOE assignment and NMR structure calculation. *Bioinformatics*, 19(2):315–316, 2003.

[61] Nir London, Barak Raveh, Eyal Cohen, Guy Fathi, and Ora Schueler-Furman. Rosetta FlexPepDock web server–high resolution modeling of peptide-protein interactions. *Nucleic Acids Research*, 39(Web Server issue):W249–253, July 2011.

[62] Giuseppe Maccari, James Robinson, Keith Ballingall, Lisbeth A. Guethlein, Unni Grimholt, Jim Kaufman, Chak-Sum Ho, Natasja G. de Groot, Paul Flicek, Ronald E. Bontrop, John A. Hammond, and Steven G. E. Marsh. IPD-MHC 2.0: an improved inter-species database for the study of the major histocompatibility complex. *Nucleic Acids Research*, 45(D1):D860–D864, January 2017.

[63] Martin Maiers, Loren Gragert, and William Klitz. High-resolution HLA alleles and haplotypes in the United States population. *Human Immunology*, 68(9):779–788, September 2007.

[64] Dermot H. Mallon, Christiane Kling, Matthew Robb, Eva Ellinghaus, J. Andrew Bradley, Craig J. Taylor, Dieter Kabelitz, and Vasilis Kosmoliaptsis. Predicting Humoral Alloimmunity from Differences in Donor and Recipient HLA Surface Electrostatic Potential. *The Journal of Immunology*, 201(12):3780–3792, December 2018. Publisher: American Association of Immunologists Section: NOVEL IMMUNOLOGICAL METHODS.

[65] Ryan J. Malonis, Jonathan R. Lai, and Olivia Vergnolle. Peptide-Based Vaccines: Current Progress and Future Challenges. *Chemical Reviews*, 120(6):3210–3229, March 2020. Publisher: American Chemical Society.

[66] Aimee H Marceau, Caileen M Brison, Santrupti Nerli, Heather E Arsenault, Andrew C McShan, Eefei Chen, Hsiau-Wei Lee, Jennifer A Benanti, Nikolaos G Sgourakis, and Seth M Rubin. An order-to-disorder structural switch activates the FoxM1 transcription factor. *eLife*, 8:e46131, May 2019. Publisher: eLife Sciences Publications, Ltd.

[67] Enrique Marcos, Tamuka M. Chidyausiku, Andrew C. McShan, Thomas Evangelidis, Santrupti Nerli, Lauren Carter, Lucas G. Nivón, Audrey Davis, Gustav Oberdorfer, Konstantinos Tripsianes, Nikolaos G. Sgourakis, and David Baker. De novo design of a non-local β-sheet protein with high stability and accuracy. *Nature Structural & Molecular*

Biology, 25(11):1028–1034, November 2018. Number: 11 Publisher: Nature Publishing Group.

[68] Guillaume Mas, Jia-Ying Guan, Elodie Crublet, Elisa Colas Debled, Christine Moriscot, Pierre Gans, Guy Schoehn, Pavel Macek, Paul Schanda, and Jerome Boisbouvier. Structural investigation of a chaperonin in action reveals how nucleotide binding regulates the functional cycle. *Science Advances*, 4(9):eaau4196, September 2018.

[69] Mary K. McCarthy and Jason B. Weinberg. The immunoproteasome and viral infection: a complex regulator of inflammation. *Frontiers in Microbiology*, 6, January 2015.

[70] Maurício Menegatti Rigo, Dinler Amaral Antunes, Martiela Vaz de Freitas, Marcus Fabiano de Almeida Mendes, Lindolfo Meira, Marialva Sinigaglia, and Gustavo Fioravanti Vieira. DockTope: a Web-based tool for automated pMHC-I modelling. *Scientific Reports*, 5, December 2015.

[71] Michele Mishto and Juliane Liepe. Post-Translational Peptide Splicing and T Cell Responses. *Trends in Immunology*, 38(12):904–915, December 2017.

[72] Annika Nelde, Tatjana Bilich, Jonas S. Heitmann, Yacine Maringer, Helmut R. Salih, Malte Roerden, Maren Lübke, Jens Bauer, Jonas Rieth, Marcel Wacker, Andreas Peter, Sebastian Hörber, Bjoern Traenkle, Philipp D. Kaiser, Ulrich Rothbauer, Matthias Becker, Daniel Junker, Gérard Krause, Monika Strengert, Nicole Schneiderhan-Marra, Markus F. Templin, Thomas O. Joos, Daniel J. Kowalewski, Vlatka Stos-Zweifel, Michael Fehr, Michael Graf, Lena-Christin Gruber, David Rachfalski, Beate Preuß, Ilona Hagelstein, Melanie Märklin, Tamam Bakchoul, Cécile Gouttefangeas, Oliver Kohlbacher, Reinhild Klein, Stefan Stevanović, Hans-Georg Rammensee, and Juliane S. Walz. SARS-CoV-2 T-cell epitopes define heterologous and COVID-19-induced T-cell recognition. *Research Square (preprint)*, June 2020.

[73] Santrupti Nerli, Viviane S. De Paula, Andrew C. McShan, and Nikolaos G. Sgourakis. Backbone-independent NMR resonance assignments of methyl probes in large proteins. *Nature Communications*, 12(1):691, January 2021. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Pub-

lisher: Nature Publishing Group Subject_term: Software;Solution-state NMR;Structural biology Subject_term_id: software;solution-state-nmr;structural-biology.

[74] Santrupti Nerli, Andrew C. McShan, and Nikolaos G. Sgourakis. Chemical shift-based methods in NMR structure determination. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 106-107:1–25, June 2018.

[75] Santrupti Nerli and Nikolaos G. Sgourakis. CS-ROSETTA. *Methods in Enzymology*, 614:321–362, 2019.

[76] Santrupti Nerli and Nikolaos G. Sgourakis. Structure-Based Modeling of SARS-CoV-2 Peptide/HLA-A02 Antigens. *Frontiers in Medical Technology*, 0, 2020. Publisher: Frontiers.

[77] Andrea T. Nguyen, Christopher Szeto, Dhilshan Jayasinghe, Christian A. Lobos, Hanim Halim, Demetra S. M. Chatzileontiadou, Emma J. Grant, and Stephanie Gras. SARS-CoV-2 Spike-Derived Peptides Presented by HLA Molecules. *Biophysica*, 1(2):194–203, June 2021. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

[78] Morten Nielsen, Massimo Andreatta, Bjoern Peters, and Søren Buus. Immunoinformatics: Predicting Peptide–MHC Binding. *Annual Review of Biomedical Data Science*, 3(1):191–215, 2020. _eprint: https://doi.org/10.1146/annurev-biodatasci-021920-100259.

[79] Morten Nielsen, Claus Lundegaard, Thomas Blicher, Kasper Lamberth, Mikkel Harndahl, Sune Justesen, Gustav Røder, Bjoern Peters, Alessandro Sette, Ole Lund, and Søren Buus. NetMHCpan, a Method for Quantitative Predictions of Peptide Binding to Any HLA-A and -B Locus Protein of Known Sequence. *PLoS ONE*, 2(8):e796, August 2007.

[80] Michael Nilges. A calculation strategy for the structure determination of symmetric dimers by 1H NMR. *Proteins: Structure, Function, and Bioinformatics*, 17(3):297–309, November 1993.

[81] Michael Nilges. Ambiguous distance data in the calculation of NMR structures. *Fold and Design*, 2:S53–S57, 1997.

[82] Benjamin North, Andreas Lehmann, and Roland L. Dunbrack. A new clustering of antibody CDR loop conformations. *Journal of molecular biology*, 406(2):228–256, February 2011.

[83] Timothy J. O'Donnell, Alex Rubinsteyn, Maria Bonsack, Angelika B. Riemer, Uri Laserson, and Jeff Hammerbacher. MHCflurry: Open-Source Class I MHC Binding Affinity Prediction. *Cell Systems*, 7(1):129–132.e4, July 2018.

[84] Mats H. M. Olsson, Chresten R. Søndergaard, Michal Rostkowski, and Jan H. Jensen. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. *Journal of Chemical Theory and Computation*, 7(2):525–537, February 2011. Publisher: American Chemical Society.

[85] Sarah A. Overall, Jugmohit S. Toor, Stephanie Hao, Mark Yarmarkovich, Sara M. O'Rourke, Giora I. Morozov, Son Nguyen, Alberto Sada Japp, Nicolas Gonzalez, Danai Moschidi, Michael R. Betts, John M. Maris, Peter Smibert, and Nikolaos G. Sgourakis. High throughput pMHC-I tetramer library production using chaperone-mediated peptide exchange. *Nature Communications*, 11(1):1909, April 2020. Number: 1 Publisher: Nature Publishing Group.

[86] Hahnbeom Park, Gyu Rie Lee, David E. Kim, Ivan Anishchenko, Qian Cong, and David Baker. High-accuracy refinement using Rosetta in CASP13. *Proteins: Structure, Function, and Bioinformatics*, 87(12):1276–1282, 2019. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.25784.

[87] Viviane S. De Paula, Kevin M. Jude, Santrupti Nerli, Caleb R. Glassman, K. Christopher Garcia, and Nikolaos G. Sgourakis. Interleukin-2 druggability is modulated by global conformational transitions controlled by a helical capping switch. *Proceedings of the National Academy of Sciences*, 117(13):7183–7192, March 2020. Publisher: National Academy of Sciences Section: Biological Sciences.

[88] Brett E. Pickett, Eva L. Sadat, Yun Zhang, Jyothi M. Noronha, R. Burke Squires, Victoria Hunt, Mengya Liu, Sanjeev Kumar, Sam Zaremba, Zhiping Gu, Liwei Zhou, Christopher N. Larson, Jonathan Dietrich, Edward B. Klem, and Richard H. Scheuermann. ViPR:

an open bioinformatics database and analysis resource for virology research. *Nucleic Acids Research*, 40(Database issue):D593–D598, January 2012.

[89] Stephen D. Pickett and Michael J. E. Sternberg. Empirical Scale of Side-Chain Conformational Entropy in Protein Folding. *Journal of Molecular Biology*, 231(3):825–839, June 1993.

[90] S. B. Piertney and M. K. Oliver. The evolutionary ecology of the major histocompatibility complex. *Heredity*, 96(1):7–21, January 2006. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 1 Primary_atype: Reviews Publisher: Nature Publishing Group.

[91] Asaf Poran, Dewi Harjanto, Matthew Malloy, Christina M. Arieta, Daniel A. Rothenberg, Divya Lenkala, Marit M. van Buuren, Terri A. Addona, Michael S. Rooney, Lakshmi Srinivasan, and Richard B. Gaynor. Sequence-based prediction of SARS-CoV-2 vaccine targets using a mass spectrometry-based bioinformatics predictor identifies immunogenic T cell epitopes. *Genome Medicine*, 12(1):70, 2020.

[92] Iva Pritišanac, Matteo T. Degiacomi, T. Reid Alderson, Marta G. Carneiro, Eiso Ab, Gregg Siegal, and Andrew J. Baldwin. Automatic Assignment of Methyl-NMR Spectra of Supramolecular Machines Using Graph Theory. *Journal of the American Chemical Society*, 139(28):9523–9533, July 2017.

[93] Srivatsan Raman, Oliver F. Lange, Paolo Rossi, Michael Tyka, Xu Wang, James Aramini, Gaohua Liu, Theresa A. Ramelot, Alexander Eletsky, Thomas Szyperski, Michael A. Kennedy, James Prestegard, Gaetano T. Montelione, and David Baker. NMR Structure Determination for Larger Proteins Using Backbone-Only Data. *Science*, 327(5968):1014–1018, 2010.

[94] H. Rammensee, J. Bachmann, N. P. Emmerich, O. A. Bachor, and S. Stevanović. SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics*, 50(3-4):213–219, November 1999.

[95] Arjun A. Rao, Ada A. Madejska, Jacob Pfeil, Benedict Paten, Sofie R. Salama, and David Haussler. ProTECT—Prediction of T-Cell Epitopes for Cancer Therapy. *Frontiers in Immunology*, 11, November 2020.

[96] Michael Rasmussen, Mikkel Harndahl, Anette Stryhn, Rachid Boucherma, Lise Lotte Nielsen, François A. Lemonnier, Morten Nielsen, and Søren Buus. Uncovering the Peptide-Binding Specificities of HLA-C: A General Strategy To Determine the Specificity of Any MHC Class I Molecule. *The Journal of Immunology*, 193(10):4790–4802, November 2014. Publisher: American Association of Immunologists Section: ANTIGEN RECOGNITION AND RESPONSES.

[97] James Robinson, Jason A. Halliwell, James D. Hayhurst, Paul Flicek, Peter Parham, and Steven G. E. Marsh. The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Research*, 43(Database issue):D423–431, January 2015.

[98] Rina Rosenzweig and Lewis E. Kay. Bringing dynamic molecular machines into focus by methyl-TROSY NMR. *Annual Review of Biochemistry*, 83:291–315, 2014.

[99] Paolo Rossi, Youlin Xia, Nandish Khanra, Gianluigi Veglia, and Charalampos G. Kalodimos. 15N and 13C- SOFAST-HMQC editing enhances 3D-NOESY sensitivity in highly deuterated, selectively [1H,13C]-labeled proteins. *Journal of biomolecular NMR*, 66(4):259–271, December 2016.

[100] Ambrish Roy, Alper Kucukural, and Yang Zhang. I-TASSER: a unified platform for automated protein structure and function prediction. *Nature Protocols*, 5(4):725–738, April 2010.

[101] Aliza B. Rubenstein, Manasi A. Pethe, and Sagar D. Khare. MFPred: Rapid and accurate prediction of protein-peptide recognition multispecificity using self-consistent mean field theory. *PLoS computational biology*, 13(6):e1005614, June 2017.

[102] Jeffrey A Ruffolo, Carlos Guerra, Sai Pooja Mahajan, Jeremias Sulam, and Jeffrey J Gray. Geometric potentials from deep learning improve prediction of CDR H3 loop structures. *Bioinformatics*, 36(Suppl 1):i268–i275, July 2020.

[103] Amy M. Ruschak and Lewis E. Kay. Methyl groups as probes of supra-molecular structure, dynamics and function. *Journal of biomolecular NMR*, 46(1):75–87, January 2010.

[104] Elena Schmidt and Peter Güntert. A New Algorithm for Reliable and General NMR Resonance Assignment. *Journal of the American Chemical Society*, 134(30):12817–12829, 2012.

[105] Andrew K. Sewell. Why must T cells be cross-reactive? *Nature Reviews Immunology*, 12(9):669–677, September 2012. Number: 9 Publisher: Nature Publishing Group.

[106] Yang Shen, Oliver Lange, Frank Delaglio, Paolo Rossi, James M. Aramini, Gaohua Liu, Alexander Eletsky, Yibing Wu, Kiran K. Singarapu, Alexander Lemak, Alexandr Ignatchenko, Cheryl H. Arrowsmith, Thomas Szyperski, Gaetano T. Montelione, David Baker, and Ad Bax. Consistent blind protein structure generation from NMR chemical shift data. *Proceedings of the National Academy of Sciences of the United States of America*, 105(12):4685–4690, 2008.

[107] John Sidney, Erika Assarsson, Carrie Moore, Sandy Ngo, Clemencia Pinilla, Alessandro Sette, and Bjoern Peters. Quantitative peptide binding motifs for 19 human and mouse MHC class I molecules derived using positional scanning combinatorial peptide libraries. *Immunome Research*, 4:2, January 2008.

[108] Fabian Sievers, Andreas Wilm, David Dineen, Toby J Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, Hamish McWilliam, Michael Remmert, Johannes Söding, Julie D Thompson, and Desmond G Higgins. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, 7:539, October 2011.

[109] Remco Sprangers and Lewis E. Kay. Quantitative dynamics and binding studies of the 20S proteasome by NMR. *Nature*, 445(7128):618–622, February 2007.

[110] Christopher Szeto, Demetra S. M. Chatzileontiadou, Andrea T. Nguyen, Hannah Sloane, Christian A. Lobos, Dhilshan Jayasinghe, Hanim Halim, Corey Smith, Alan Riboldi-Tunnicliffe, Emma J. Grant, and Stephanie Gras. The presentation of SARS-CoV-2 peptides by the common HLA-A□02:01 molecule. *iScience*, 24(2):102096, February 2021.

[111] Martin Christen Frølund Thomsen and Morten Nielsen. Seq2Logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including

sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion. *Nucleic Acids Research*, 40(Web Server issue):W281–W287, July 2012.

[112] Martin Christen Frølund Thomsen and Morten Nielsen. Seq2Logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion. *Nucleic Acids Research*, 40(Web Server issue):W281–W287, July 2012.

[113] Sarah J. Todman, Mark D. Halling-Brown, Matthew N. Davies, Darren R. Flower, Melis Kayikci, and David S. Moss. Toward the atomistic simulation of T cell epitopes automated construction of MHC: peptide structures for free energy calculations. *Journal of Molecular Graphics & Modelling*, 26(6):957–961, February 2008.

[114] Rudi Tong, Rebecca C. Wade, and Neil J. Bruce. Comparative electrostatic analysis of adenylyl cyclase for isoform dependent regulation properties. *Proteins: Structure, Function, and Bioinformatics*, 84(12):1844–1858, 2016. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.25167.

[115] Jugmohit S. Toor, Arjun A. Rao, Andrew C. McShan, Mark Yarmarkovich, Santrupti Nerli, Karissa Yamaguchi, Ada A. Madejska, Son Nguyen, Sarvind Tripathi, John M. Maris, Sofie R. Salama, David Haussler, and Nikolaos G. Sgourakis. A Recurrent Mutation in Anaplastic Lymphoma Kinase with Distinct Neoepitope Conformations. *Frontiers in Immunology*, 9:99, 2018.

[116] Matthias Trautwein, Kai Fredriksson, Heiko M. Moller, and Thomas Exner. Automated assignment of NMR chemical shifts based on a known structure and 4D spectra. *Journal of Biomolecular NMR*, 65(3):217–236, 2016.

[117] Thomas Trolle, Curtis P. McMurtrey, John Sidney, Wilfried Bardet, Sean C. Osborn, Thomas Kaever, Alessandro Sette, William H. Hildebrand, Morten Nielsen, and Bjoern Peters. The length distribution of class I restricted T cell epitopes is determined by both peptide supply and MHC allele specific binding preference. *Journal of immunology (Baltimore, Md. : 1950)*, 196(4):1480–1487, February 2016.

[118] Oleg Trott and Arthur J. Olson. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 31(2):455–461, 2010. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.21334.

[119] Vitali Tugarinov, Peter M. Hwang, Jason E. Ollerenshaw, and Lewis E. Kay. Cross-correlated relaxation enhanced 1H[bond]13C NMR spectroscopy of methyl groups in very high molecular weight proteins and protein complexes. *Journal of the American Chemical Society*, 125(34):10420–10428, August 2003.

[120] Vitali Tugarinov and Lewis E. Kay. Ile, Leu, and Val methyl assignments of the 723-residue malate synthase G using a new labeling strategy and novel NMR methods. *Journal of the American Chemical Society*, 125(45):13868–13878, November 2003.

[121] Michael D. Tyka, Daniel A. Keedy, Ingemar André, Frank Dimaio, Yifan Song, David C. Richardson, Jane S. Richardson, and David Baker. Alternate states of proteins revealed by detailed energy landscape mapping. *Journal of Molecular Biology*, 405(2):607–618, January 2011.

[122] Pramodh Vallurupalli, Guillaume Bouvignies, and Lewis E. Kay. Studying ”invisible” excited protein states in slow exchange with a major state conformation. *Journal of the American Chemical Society*, 134(19):8148–8161, May 2012.

[123] Randi Vita, Swapnil Mahajan, James A. Overton, Sandeep Kumar Dhanda, Sheridan Martini, Jason R. Cantrell, Daniel K. Wheeler, Alessandro Sette, and Bjoern Peters. The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Research*, 47(D1):D339–D343, January 2019.

[124] R. C. Wade, R. R. Gabdoulline, and F. De Rienzo. Protein interaction property similarity analysis. *International Journal of Quantum Chemistry*, 83(3-4):122–127, 2001. _eprint: https://www.onlinelibrary.wiley.com/doi/pdf/10.1002/qua.1204.

[125] Andrew Waterhouse, Martino Bertoni, Stefan Bienert, Gabriel Studer, Gerardo Tauriello, Rafal Gumienny, Florian T. Heer, Tjaart A. P. de Beer, Christine Rempfer, Lorenza Bordoli, Rosalba Lepore, and Torsten Schwede. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Research*, 46(W1):W296–W303, 2018.

[126] Kathy Y. Wei, Danai Moschidi, Matthew J. Bick, Santrupti Nerli, Andrew C. McShan, Lauren P. Carter, Po-Ssu Huang, Daniel A. Fletcher, Nikolaos G. Sgourakis, Scott E. Boyken, and David Baker. Computational design of closely related proteins that adopt two well-defined but structurally divergent folds. *Proceedings of the National Academy of Sciences of the United States of America*, 117(13):7208–7215, 2020.

[127] Fan Wu, Su Zhao, Bin Yu, Yan-Mei Chen, Wen Wang, Zhi-Gang Song, Yi Hu, Zhao-Wu Tao, Jun-Hua Tian, Yuan-Yuan Pei, Ming-Li Yuan, Yu-Ling Zhang, Fa-Hui Dai, Yi Liu, Qi-Min Wang, Jiao-Jiao Zheng, Lin Xu, Edward C. Holmes, and Yong-Zhen Zhang. A new coronavirus associated with human respiratory disease in China. *Nature*, 579(7798):265–269, 2020.

[128] Kurt Wüthrich. *NMR of Proteins and Nucleic Acids*. Wiley, New York, 1986.

[129] Chen Yanover and Philip Bradley. Large-scale characterization of peptide-MHC binding landscapes with structural simulations. *Proceedings of the National Academy of Sciences*, 108(17):6981–6986, April 2011. Publisher: National Academy of Sciences Section: Biological Sciences.

[130] Mark Yarmarkovich, Quinlen F. Marshall, John M. Warrington, Rasika Premaratne, Alvin Farrel, David Groff, Wei Li, Moreno di Marco, Erin Runbeck, Hau Truong, Jugmohit S. Toor, Sarvind Tripathi, Son Nguyen, Helena Shen, Tiffany Noel, Nicole L. Church, Amber Weiner, Nathan Kendsersky, Dan Martinez, Rebecca Weisberg, Molly Christie, Laurence Eisenlohr, Kristopher R. Bosse, Dimiter S. Dimitrov, Stefan Stevanovic, Nikolaos G. Sgourakis, Ben R. Kiefel, and John M. Maris. Cross-HLA targeting of intracellular oncoproteins with peptide-centric CARs. *Nature*, pages 1–8, November 2021. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Cancer immunotherapy;MHC class I Subject_term_id: cancer-immunotherapy;mhc-class-i.

[131] Jonathan W. Yewdell and Jack R. Bennink. Immunodominance in major histocompatibility complex class i–restricted t lymphocyte responses. *Annual Review of Immunology*, 17(1):51–88, April 1999. Publisher: Annual Reviews.

[132] Yiyuan Yin and Roy A. Mariuzza. The multiple mechanisms of T cell receptor cross-reactivity. *Immunity*, 31(6):849–851, December 2009.

[133] Ian A. York, Michael A. Brehm, Sophia Zendzian, Charles F. Towne, and Kenneth L. Rock. Endoplasmic reticulum aminopeptidase 1 (ERAP1) trims MHC class I-presented peptides in vivo and plays an important role in immunodominance. *Proceedings of the National Academy of Sciences*, 103(24):9202–9207, June 2006. Publisher: National Academy of Sciences Section: Biological Sciences.

[134] Lingxiao Zhao, Min Zhang, and Hua Cong. Advances in the study of HLA-restricted epitope vaccines. *Human Vaccines & Immunotherapeutics*, 9(12):2566–2577, December 2013.

Appendix A

Algorithms

Algorithm 1: Algorithm to enumerate solutions and create support sets of each residue.

Data: Initial support sets of all residues, Φ
Data: Propositional formula
Result: Support sets of all the residues, Φ'
while $\Phi \neq \emptyset$ **do**
  $s_j \leftarrow FetchRotamer(\Phi)$ ▷ Fetch a rotamer from set Φ for some residue
  $aux \leftarrow (\wedge_{r_i \in R} \neg X_{ij})$;
  $result \leftarrow solve(F \wedge aux)$;
  if $result = UNSAT$ **then**
    $\Phi \leftarrow \Phi - \{s_j\}$;
    $Formula \leftarrow Formula \wedge (\vee_{r_i \in R} X_{ij})$;
  else
    for $r_i \in R$ **do**
      $\Phi' \leftarrow \Phi' \cup \{s_j\}$
    end
  end
end

Algorithm 2: Algorithm to encode a rotamer as exactly one clause.

Data: Literal, X_{ij}
Data: Formula
Result: Formula
$Formula \leftarrow AtMostOne(X_{ij}) \wedge AtLeastOne(X_{ij})$;

Algorithm 3: Algorithm to encode a rotamer as atmost one clause.

Data: Literal, X_{ij}
Data: Formula
Result: Formula
for $s_j^{r_i} \in \Phi^{r_i} : j \leftarrow 1\ to\ m-1$ **do**
  for $s_k^{r_i} \in \Phi^{r_i} : k \leftarrow j+1\ to\ m$ **do**
    $Formula \leftarrow Formula \wedge (\neg X_{ij} \vee \neg X_{ik})$;
  end
end

Algorithm 4: Algorithm to encode a rotamer as atleast one clause.

Data: Literal, X_{ij}
Data: Formula
Result: Formula
for $s_j^{r_i} \in \Phi^{r_i}$ **do**
  $F \leftarrow F \vee X_{ij}$;
end
$Formula \leftarrow Formula \wedge F$

Algorithm 5: Overview of our algorithm used to shrink side chain rotamer space for an input protein.

Data: Protein structure
Result: Reduced super support set Φ' containing support sets for all the residues $r_i \in R$
▷ Eliminate side chain rotamers if they clash with their backbone.
for $r_i \in R$ **do**
  for $s_j^{r_i} \in \Phi^{r_i}$ **do**
    if *(Clash($s_j^{r_i}$))* **then**
      EliminateRotamer($s_j^{r_i}$)
    end
  end
end
▷ Encode side chain rotamers as boolean variables and add clauses stating that one and only one rotamer can be used for each residue at a time.
for $r_i \in R$ **do**
  for $s_j^{r_i} \in \Phi^{r_i}$ **do**
    $Formula \leftarrow ExactlyOne(X_{ij})$
  end
end
▷ Encode pairwise side chain rotamers as hard constraints stating that no two rotamers can clash with one another.
if *(Clash($s_j^{r_i} \in \Phi^{r_i}$* ***and*** *$s_m^{r_n} \in \Phi^{r_n}$))* **then**
  $Formula \leftarrow Formula \cup \{\neg X_{ij} \vee \neg X_{mn}\}$
end
$\Phi' \leftarrow EnumerateSolutions(Formula)$

Appendix B

Tables

Table B.1: HLA alleles in PDB.

Allele	Total number of structures in PDB	Peptide length
B*35:08	1	8
B*08:01	6	8
B*35:01	1	8
B*51:01	1	8
A*24:02	3	8
A*02:01	1	8
B*39:01	1	8
B*52:01	1	8
B*18:01	2	8
B*27:05	11	9
A*02:01	97	9
A*24:02	2	9
B*44:02	7	9
B*81:01	3	9
B*27:09	8	9
B*58:01	5	9

B*46:01	1	9
A*01:01	3	9
E*01:03	7	9
B*53:01	2	9
A*11:01	3	9
B*51:01	1	9
B*07:02	6	9
B*35:01	11	9
B*57:03	3	9
A*68:02	2	9
B*15:01	4	9
A*30:03	3	9
B*57:01	7	9
C*04:01	1	9
B*42:01	3	9
B*44:05	4	9
B*14:02	2	9
B*37:01	3	9
B*27:06	1	9
B*08:01	4	9
A*68:01	1	9
A*03:01	4	9
B*27:04	1	9
B*44:03	4	9
C*08:01	1	9
C*06:02	2	9
B*27:03	1	9
G*01:01	3	9
B*40:02	1	9

B*18:01	1	9
B*39:01	1	9
A*30:01	1	9
C*05:01	1	9
E*01:01	3	9
C*07:02	1	9
B*58:03	1	9
B*57:01	6	10
A*68:01	1	10
A*02:01	34	10
A*02:06	1	10
A*03:01	1	10
B*44:05	1	10
B*58:01	1	10
A*02:03	1	10
A*01:01	1	10
B*35:08	1	10
B*27:09	1	10
A*11:01	5	10
B*27:03	1	10
B*44:02	1	10
B*44:03	1	10
B*07:02	1	10
B*35:01	1	10
B*27:05	2	10
A*24:02	6	10
B*40:01	1	10
B*15:01	1	10
A*02:07	1	10