Structure-based methods for binding mode and binding affinity prediction for peptide-MHC complexes

Dinler A. Antunes\textsuperscript{a}, Jayvee R. Abella\textsuperscript{a}, Didier Devaurs\textsuperscript{a}, Maurício M. Rigo\textsuperscript{b} and Lydia E. Kavraki\textsuperscript{a}  

\textsuperscript{a}Computer Science Department, Rice University, Houston, Texas, USA;  
\textsuperscript{b}School of Medicine, Pontifical Catholic University of Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, Brazil

Abstract: Understanding the mechanisms involved in the activation of an immune response is essential to many fields in human health, including vaccine development and personalized cancer immunotherapy. A central step in the activation of the adaptive immune response is the recognition, by T-cell lymphocytes, of peptides displayed by a special type of receptor known as Major Histocompatibility Complex (MHC). Considering the key role of MHC receptors in T-cell activation, the computational prediction of peptide binding to MHC has been an important goal for many immunological applications. Sequence-based methods have become the gold standard for peptide-MHC binding affinity prediction, but structure-based methods are expected to provide more general predictions (i.e., predictions applicable to all types of MHC receptors). In addition, structural modeling of peptide-MHC complexes has the potential to uncover yet unknown drivers of T-cell activation, thus allowing for the development of better and safer therapies. In this review, we discuss the use of computational methods for the structural modeling of peptide-MHC complexes (i.e., binding mode prediction) and for the structure-based prediction of binding affinity.

INTRODUCTION

Although often imagined as a defense system waiting for an infection, our immune system is also constantly engaged in surveillance and maintenance of a complex microbiome (1, 2). While effective responses must be triggered against cancer cells and dangerous bacteria, harmful responses against healthy cells and gut bacteria must be avoided (3). Other potentially harmful impacts of immune responses that are undesirable include autoimmune reactions (4), as well as reactions to therapeutic products (5, 6) or tissue transplantation (7). For all these reasons, the ability to predict what triggers an immune response is of great biomedical interest.

The ability of a given substance to trigger an immune response is referred to as immunogenicity (6, 8). In a broader sense, immunogenicity refers to the activation of both sides of adaptive immunity: cellular response (mediated by cytotoxic cells) and humoral response (mediated by antibodies). The activation of T-cells is a decisive step in both cases (8, 9), and it will also be referred to as immunogenicity here. T-cells are a special type of lymphocyte that undergo a complex maturation and selection process, which makes them capable of recognizing “non-self” peptides (10). Note that we are referring to T-cells in general; different subtypes of T-cells are involved on each side of adaptive immunity (Fig. 1).

T-cells only recognize peptides displayed by Major Histocompatibility Complex (MHC) receptors (8, 11). Specifically, the T-cell receptors (TCRs) of cytotoxic T-cells can only recognize peptides displayed by class I MHC receptors (MHC-I), while the TCRs of helper T-cells can only recognize peptides displayed by class II MHC receptors (MHC-II) (Fig. 1). Given their central role in both types of responses, MHC receptors have long been the focus of many studies in computational biology (12–16).

Binding to MHC receptors is a prerequisite for peptide immunogenicity (9, 17, 18). In turn, immunogenic peptides are needed for peptide-based vaccine design and cancer immunotherapy. Additional information on this topic can be found in reviews on epitope discovery (19, 20) and reverse vaccinology (21, 22). In this context, sequence motifs (14) and scoring matrices (23, 24) were among the first computational methods used to perform sequence-based binding affinity prediction. They were quickly overpowered by statistical learning algorithms (25–27), which remain the gold standard in the field (28–30).

Despite their unquestionable usefulness, sequence-based methods have known limitations. For instance, statistical learning methods require an experimental dataset for training, and predictions can be biased by the composition of this training dataset (22, 31). Therefore, predictions for MHC variants (i.e., allotypes) with larger datasets available for training tend to be more reliable than predictions for less studied allotypes. These gaps in the training data can be a limitation for some of the most interesting medical applications, such as personalized can-
lyze the impact of post-translational amino acid modifications, by sequence analysis alone. For instance, it can be used to explore many other questions that cannot be addressed without access to structural information about pMHC complexes can be optimized for allotype-specific training datasets (22). In addition, prediction relies on the biochemical properties of the amino acids involved in pioneer studies in the 90’s (12, 34, 35), structure-based prediction is underpinned by structure-based methods (22). As discussed in (44) producing several approaches for structural prediction, analysis, and refinement (45, 46). A particular domain of molecular modeling relates to the prediction of the bound structure of protein-ligand complexes; a problem usually addressed with computational methods known as molecular docking tools (47–50). The first application focuses on accurately predicting the 3D conformation of the ligand, upon binding to the target receptor. The second one focuses on checking a large number of potential ligands and selecting the ones that can bind to the target receptor. Both applications share a central challenge: accounting for ligand flexibility. The greater the number of flexible bonds in a ligand, the greater the number of “shapes” (i.e., conformations) it can adopt. To determine the best possible binding mode, a docking method must consider these alternative conformations, in addition to the position and orientation of the ligand inside the receptor’s binding cleft. This search process is referred to as sampling (48).

Fig. 1. Schematic view of the role of MHCs in T-cell activation. Class I Major Histocompatibility Complexes (MHC-I) are present in almost every cell and involved in the surface presentation of peptides derived from intracellular proteins. On the other hand, class II MHCs are present only in “professional” antigen-presenting cells (phagocytes) and involved in the surface presentation of peptides derived from extracellular proteins. The recognition of displayed peptide-MHC complexes by the T-cell receptor (TCR) triggers T-cell activation, clonal expansion and immunological memory. While cytotoxic T-cells (CD8+) mediate cellular immunity, helper T-cells (CD4+) control the humoral response and have other regulatory roles. CD stands for cluster of differentiation.
As discussed in previous publications, sampling algorithms can be divided into three general categories: shape matching, systematic search and stochastic search (48, 53, 54). Briefly, matching algorithms perform geometric-based evaluations on how much the shape of the ligand fits the shape of the receptor’s binding cleft (47), often using a graph-based representation of the ligand’s structure (47, 55). These methods are usually applied to perform a first exploration of the ligand’s rotational and translational degrees of freedom, without exploring its conformational flexibility (which is known as rigid docking) (55). On the other hand, systematic search algorithms explore all the degrees of freedom of the ligand (e.g., through exhaustive search, fragment-based search, or conformational ensemble search) (48, 53, 54). These methods are much more accurate than matching algorithms, but their computational cost prevents them from being used for larger ligands. Finally, stochastic search algorithms randomly explore the degrees of freedom of the ligand, using different heuristics to guide the exploration (e.g., Monte Carlo, genetic algorithms, tabu search or swarm optimization) (48, 53). As further discussed in the following sections, the size and flexibility of MHC-binding peptides represents a challenge that could not be efficiently handled even by stochastic algorithms, thus requiring additional strategies to make the sampling problem computationally tractable.

Regardless of the sampling method, some kind of ranking of the sampled conformations is needed to guide the sampling and select the best binding mode. This ranking is based on a “quality” assessment of the ligand conformations, which is referred to as scoring. Note that the scores used to rank conformations do not necessarily correspond to accurate binding affinity estimates. In fact, as the number of evaluated poses can be extremely large, scoring functions usually favor computational efficiency over accuracy (56). To achieve that, numerous scoring methods depart from explicitly calculating all relevant interactions between ligand and receptor at the atomic level.

Besides allowing for the assessment and comparison of different conformations of a given ligand to a given protein, a scoring method can also be used for screening (i.e., to assess how strongly different ligands might bind to a given protein). First, scoring methods can help distinguish between ligands that bind and ligands that do not bind the protein, in a purely qualitative manner. This requires performing a binary classification to separate so-called binders from non-binders, based on their respective scores. In this case, ligand scores do not have to correspond to actual binding affinities, as only relative differences between these scores are evaluated. Second, when using scoring methods that are biophysically accurate, one can quantitatively predict actual binding affinities. The capability of a scoring method to do that is usually assessed by evaluating the correlation between these predicted binding affinities and experimental binding affinities, and not by evaluating whether they match exactly. The differences between the qualitative and quantitative applications of scoring functions are clearly illustrated in Figure 3 of (57).

Note that the main assumption underlying a docking-based binding affinity prediction is that the binding free energy of the complex can be approximated by the minimum internal energy of the system (58). In turn, the internal energy of the system is estimated by the scoring function, for each sampled conformation of the complex. Therefore, the accuracy of the binding affinity predicted by a molecular docking tool depends on the quality of both the sampling and the scoring. First, the sampling algorithm have to succeed in generating a conformation of the complex that presents the native set of stable interactions between ligand and receptor. Then, these key interactions must be identified by the scoring function, and properly summarized into an approximated binding affinity. In other words, insufficient sampling can hinder the docking prediction as much as an inaccurate scoring function. Insufficient sampling becomes an even greater issue in the case of highly-flexible ligands or flexible binding sites, since the search space becomes even larger and there is less confidence that the best values of the scoring function can be reached. It is also important to note that these components (i.e., sampling and scoring) and applications of structural prediction methods (i.e., geometry optimization and virtual screening) can be explored separately, or in a combined manner. In the context of docking-based virtual screening, for instance, a given scoring function can be used to (i) rank different conformations of each ligand to guide the sampling, (ii) rank different ligands to identify strong binders, and (iii) estimate the binding affinity of selected ligands. Here, we will discuss how each one of these components/applications was explored in the context of pMHC structural analysis.

**COMPUTATIONAL METHODS FOR BINDING MODE PREDICTION**

In this section we review publications focused on describing and validating methods for accurate binding mode prediction of pMHC complexes. Note that many additional publications report ad-hoc approaches to predict the structure of pMHC complexes as part of larger pipelines for epitope discovery or rational vaccine design (59–63), without focusing on accurate and reproducible binding mode prediction.

**Evaluation of sampling methods**

The two standard experiments for validating a docking method are self-docking and cross-docking (Fig. 2). Both methods rely on the use of experimentally-determined crystal structures of known complexes as controls. The accuracy of the method can be measured through the deviation (i.e., the “error”) between the predicted binding mode and the corresponding crystal structure. This error is usually assessed by calculating the Root Mean Square Deviation (RMSD) for the peptide.
An all-atom RMSD below 2 Å is classically considered a successful reproduction of the native binding mode (64, 65).

Self-docking, also known as re-docking, focuses solely on sampling the ligand. For each target complex, the structures of the ligand and of the receptor are separated, the conformation of the ligand is randomized, and the method under evaluation is used to predict the best binding mode of the ligand to the receptor from the same (self) co-crystallization. On the other hand, cross-docking consists of predicting the binding mode of the ligand to a different conformation of the receptor (e.g., a model or a structure from a different co-crystallization). Therefore, cross-docking usually requires some type of relaxation or sampling of the receptor, in addition to that of the ligand.

Common strategies to make sampling tractable

Binding mode prediction for pMHC complexes is more challenging than most docking problems in drug discovery. Indeed, most drug-like ligands have less than 10 flexible bonds, while MHC-binding peptides usually have more than 30 flexible bonds (even more than 50 for MHC-II). Interestingly, data from the first crystal structures of pMHC-I complexes suggested the existence of conserved structural patterns (12, 17), which were imposed by structural constraints in the binding cleft (Fig. 3). Aiming at leveraging these structural constraints and limiting the computational cost of sampling, three strategies have been devised to predict the binding modes of pMHC complexes: constrained backbone prediction, constrained termini prediction, and incremental prediction (Table 1).

Constrained backbone prediction

The binding cleft of MHC-I receptors is “closed” at both ends (Fig. 3A and 3C), with deeper “pockets” allowing for key interactions with the “anchor” residues of peptides. Analysis of the first pMHC-I crystal structures suggested a conserved conformation of the peptide’s backbone, despite the diversity of amino acid sequences (i.e., the diversity of side chains) (34, 70). These observations justified the use of a backbone template that is kept rigid or constrained during docking (Table 1). Although a backbone template simplifies the problem, the same template cannot be used for MHC-I and MHC-II (Fig. 3), or even for different MHC allotypes. The conformation of the peptide’s backbone is impacted by the composition of the different pockets inside the binding cleft, and the presence of alternative anchor residues. In addition, different templates are required for peptides of different lengths binding to a given MHC allotype (34, 70).

In this context, the work in (17) proposed using a library of crystallographic templates. They utilized the backbone of both the peptide and the MHC as a template, filling in the side-chains of the target sequence using rotamer libraries and the MOIL package (101). The method of utilizing a “clean” backbone to which desired side-chains are added is known as threading (102). Note that the term threading is also used to refer to another molecular modeling method, applied by tools such as I-TASSER (103). Despite promising results on the prediction of the buried side chains of the peptide, they noticed that a general rotamer library from PDB-deposited structures did not include some side-chain conformations observed in pMHC crystal structures (17). Therefore, the generality and accuracy of their predictions was to some extent limited by the small dataset of available pMHC crystal structures.
The growing number of pMHC crystal structures continued to reveal additional backbone variation. To try and reduce biases introduced by the template, other methods added steps of backbone sampling or refinement to the docking process. For instance, the work in (71) proposed pDOCK. This method combines homology modeling of the MHC receptor, positioning of the peptide based on crystal structures, and refinement of the binding site residues using the Internal Coordinate Mechanics (ICM) docking algorithm (104) and a biased Monte Carlo procedure. pDOCK was validated in a self-docking experiment with 186 pMHC complexes (149 MHC-I and 37 MHC-II), reporting average backbone-atom RMSDs of only 0.32 Å (computed for the 9-mer “core” residues). The accuracy of pDOCK in terms of all-atom RMSD was not reported.

Another method using a backbone template was proposed in (73). This approach was based on the Rosetta FlexPepDock refinement protocol (105) and validated through a series of cross-docking experiments using 30 selected crystal structures. Interestingly, the authors report good results even when the template is known to come from a peptide bound to a different MHC allotype (best all-atom RMSD of 1.8 Å among the top 5 ranking conformations). However, the selected dataset was limited to 9-mers bound to MHC-I, and presented small backbone RMSD differences between template and target (the largest difference being 1.35 Å).

Most of these methods were not made available as software or webserver, which might have limited their use by other groups. The first webserver for the structural prediction of pMHC complexes was MHCSim (68). MHCSim relied on sequence alignment to find the closest template from a curated dataset of crystal structures, and side chains were mutated on both ligand and receptor. Rather than providing binding mode prediction, the goal of MHCSim is to generate initial pMHC structures for molecular dynamics (MD) simulations. More recently, the DockTope webserver was proposed in (65) (soon to be available at tools.iedb.org/docktope). DockTope relies on a template-based docking with AutoDock Vina (106), and a refinement loop involving energy minimization followed by a new round of docking (65, 70). DockTope was validated through the cross-docking of 135 non-redundant pMHC-I structures, reporting an average all-atom RMSD of only 1.96 Å. These results present DockTope as a valuable tool for the geometry optimization of pMHC-I complexes. Unfortunately, it only provides predictions for key allotypes for which conserved backbone conformations of the peptide have been observed.

Constrained termini prediction

An alternative assumption that is potentially more general, is that the locations of termini residues are more conserved than the conformation of the backbone (12). Depending on their implementation, “constrained termini” approaches can generalize across MHC-I allotypes because MHC-I binding clefts all have approximately the same length, and termini residues will be constrained by the same pockets. That was the rationale behind the pioneer studies in (34) and (35). The work in (35) proposed a modeling method involving a multiple-copy algorithm (107) to dock the termini residues, followed by a loop closure algorithm to fill the middle residues (108). This general strategy was further explored and perfected by others (69, 75–78).

The most recent implementation of the “constrained termini” strategy is GradDock (79). GradDock combines a fast peptide binding simulator with a Rosetta-based ranking function specifically designed for pMHC-I, and it is available for download (bel.kaist.ac.kr/research/GradDock). This method was tested through both self-docking (107 complexes) and cross-docking (70 complexes), providing impressive results (average all-atom RMSD around 2.5 Å). GradDock results suggest that fast virtual screening of pMHC complexes might be possible, and that the conserved termini assumption might be general enough to provide predictions across MHC-I allotypes. On the other hand, the greater all-atom RMSD observed in some cases suggest this might not be the best tool for geometry optimization.

The authors of GradDock also discuss the limitations imposed by the constrained termini strategy, having excluded from their analysis known cases of alternative binding modes. A notable example is that of a melanoma-derived 9-mer peptide bound to a highly prevalent human MHC (HLA-A*02:01), which uses an alternative anchor and has an unusual backbone conformation (PDB code 2GTW). Being an exception to observed patterns, this complex cannot be predicted by methods relying on constrained termini or constrained backbone strategies (65, 79). Since experimental data on alternative binding modes is still limited, especially considering the diversity of MHC allotypes, it is difficult to evaluate the actual impact of imposing such constraints.

Incremental prediction

As considering the entire conformational space of the peptide was impractical without constraints, another proposed strategy focused on incrementally exploring the flexibility of the ligand (e.g., one residue at a time). A fragment-based docking strategy was first proposed by (80) using the package CONGEN (109) and a similar approach was proposed by (81), using the BRUGEL package (110). This incremental strategy was later revisited with the publication of DynaPred (82). Instead of using a docking tool, DynaPred relies on a short MD simulation to sample each peptide residue inside the binding cleft. DynaPred uses a backbone template from crystal structures to help position amino acids in the binding cleft, but allows for the flexibility of this backbone during the simulation. Conformations from independent residues are then “stitched” together, and a minimization protocol is used to generate the final conformation.

More recently, the work in (43) proposed the use of an incremental meta-docking approach called DINC. DINC is not
a traditional fragment-based docking tool (111), and does not explore the residues independently. Instead, DINC involves incrementally docking overlapping fragments with a growing number of atoms, while maintaining the number of flexible bonds constant during this incremental process (43, 111, 112). DINC handles the fragment expansion and the parallelization of the search, while relying on a regular docking tool, such as AutoDock4 (113), to perform the sampling. DINC was developed as a general tool for docking large ligands, and is available as a webserver (http://dinc.kavrakilab.org/). In the context of pMHC structural prediction, a customized version of DINC was tested through self-docking of a diverse set of known structures. Despite being a small dataset (25 structures), it included very different binding modes (e.g., 10 different human MHC-I allotypes and peptides of different lengths), and very challenging complexes (e.g., the unusual conformation under PDB code 2GTW). The reported average all-atom RMSD was 1.92 Å, and the results were presented as a proof-of-concept for a prediction method that could generalize across MHC-I allotypes. However, broader benchmarking of DINC is needed to evaluate its performance and accuracy across known MHC-I and MHC-II allotypes.

Additional challenges for modeling pMHC-II complexes

Although some of the aforementioned methods were applied to both classes of MHCs, MHC-II complexes represent a more challenging problem for computational modeling. MHC-I and MHC-II receptors have analogous functions and share general structural features, such as having a peptide-binding cleft limited by two parallel α-helices and a floor of β-sheets. A closer look, however, reveals key structural differences (Fig. 3). For instance, while the MHC-I cleft is “closed” at both ends and the peptides are forced to adopt a bulged conformation to fit in, the MHC-II cleft is shallower and allows longer peptides to go beyond both ends of the cleft (Fig. 3B and 3D). As a consequence, a given MHC-II allotype can bind to different portions of the same peptide (i.e., have different binding registers) (114). The portion of the peptide binding to the MHC-II is usually 9 amino acids long (84), but MHC-II receptors can bind peptides with up to 25 amino acids (76).

Longer peptides have a greater number of possible registers, but not all possible registers can bind. Similar to MHC-I receptors, there are key “pockets” that are primarily responsible for the binding of “anchor” peptide residues (Fig. 3D). In MHC-II receptors, pockets 1, 4, 6 and 9 appear to be the most crucial determinants for binding (115, 116). These pockets are hydrophobic cavities that favor hydrophobic side chains (116, 117). Nonetheless, the structural prediction of pMHC-II complexes is a challenging task because it entails simultaneously predicting the binding register and the corresponding binding mode (69, 114).

Another peculiarity is that MHC-II receptors are hetero-dimers formed by two analogous chains (α and β), each one encoded by a different gene. Despite not having as many allotypes as MHC-I genes, the binding cleft of MHC-II receptors can be formed by the combination of α and β chains from different genes, which increases the diversity of MHC-II receptors at the cell surface, each one with slightly different peptide-binding requirements.

Despite these additional levels of diversity, the existence of termini anchor residues and the more linear conformation of the core 9-mer allowed for some of the aforementioned modeling methods to be applied to pMHC-II complexes. Most notably, the validation datasets used by (71) and (76) included MHC-II allotypes. In both cases, the validation was focused on the accuracy of the backbone prediction for the binding core (Table 1). Finally, (69) has discussed the potential generality of a docking-based binding mode prediction method for pMHC-II complexes, reporting very good results in both self-docking and cross-docking experiments (with 9-mer core mean all-atom RMSD of 0.73 Åand 1.37 Å, respectively).

COMPUTATIONAL METHODS FOR BINDING AFFINITY PREDICTION

In this section we review methods previously applied for structure-based binding affinity prediction for pMHC complexes. We discuss the differences between methods for qualitative ranking/classification, and methods for quantitative binding affinity prediction.

Qualitative ranking and ligand classification

Scoring methods used to guide sampling (i.e., rank conformations) are very general in nature: they are usually developed to score any protein-ligand complex. However, in order to improve accuracy, some scoring methods are intended for specific groups of ligands and receptors. For instance, a scoring method can be specific to peptides (as opposed to drug-like ligands), or designed specifically for pMHC complexes (Table 2).

Scoring functions for protein–peptide docking

Most protein–peptide docking tools involve energy-based scoring functions. These scoring functions have been previously classified into three main categories: Empirical, semi-empirical and knowledge-based (56).

Empirical scoring functions are inspired by the quantum mechanics / molecular mechanics (QM/MM) formalism, which allows calculating potential energies (118, 119). Such calculation relies on the definition of a force field as a sum of energy terms corresponding to both covalent and non-covalent interactions within and between molecules. Typical energy terms evaluate the bond stretching, angle bending and torsional angles of covalent interactions, as well as Van der Waals and electrostatic
contributions of non-covalent interactions (120). Some studies have shown that the most important term is often the electrostatic one (121). Empirical scoring functions have long been involved in popular docking tools, such as AutoDock, Glide and DOCK, as well as many others (122, 123). To achieve computational efficiency, energy terms assessing atomic interactions can be replaced by terms derived from coarse-grained potentials, such as the Go potential, that evaluate interactions between large pseudo-atoms representing entire amino acids (124). Also for the sake of computational efficiency, rather than considering explicit water molecules, one can implicitly evaluate solvent effects by using Poisson–Boltzmann surface area (PBSA) or generalized Born surface area (GBSA) energy terms (121). By adding an energy term evaluating entropic effects, for example using empirical conformational energy analysis (CFEA) (125) or normal mode analysis (NMA) (121), a force field allows calculating free energies. One can also directly calculate free energies by using free energy perturbation techniques (126).

Semi-empirical scoring functions differ from purely empirical scoring functions in that they do not attempt to include all physical interactions of protein–peptide poses or to recapitulate biophysically-relevant energies (56). Nonetheless, they include biophysically-plausible energy terms that correspond to physical properties describing the protein–peptide interface. The physical properties that are typically considered correspond to non-covalent interactions between peptides and proteins, such as hydrogen bonds, electrostatic and van der Waal interactions, hydrophobic interactions (127), as well as solvation (128, 129) and entropic effects. These energy terms are then added together with multiplicative weights assigned to them. These weights are usually tuned to optimize binding affinity predictions given a dataset of protein-ligand complexes with known structure (130). Classical examples of semi-empirical scoring functions are ChemScore and X-score, but others have been developed (122, 131).

Knowledge-based scoring functions calculate pseudo-energies that are not biophysically meaningful, but that reflect the likelihood for protein–peptide interface properties to be native or native-like (56). These functions are trained (i.e., calibrated) by performing a statistical analysis on available structural data contained in reported protein–peptide complexes (131). More precisely, an interaction potential is calculated by implicitly estimating the change in energy associated with a change in the distance between atoms of a specific type in a peptide and atoms of a specific type in a protein (123). Examples of popular knowledge-based scoring functions are DrugScore, PMF-score and SMoG (130). Note that this methodology can also be applied in a coarse-grained fashion, by considering distances between pairs of residues. Recently, going away from the classical linear regression approach, new scoring functions have been proposed, using a machine-learning approach based on nonlinear regression (132).

Instead of using scoring functions that are only based on energy calculations, attempts have been made to enhance them with additional information, such as co-evolutionary or mutagenesis data (133). Other approaches complement the energetic analyses with structural clustering or sequence-based predictions. In addition, it has often been observed that combining several scoring functions can improve their accuracy (134).

**Ranking of pMHC binding modes**

For sake of simplicity, we will treat the ranking of conformations during sampling as being merely qualitative, and we will discuss quantitative binding affinity prediction of pMHC complexes in a separate section. One of the early approaches used to guide sampling of pMHC complexes was based on empirically-derived residue-contact matrices (83, 127, 135–137). These matrices, also known as statistical-pair potentials, encode how favorable the interaction is between two given residues (138, 139). In (127), it was found that the so-called MJ matrix (138) only worked for MHC allotypes with hydrophobic binding pockets. Therefore, in follow-up studies (83, 127), the parameters of the newer BT matrix (139) were tuned to improve performance across all allotypes. Through a webserver (137), one can use these matrices or one’s own scoring potentials.

FRESNO (75, 140) was one of the first scoring functions specifically developed for pMHC complexes. This scoring function accounts for hydrogen bonding, lipophilic interactions, rotational entropy loss, buried polar-apolar contacts, and desolvation energies. FRESNO initially allowed making accurate predictions for the HLA-A*02:01 allotype. It was then re-implemented with open-source software (86), and its weights were re-calculated, using a more diverse training set including class II HLAs. This allowed making accurate predictions for the class II HLA-DR15 allotype.

Some studies have used statistical learning methods to optimize the weights of the scoring functions considering specific subsets of complexes (e.g., pMHC-II structures) (69, 78, 141, 142), or to predict the correct register of MHC-II-binders (69). For instance, the scoring function used in GradDock was optimized to guide the sampling of pMHC-I complexes (79). The authors improved a scoring function from the popular modeling library Rosetta, testing different combinations of terms and weight values while performing self-docking and cross-docking experiments.

**Peptide classification (binders vs non-binders)**

Structure-based methods have also been used to classify peptides into binders vs non-binders, considering specific MHC allotypes. For instance, the work in (84) reports high predictive power for HLA-DRB1*0402 and HLA-DQB1*0503, while the study in (85) reports an AUC of 0.9 for HLA-DQ3.2/3. In another study, AUC values in the range 0.632–0.821 are reported.
Quantitative binding affinity prediction

Going beyond qualitative ranking of conformations and classification of binders vs non-binders, some methods aim at predicting realistic values of binding affinity. Two approaches for quantitative binding affinity prediction are discussed here (Table 2). First, we describe data-driven methods to derive binding affinity from a single pMHC conformation. Then, we describe simulation-based methods to derive binding affinity from an ensemble of conformations.

Data-driven predictions

Statistical learning methods mentioned in previous sections were applied to learn the weights of a given scoring function. In this section, statistical learning methods are used to predict binding affinities directly from structural features.

The two defining characteristics of data-driven methods are the representation of the dataset (through features or descriptors) and the type of statistical learning model. First, different types of structural features have been investigated: residue-residue contacts (91, 92, 94), general physical-chemical descriptors (90, 93, 146), energy terms from semi-empirical scoring functions (78), and features derived from molecular dynamics simulations (82). Second, several statistical learning models have been used: partial least squares (90, 91, 93), support vector machines (78, 82, 146), and random forests (94). All these methods report high prediction accuracies for their datasets which consist of one or several MHC allotypes.

A more recent structural data-driven approach is the method HLaFFy (95). A statistical pair potential was constructed using the frequency of residue contacts present in the modeled structures of known binders. When the input is a sequence that was not explicitly modeled, a linear optimization problem allows maximizing the constructed statistical pair potential. Finally, a Gaussian process regression scheme is used to go from interaction profiles (encoded by the statistical pair potential) to predicted binding affinity values. HLaFFy had an average prediction accuracy of 82.5% using 5-fold cross-validation.

Note that the limitations of statistical learning methods for sequence-based predictions, mentioned in section 1, also apply to structure-based data-driven predictions. In fact, biases are even greater given the small number of available pMHC structures. Despite the existence of nearly 10,000 MHC allotypes in humans, there are less than 650 pHLA structures in the PDB. In addition, more than one third of these structures relate to the same HLA allotype (HLA-A*02:01). Therefore, a scoring function using weights learned from all available pHLA structures would certainly overfit HLA-A*02:01, while lacking training on most other allotypes.

Simulation-based predictions

Other methods for binding affinity prediction are based on simulations, such as MD and Monte Carlo. MD simulations track the time evolution of a molecular system using a potential energy function, also known as a force field. MD-based methods hold a lot of promise in that they are completely model-based and do not require any experimental data for training. However, they are the most demanding computationally, and as such have most often been used to analyze important contacts in pMHC complexes with known affinity/immunogenicity (141, 147–154). We direct the interested reader to a dedicated review on the use of MD for pMHC systems (155). Here we highlight some recent work not covered in that review.

By far, the most extensive use of MD for computing pMHC binding affinities has been through the protocol named ES-MACS: enhanced sampling of molecular dynamics with the approximation of continuum solvent (99). This technique relies on calculating free energy with the MM/PB/GB/SA method, which was mentioned in section 4.1.1 (156). The free energy of binding is computed as the free energy of the complex minus the free energies of the peptide and receptor (121). Typically, these free energies are derived from conformations sampled from a single MD simulation of the whole complex. This is in contrast with the so-called 3-trajectory variant, where the free energies
are derived from three separate simulations: for the complex, the protein alone, and the peptide alone. Not performing the 3-trajectory variant means assuming that using the conformations sampled from the peptide bound to the receptor allow computing the free energy of the peptide alone, which neglects important changes in entropy between the bound and unbound states of the peptide.

Large variations in computed free energies have been observed when sampling from single, long MD trajectories, as opposed to ensembles of short simulations (156). Therefore, the study in (99) aimed to produce precise estimates of binding free energies by performing the 3-trajectory variant using ensembles of short MD simulations. For a given system (pMHC complex, MHC, or peptide), 50 replica MD simulations, 4 ns each, were run using different initial velocities starting from a given crystal structure. For 12 diverse peptides bound to HLA-A*02:01, computed binding affinities had a Pearson correlation coefficient of 0.80 with experimental binding (99).

Another interesting work features hierarchical natural move Monte Carlo simulations to explore pMHC detachment processes (100). The pMHC system is represented in a coarse-grained manner: each amino acid is modeled by its alpha carbon, its carbonyl oxygen, and the center of its side chain. The study involves 32 peptides bound to HLA-A*02:01: 100 independent simulations were performed for each peptide until detachment. The authors found that the average time it took for a simulation to completely sample the peptide detaching correlated with experimental binding affinity. Using these average simulation times and an appropriate cutoff, their method achieved an AUC of 0.85 when discriminating binders from non-binders.

Simulation-based predictions have also been applied to MHC-II systems. For instance, the work in (96) used simulated annealing of pMHC-II models and correlated the interaction terms from the AMBER force field with experimental affinity. The study in (97) compares three different ways of computing binding affinity for pMHC-IIs, including simulation, statistical pair potentials and contact analysis (97). However, the conclusion was that, while predictions made by these structure-based methods are significantly better than random, they are still not on par with the leading sequence-based methods. More recently, a computational suite for the optimization of protein and ligand conformations – Proteus – was proposed, featuring MM/PB(GB)SA calculations for pMHC-IIs and considering different pH values (98).

**FUTURE CHALLENGES IN THE SEARCH FOR IMMUNOGENICITY**

A recent study described the use of an ensemble refinement approach to reveal hidden dynamics in crystal structures of pMHC and TCR/pMHC complexes (157). By generating an ensemble of conformations (Fig. 4), all compatible with data from a given X-ray crystallography experiment, the authors illustrated how differing interpretations can be made using a single conformation as opposed to the whole ensemble. In fact, the ensemble derived from some pMHC crystal structures contained not only alternative peptide side chain configurations, but even alternative binding modes, with significantly different backbone conformations and coordination networks. These results are compatible with findings from molecular dynamics studies and highlight the need to consider the whole ensemble of peptide conformations, rather than a single binding mode. This suggests the need for a new paradigm in structural predictions. Instead of trying to find a single top scoring conformation that matches the corresponding binding mode observed in a crystal structure (e.g., cross-docking), a more reasonable goal would be to find an ensemble of peptide conformations that is “equivalent” to that of the crystal structure. This new goal would also require new metrics that could evaluate the accuracy of binding mode predictions in terms of the generated ensemble (157).

![Figure 4. Side view of a pMHC complex after ensemble refinement.](image)

Alternative conformations of selected MHC side chains are depicted in sticks, as well as alternative conformations of the ligand. All alternative conformations are compatible with the x-ray experimental data, and were obtained through a procedure of ensemble refinement (157). The single peptide conformation displayed in the crystal structure is represented in a darker shade of grey. Part of the conformational “frame” of the binding site can also be observed (i.e., a lateral alpha-helix and a floor of beta-sheets, both depicted in cartoon). Graphics were obtained with UCSF Chimera (158).

A paradigm shift from single binding modes to ensembles of conformations would also require new methods for ranking and binding affinity prediction, adding complexity to an already difficult problem. As discussed in previous sections, there is an inherent tradeoff between accuracy and scalability of binding prediction methods. Simulation-based affinity prediction tools can be more accurate and provide additional interpretability of molecular interactions, but cannot be used for peptide screening (97). On the other hand, docking-related scoring functions and data-driven approaches may be scalable, but their accuracy is not yet at the level of sequence-based methods. For instance, scoring functions notoriously suffer from a strong lack of accuracy, which is only partially explained by their focus on
computational efficiency. Several evaluation studies of docking tools (some involving datasets containing pMHC complexes) have shown that existing scoring functions often fail to identify near-native poses (52, 53, 64, 159). Scoring functions also show limitations in distinguishing binders from non-binders, recapitulating known rankings, and reproducing experimental binding affinities (160, 161).

In a broader context, recent findings in drug discovery reveal that the binding kinetics is a decisive factor for drug efficacy and safety (162, 163). In this context, the dissociation rate constant \( K_{\text{off}} \) becomes a more important measure than binding affinity. However, this value is much harder to estimate computationally. It requires extensive sampling of transition states, derived from multiple paths from bound to unbound states (162). Considering that \( K_{\text{off}} \) is impacted by the residence time (i.e., the average time that the ligand stays in the binding site), and that the residence time is usually at the scale of minutes or hours, computing transition states becomes almost impossible with conventional MD simulations (with time scales of \( ns \) to \( \mu s \)). However, recent advances in the development of “optimized” MD sampling algorithms and the fast growing processing capacity of modern CPUs and GPUs are opening new avenues for \( K_{\text{off}} \) prediction (162, 163).

In the context of pMHC complexes, there is also evidence that pMHC stability is a better predictor of immunogenicity than pMHC binding affinity (164). As expected, predicting the \( K_{\text{off}} \) of peptides is a much harder problem as compared to drug-like ligands. Peptides are bigger and have a more complex network of interactions, which results in a much longer unbinding process (100). However, the advances in sampling methods discussed in this review could allow for the fast generation of a meaningful ensemble of bound peptide conformations, despite the use of imperfect scoring functions. In turn, these conformations could be used as seeds for “optimized” MD sampling algorithms, such as adaptive sampling (165–167), allowing for a more efficient prediction of \( K_{\text{off}} \) rates for pMHC complexes.

CONCLUSION

Over the past decades, different sampling methods have been applied to the structural prediction of pMHC complexes (Table 1). These methods are very diverse, relying on a variety of tools and procedures. In addition, research groups have used different datasets and metrics to report their results (e.g., \( \alpha \)-carbon RMSD or all-atom RMSD). Therefore, a fair comparison of all these approaches would be difficult and beyond the scope of this review. Instead, we have focused on highlighting the main assumptions and trade-offs behind these approaches. We have proposed a classification of these methods into three general strategies: constrained backbone prediction, constrained termini prediction and incremental prediction. Note that each strategy solves a different formulation of the pMHC structural prediction problem. For instance, using a backbone template dramatically reduces the conformational space that must be explored, as compared to docking the entire ligand with full flexibility. On the other hand, relying on a template reduces the generality of the proposed method. This is particularly relevant for peptides with unusual binding modes and for MHC allotypes lacking experimental data.

Despite all the challenges, the latest publications show impressive results and suggest that sampling is not anymore a limitation for pMHC binding mode prediction. It is also worth noting that different methods might be better suited for different docking applications (51). For instance, the efficient sampling of GradDock (79) combined with a pMHC-specific scoring function makes this tool potentially useful for large-scale virtual screening, while the higher accuracy of DockTope (65) and DINC (43) make them more suitable for geometry optimization.

On the other hand, the development of fast and accurate scoring functions represent an unmet need for both pMHC binding mode prediction and structure-based binding affinity prediction. Similar to what was discussed for binding mode prediction, different approaches for binding affinity prediction make different assumptions and trade-offs. Fortunately, recent advances in high performance computing are allowing for computation intensive applications, which are expected to have a huge impact on both binding mode and binding affinity prediction. Although no single tool is yet capable of solving all these problems, we are finally getting closer to the point where different structure-based methods can be used to address specific problems in medicinal chemistry. Hopefully, structural methods will soon be combined with sequence-based methods, providing general and accurate predictions that will help researchers and physicians to tackle some of the most challenging health care problems of our time.

ACKNOWLEDGEMENTS

This work was supported by NIH (grant number 1R21CA209941-01), through the Informatics Technology for Cancer Research (ITCR) initiative of the National Cancer Institute (NCI), and by two training fellowships from the Gulf Coast Consortia, through the NLM Training Program in Biomedical Informatics and Data Science (T15LM007093) and through the Computational Cancer Biology Training Program (CPRIT Grant No. RP170593). This study was also financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. Molecular graphics were obtained with UCSF Chimera and UCSF ChimeraX, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from NIH R01-GM129325 and P41-GM103311.
References


Structure-based methods for binding mode and binding affinity prediction for peptide-MHC complexes


Structure-based methods for binding mode and binding affinity prediction for peptide-MHC complexes


Structure-based methods for binding mode and binding affinity prediction for peptide-MHC complexes


### Table 1. Key methods applied to pMHC binding mode prediction

<table>
<thead>
<tr>
<th>Publication (Ref.)</th>
<th>Tool</th>
<th>Method</th>
<th>Validation (size)</th>
<th>Dataset Composition</th>
<th>Best Results (RMSD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schueler-F. et al., 1998 (17)</td>
<td>rotamer libraries for side-chains with MOIL</td>
<td>constrained backbone self-docking (23)</td>
<td>8-mers to 10-mers, 9 class I MHCs</td>
<td>Buried side-chains: 1.2˚A</td>
<td></td>
</tr>
<tr>
<td>Ota et al., 2001 (67)</td>
<td>MCSA-PCR (randomized crystal template)</td>
<td>constrained backbone self-docking (1)</td>
<td>only one 8-mer peptide, 1 class I MHC</td>
<td>Backbone: 0.76˚A, All-atom: np</td>
<td></td>
</tr>
<tr>
<td>Bordner et al., 2010 (69)</td>
<td>ICM + monte carlo + modeling</td>
<td>constrained backbone self-docking (17), cross-docking (18)</td>
<td>11-mers to 15-mers, 17 class II MHCs</td>
<td>Backbone (core): 0.68˚A, All-atom (core): 1.37˚A</td>
<td></td>
</tr>
<tr>
<td>Khan et al., 2010 (71)</td>
<td>pDOCK (modeling + ICM + monte carlo)</td>
<td>constrained backbone self-docking (186)</td>
<td>8-mers to 10-mers, 12 class I MHCs and 9 class II MHCs</td>
<td>C-alpha (core): 0.32˚A, All-atom: np</td>
<td></td>
</tr>
<tr>
<td>Rigo et al., 2015 (65)</td>
<td>DockTope (D1-EM-D2)</td>
<td>constrained backbone cross-docking (135)</td>
<td>8-mers and 9-mers, only 4 class I MHCs</td>
<td>C-alpha: 0.88˚A, All-atom: 1.96˚A</td>
<td></td>
</tr>
<tr>
<td>Fagerberg et al., 2006 (74)</td>
<td>near native MD with simulated annealing steps</td>
<td>constrained backbone self-docking (41), cross-docking (12)</td>
<td>8-mers to 10-mers, only 1 class I MHC</td>
<td>Backbone: 1˚A*, All-atom: 1.5˚A*</td>
<td></td>
</tr>
<tr>
<td>Rosenfield et al., 1993 (35)</td>
<td>multiple copy algorithm, loop closure, minimization</td>
<td>constrained termini self-docking (1)</td>
<td>only 1 9-mer peptide, 1 class I MHC</td>
<td>Backbone: 2.7˚A, All-atom: np</td>
<td></td>
</tr>
<tr>
<td>Sezerman et al., 1993 (34)</td>
<td>rigid-body grid search with EM</td>
<td>constrained termini self-docking (1), cross-docking (4)</td>
<td>1 4-mer, 1 9-mer, 1 13-mer, 2 class I MHCs</td>
<td>Backbone: 0.83˚A, All-atom: np</td>
<td></td>
</tr>
<tr>
<td>Rognan et al., 1999 (75)</td>
<td>modeling, rotamer library, loop closure, EM, SA</td>
<td>constrained termini self-docking (5)</td>
<td>8-mers and 9-mers, 11 class I MHCs</td>
<td>Backbone: 1.2˚A, All-atom: np</td>
<td></td>
</tr>
<tr>
<td>Tong et al., 2004 (76)</td>
<td>docking with ICM + loop closure</td>
<td>constrained termini self-docking (40), cross-docking (15)</td>
<td>8-mers to 10-mers, 8 class I and 6 class II MHCs</td>
<td>C-alpha: 2.2˚A, All-atom: np</td>
<td></td>
</tr>
<tr>
<td>Bui et al., 2006 (77)</td>
<td>PePSSI (computed library of backbones)</td>
<td>constrained termini self-docking (8)</td>
<td>only 9-mers, only 1 class I MHC</td>
<td>All-atom: 1.7˚A</td>
<td></td>
</tr>
<tr>
<td>Bordner et al., 2006 (78)</td>
<td>homology modeling + ICM + monte carlo</td>
<td>constrained termini self-docking (3), cross-docking (23)</td>
<td>8-mers to 10-mers, 7 class I MHCs</td>
<td>Backbone: 0.93˚A, All-atom: np</td>
<td></td>
</tr>
<tr>
<td>Kyeong et al., 2018 (79)</td>
<td>GradDock (gradient-based method)</td>
<td>constrained termini self-docking (107), cross-docking (70)</td>
<td>8-mers to 10-mers, 82 class I MHCs</td>
<td>Backbone: 1.2˚A**, All-atom: 2.5˚A**</td>
<td></td>
</tr>
<tr>
<td>Sezerman et al., 1996 (80)</td>
<td>fragment-based docking, rotamer (CONGEN)</td>
<td>incremental prediction cross-docking (4)</td>
<td>only 4 9-mer peptides, 3 class I MHCs</td>
<td>All-atom: 1.6˚A</td>
<td></td>
</tr>
<tr>
<td>Desmet et al., 1997 (81)</td>
<td>fragment-based docking, rotamer (BRUGEL)</td>
<td>incremental prediction self-docking (1), cross-docking (1)</td>
<td>1 8-mer and 1 9-mer, 1 class I MHC</td>
<td>Backbone: 0.8˚A/1.3˚A, All-atom: np</td>
<td></td>
</tr>
<tr>
<td>Antunes et al., 2017 (43)</td>
<td>DINC (AutoDock4 + incremental method)</td>
<td>incremental prediction self-docking (25)</td>
<td>8-mers to 10-mers, 10 class I MHCs</td>
<td>C-alpha: 0.99˚A, All-atom: 1.92˚A</td>
<td></td>
</tr>
</tbody>
</table>

**RMSD, Root Mean Square Deviation. np, not provided. *Upper threshold for 92% of the dataset. **Reported averages vary depending on sample size, but are similar between self-docking and cross-docking.**

### Table 2. Key methods applied to pMHC binding affinity prediction

<table>
<thead>
<tr>
<th>Publication (Ref.)</th>
<th>Tool</th>
<th>Method</th>
<th>Validation (size)</th>
<th>Dataset Composition</th>
<th>Best Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rognan et al., 1999 (75)</td>
<td>FRESNO scoring function</td>
<td>Correlation (84)</td>
<td>Class I: HLA-A<em>02:01, HLA-A</em>02:04, H-2Kb</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tong et al., 2006 (84)</td>
<td>calibrated scoring function</td>
<td>Classification (139)</td>
<td>Class II: HLA-DQ3.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tong et al., 2006 (85)</td>
<td>calibrated scoring function</td>
<td>Classification and Correlation (84)</td>
<td>Class II: HLA-DRB1<em>0402, HLA-DQB1</em>0503</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Liao et al., 2011 (86)</td>
<td>Modified version of FRESNO scoring function</td>
<td>Correlation (100)</td>
<td>Class I and II: HLA-A2, HLA-DR15, HLA-DR1, and HLA-DR4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Atanasova et al., 2011 (89)</td>
<td>EpiDOCK, generated quantitative matrices</td>
<td>Classification (4540)</td>
<td>Class II: 12 HLA-DRB1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doytchinova et al., 2002 (90)</td>
<td>QSAR</td>
<td>Correlation (266)</td>
<td>Class I: HLA-A*02:01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doytchinova et al., 2004 (91)</td>
<td>QSAR</td>
<td>Correlation (90)</td>
<td>Class I: HLA-A<em>02:01, HLA-A</em>03:01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bordner et al., 2006 (78)</td>
<td>SVM using scoring function terms</td>
<td>Classification (331)</td>
<td>Class I: HLA-A*02:01, H-2Kb</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tian et al., 2009 (93)</td>
<td>QSAR</td>
<td>Correlation (152)</td>
<td>Class I: HLA-A*02:01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Saethang et al., 2013 (94)</td>
<td>random forest using residue-residue contacts and topological descriptors</td>
<td>Classification (1000)</td>
<td>Class I: HLA-A2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mukherjee et al., 2016 (95)</td>
<td>learning statistical pair potentials to use as features for Gaussian process regression</td>
<td>Classification and Correlation (10000)</td>
<td>Any Class I with experimental binding affinity data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Davies et al., 2003 (96)</td>
<td>simulated annealing, AMBER force field</td>
<td>Classification (10)</td>
<td>Class II: 4 HLA-DR1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zhang et al., 2010 (97)</td>
<td>position specific free energy contributions using MD and MM/PBSA</td>
<td>Correlation (3882)</td>
<td>Class II: HLA-DRB1*0101</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Polydorides et al., 2016 (98)</td>
<td>Proteus, computational suite for the optimization of protein and ligand conformations</td>
<td>Correlation (10000)</td>
<td>Any Class II, HLA-DR1, HLA-DQ8</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Correlation: study reports computing affinity values that can be directly compared with experiment. Classification: study reports affinity predictions for the purpose of classifying peptides as binders or non-binders given an appropriate threshold.**

*These studies searched for strong binding peptides, instead of producing a score related to affinity. See text.** Only reported an example use case using their tool.