Published in final edited form as:

J Immunol. 2020 October 01; 205(7): 1962–1977. doi:10.4049/jimmunol.1900918.

Combining 3D modeling with artificial intelligence to increase specificity and precision in peptide:MHC binding predictions

Michelle P. Aranha^{1,2}, Yead S. M. Jewel^{1,2}, Robert A. Beckman^{3,4}, Louis M. Weiner⁴, Julie C. Mitchell⁵, Jerry M. Parks², Jeremy C. Smith^{1,2,*}

¹Department of Biochemistry and Cellular and Molecular Biology, University of Tennessee, Knoxville, TN, 37996

²UT/ORNL Center for Molecular Biophysics, Oak Ridge National Laboratory, Oak Ridge, TN, 37830

³Department of Biostatistics, Bioinformatics, and Biomathematics and Innovation Center for Biomedical Informatics, Georgetown University Medical Center, Washington, DC, USA

⁴Department of Oncology and Lombardi Comprehensive Cancer Center, Georgetown University Medical Center, Washington, DC, USA

⁵Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, 37830

Abstract

The reliable prediction of the affinity of candidate peptides for the major histocompatibility complex (MHC) is important for predicting their potential antigenicity, and thus influences medical applications, such as decisions on their inclusion in T cell-based vaccines. Here, we present a rapid, predictive computational approach that combines a popular, sequence-based artificial neural network method, NetMHCpan 4.0, with 3D structural modeling. We find that the ensembles of bound peptide conformations generated by the programs MODELLER and Rosetta FlexPepDock are less variable in geometry for strong binders than for low-affinity peptides. In tests on 1,271 peptide sequences for which the experimental dissociation constants for binding to the well-characterized murine MHC allele, H-2D^b are known, by applying thresholds for geometric fluctuations the structure-based approach in a stand-alone manner drastically improves the statistical specificity, reducing the number of false positives. Furthermore, filtering candidates generated with NetMHCpan 4.0 with the structure-based predictor led to an increase in the positive predictive value (PPV) of the peptides correctly predicted to bind very strongly (i.e., $K_D < 100$ nM), from 40% to 52% ($p=0.027$). The combined method also significantly improved the PPV when tested on five human alleles, including some with limited data for training. Overall, an average increase of 10% in the PPV was found over the standalone sequence-based method. The combined method should be useful in the rapid design of effective T cell-based vaccines.

*Correspondence to Jeremy C. Smith: smithjc@ornl.gov.

Introduction

Presentation of antigenic peptides by major histocompatibility complex (MHC) proteins is essential for cell-mediated adaptive immunity (1, 2). MHC-bound peptide antigens may be derived from cellular ‘housekeeping’ proteins, tumor cells or from pathogenic viruses and bacteria (3). T cells scan antigen-presenting cells for MHC-bound peptides, discriminating between self and non-self and may activate the immune response upon recognizing a non-self-antigenic peptide (4–6). T-cell binding is extremely specific, and it is estimated that only 50% of the peptides that bind to MHC will be recognized by the T cell and that only about a third of those peptides will be immunodominant (7).

An effective T cell-mediated immune response depends on three main events: peptide processing from its source, stable binding of the peptide (p) to the MHC to form the p-MHC binary complex, and the recognition of the MHC-bound peptide by the T cell receptor (8). Peptide binding to the MHC, a prerequisite for T cell recognition, appears to be particularly important in defining T cell epitopes (8, 9). Immunization with peptides known to bind to an MHC suggests that vaccine peptides with the highest binding affinities to MHC are not only immunogenic but also immunodominant (10–13). Moreover, cancer immunotherapy studies have reported a correlation of high p-MHC binding affinities ($k_D < 10$ nM) with tumor eradication, whereas lower binding affinities ($k_D > 100$ nM) were associated with tumor relapse (14, 15). Distinguishing tight-binding peptides from weak and non-binders is therefore of fundamental importance for the development of epitope-based vaccines and immunotherapies targeting infectious diseases and cancers (16, 17).

Experimental determination and discrimination of binding affinities of large numbers of peptide sequences to highly polymorphic MHC molecules is both expensive and time-consuming. As an alternative, *in silico* approaches for selecting strongly binding antigenic peptides as vaccine candidates have been found to be effective in reducing the burden of experimental validation (18).

As an illustration of the clinical relevance of peptide-MHC binding prediction, a common workflow used to develop cancer vaccines involves the identification of missense mutations in the tumor-expressed proteins, followed by prediction of which mutated proteins will be cleaved into 8- to 11-mer neoantigenic peptides by proteasomes (19–21). Computational prediction of the binding affinities between the specified 8- to 11-mer peptides and one of the many polymorphic MHC alleles of interest then facilitates the selection of the strongest binding neoantigens as vaccine candidates. These peptides, when loaded onto the MHC, are predicted to yield an immune response through recognition by CD8⁺ T-cells. This kind of targeted approach has been found to be effective against murine sarcoma (22), non-small cell lung cancer (23, 24) and melanoma (25–27).

Popular predictive tools for peptide-MHC qualitative and quantitative binding affinities, such as NetMHC (28, 29), NetMHCpan (30, 31), MHCflurry (32), IEDB tools (33) and SYFPEITHI (34, 35), use artificial intelligence (AI) techniques such as artificial neural networks or linear regression-based stabilized matrix methods. These tools support a large number of MHC alleles and peptides 8-15 residues long, and have been developed to predict

peptide-MHC binding affinities by training on sequence-derived features of peptides, position-specific matrices, and binding affinity measurements deposited in the Immune Epitope Database (IEDB) (33) and other sources (34). In addition, in some of these tools, mass spectrometric (MS) peptide elution data were also included in training (31, 32). Of the machine learning tools, those that use artificial neural networks and MS data along with binding affinity measurements have hitherto been more accurate than those using other machine learning methods (36, 37).

Although successful, the machine learning methods have some drawbacks. Firstly, to work well they require a large amount of training data with even sequence coverage and expansive binding affinity ranges for canonical and non-canonical peptide lengths, and do not work as well for MHC alleles for which these data are scarce (38). Sequence-based methods can thus predict peptide binding to only a very small fraction of the highly polymorphic MHC alleles (there are more than 12000 MHC class I alleles in the case of humans). In an effort to alleviate the problem of data scarcity, NetMHCpan uses a “pan-allele” approach, instead of training on separate predictors for each allele. Furthermore, both NetMHCpan and MHCflurry have improved their predictive power by using MS data that takes into account human leukocyte antigen (HLA)-peptide processing steps in addition to the binding affinity data.

Ideally, for a p-MHC binding prediction approach to be most useful in antigen identification and vaccine development, it would have both high statistical sensitivity (i.e., would predict a large fraction of the true positives to bind) and high statistical specificity (i.e., would predict a large fraction of true negatives not to bind) (39). However, of particular relevance to the present study is the finding that, even with the above-mentioned improvements the performance of NetMHCpan and MHCflurry evaluated on independent datasets for seven major MHC alleles revealed variability in the accuracy of the predictions amongst the different alleles and peptide lengths that can often result in a considerable number of false negatives and false positives (40). Consistent with this observation, evaluation of results of the automated benchmarking of the predictive algorithms by the IEDB found that, although the algorithms are a valuable predictive resource, even for common alleles they can suffer from problems of low sensitivity, specificity and precision (Table S1) (41). Many studies have focused on increasing the statistical sensitivity of the computational prediction in order to not lose any potential binders. However, a drawback of this approach has been a higher false positive rate (38, 40). A high proportion of false positives or low specificity and precision indicates that a relatively high proportion of peptides predicted to bind strongly, in fact, will not. In applications to vaccine design, this implies that a relatively high proportion of candidate antigens in a vaccine will not be effective.

Here, in an attempt to alleviate the above difficulties, we consider augmenting the successful machine learning-based approaches with 3D structural modeling. Atomistically detailed structural models can be complementary to sequence-based methods while also providing mechanistic insight into peptide-MHC interactions (42). These methods model the conformation of the peptide in the MHC binding groove and try to discriminate between strong and weak binding peptides using various methods, such as correlating binding affinities with physicochemical descriptors in quantitative-structure activity relationship

(QSAR) studies (43), calculating free energies of peptide-MHC binding from molecular dynamics (MD) simulations (44), using statistical pairwise potentials or energetic scoring terms of peptide-MHC complexes from MD, docking or threading approaches (45–47), and expressing conformational stability as root-mean-square fluctuations (RMSF) obtained from MD (48, 49). An advantage of structure-based methods (apart from those using QSAR) is independence from datasets of measured binding affinities. However, a disadvantage has hitherto been that structure-based methods can be computationally relatively expensive, especially when they rely on MD simulations. For potential industrial application in high-throughput personalized cancer vaccine design it may be important that sequenced tumor genomes be processed rapidly. Furthermore, structure-based methods have not yet been clearly demonstrated to be more generally accurate than the more rapid, sequence-based methods. Some approaches use a combination of sequence- and structure-based approaches (50–52), a notable example being the use of the agretopicity index, defined as the difference in the NetMHC scores of putative neoepitopes with their unmutated counterparts, together with the conformational stability of epitopes from MD simulation to predict peptide-MHC binding affinities (48).

In this paper, we present an accurate and rapid structure-based approach for distinguishing strong from weak or non-binding peptides. We use initially as a test case nonameric peptides interacting with a commonly studied murine MHC allele, H-2D^b. The approach consists of generating an ensemble of docked peptide-MHC models and quantifying the backbone conformational variation among the resulting docked poses of the peptide. We demonstrate its effectiveness by predicting the binding of 1,271 peptides to the MHC and comparing with experiment. When applied as a filter to antigens predicted to bind by a popular sequence-based method, NetMHCpan 4.0, the structure-based calculations are found to yield a significant decrease in the fraction of false positives. Further, when tested on human alleles, this approach yielded a consistent increase in the positive predictive value (true positives / (true positives + false positives)) for the prediction of very tightly binding peptides as compared to the sequence-based method alone. Hence, a scheme involving primary screening by a sequence-based method followed by the present structure-based screening method has the potential to increase the proportion of effective antigens in an experimental vaccine.

Methods

As a test case for the methodology, we initially selected the well-characterized mouse MHC allele H-2D^b. A subset of 1271 peptides for which experimental dissociation constants of binding to H-2D^b are available was selected from the IEDB (33). This subset was chosen to have balanced representations of different ranges of binding constants. p-MHC models were constructed for these peptides bound to the H-2D^b MHC allele. Initial 3D models were generated with the homology modeling program MODELLER (53–55) starting from an X-ray crystal structure (PDB ID: 5TIL) of H-2D^b to which a high-affinity peptide (KAPYNFATM) is bound.

Here, 500 models per p-MHC complex were produced and the top-ranked model of each complex was selected based on the MODELLER DOPE score (56). These top-ranked

peptide-MHC models were then refined with FlexPepDock (57) from the Rosetta modeling suite (58). The refinement protocol of Rosetta FlexPepDock involves optimizing the positions of the peptide backbone and side chains together with the receptor side chains.

It has been shown that FlexPepDock can create near-native structures if the input model is within 5.5 Å RMSD of the crystal structure (57, 59). Peptides bound to the MHC have restricted conformational freedom in the binding groove, and hence substituting the residues of the peptide in an existing template of a p-MHC complex results in reasonable initial conformation for any strongly binding peptide, which should allow the FlexPepDock refinement protocol to generate sound p-MHC models. Therefore, the knowledge of the binding pocket, the large degree of structural conservation among peptides bound to a specific MHC allele, and the availability of crystal structures that can be used as templates make p-MHC complexes especially good candidates for refinement by Rosetta FlexPepDock. Indeed, benchmarking results reported for the refinement protocol of Rosetta FlexPepDock using 30 peptide-human leukocyte antigen (HLA) crystal structures showed that the peptides in the top-ranked models had backbone root mean square deviations (RMSDs) of less than 2 Å from those in crystal structures, confirming that Rosetta FlexPepDock can be used to generate useful models of p-MHC complexes (60).

Although the Rosetta FlexPepDock protocol has been extensively benchmarked for p-HLA and other peptide-protein complexes (57, 59, 60), no specific validation exists for FlexPepDock refinement for modeling peptide-H-2D^b complexes. We thus tested the docking protocol by comparing the crystal structures of the 38 non-redundant peptide-H-2D^b complexes with cross-docked modeled complexes. We found that the peptide C_α root-mean-square deviations (RMSDs) between the top-ranked docked models and the corresponding crystal structures for the overall superposition of the docked p-MHC model to its crystal structure were around 1.2 Å (Table 1), giving us further confidence that Rosetta FlexPepDock can produce near-native and reasonably accurate models of p-MHC complexes.

The sequential use of MODELLER and FlexPepDock was found to be computationally efficient and accurate. To achieve high-quality conformational sampling, Rosetta FlexPepDock applies small perturbations to the structure and reoptimizes, using Monte Carlo sampling and energy minimization (57) (61).

We generated 1,000 models per p-MHC complex with FlexPepDock and ranked them based on their Rosetta REF2015 score, which approximates the energy of biomolecular conformations with a linear combination of knowledge- and physics-based potential terms (62). The relationship of such a linear combination of terms to actual binding affinities is unclear a priori. We did not find any distinguishable difference in the Rosetta energy ranges across the strong, weak and non-binding peptides (Fig. S2 A – C). In addition, when sorted by the Rosetta energy, irrespective of the binding class to which a given peptide belonged, the difference in the energy between the top-ranked and 50th ranked p-MHC model was modest (~12 Rosetta energy units) (Fig. S2 B). Furthermore, the scoring terms used in Rosetta FlexPepBind (61, 63–65), a protocol developed to predict peptide binding specificity, did not correlate well with the p-MHC binding affinities (Fig. S2 D – F).

However, during preliminary testing we found that the variation among the peptide backbone poses did correlate with the binding affinities, with tighter binding correlating with less variation in the backbone poses. We therefore selected the top 50 models of each p-MHC complex for conformational analysis.

Example scripts for generating p-MHC models with MODELLER and Rosetta FlexPepDock and the dataset of 1271 peptides that were modeled will be provided upon request. The input parameters used for Rosetta FlexPepDock are provided in the Supporting Information. The software packages used for the molecular modeling are freely available for academic use and Rosetta FlexPepDock is also available in web server form (66).

Training and testing datasets for human alleles

In a second stage of method development, to test the broadness of the applicability of the structure-based approach we chose human alleles HLA-A*02-01, HLA-B*44-03, HLA-B*27-05 to represent cases with varying amount of data availability and HLA-A*23-01 and HLA-A*32-01 to represent alleles that have no template crystal structure. Here, we have constrained the structure-based approach to peptides equal in length to the peptides in the templates, i.e., predominantly 9- and 10-mers. However, to determine the core region of longer peptides, the sequence can be parsed by NetMHCpan, which uses NNAlign (algorithm that determines the linear sequence motif) in its training model to determine the binding core, and this information can be used to guide alignment to the template and docking.

Case 1: Template crystal structure available and large training dataset exists

HLA-A*02-01 with 9-mers: HLA-A*02-01 is a well-studied MHC allele with numerous crystal structures available of the protein in complex with different binding peptides. The prediction performance of many of the sequence-based methods for this allele are excellent, with the reported area under the curve (AUC) values greater than 90% owing to the abundant experimentally determined binding affinity data (>9000 for nonameric peptides) deposited in the IEDB. To make structure-based peptide:HLA-A*02-01 binding predictions, nonameric peptide sequences for which the binding to HLA-A*02-01 allele is known were filtered from the IEDB dataset. From these sequences, 100 very strongly binding peptides ($k_D \leq 100$ nM) were selected as the training set to generate ΔG RMSF thresholds. Structural models were generated as described in the Methods section using PDB ID: 1HHG as the template, which corresponds to 9-mer peptide TLTSNTSV bound to the HLA-A*02-01 allele. Next, a balanced held-out test dataset of 460 peptides was created from the nonamer peptides available in the IEDB HLA-A*02-01 dataset by randomly selecting 115 peptides each from the very strong binding ($k_D \leq 100$ nM), strong binding ($100 \text{ nM} < k_D \leq 1000$ nM), weak binding ($1000 \text{ nM} < k_D \leq 6500$ nM) and non-binding ranges ($k_D > 6500$ nM). There was no overlap between sequences in the training and test set.

Case 2: Template crystal structure available and only a small training dataset exists—Alleles HLA-B*44-03 and HLA-B*27-05 are two instances for which a template crystal structure is available but fewer than 1000 peptide binding observations are available in the IEDB.

HLA-B*44-03 with 9-mers: HLA-B*44-03 has 733 nonameric sequences for which experimental binding affinity data is available in the IEDB (120 sequences with $k_D \leq 100$ nM, 114 sequences with $100 \text{ nM} < k_D \leq 1000$ nM, 94 sequences with $1000 \text{ nM} < k_D \leq 6500$ nM and 405 sequences with $k_D > 6500$ nM). Only 40 very strongly binding peptides were selected as the training set to generate Ca RMSF thresholds. HLA-B*44-03 bound to the 9-mer peptide EEYLQAFTY (PDB ID: 3KPN) was used as a template. A balanced held-out test dataset was created by randomly selecting 80 peptides from each binding class with no overlap between sequences in the training and test set.

HLA-B*27-05 with 10-mers: HLA-B*27-05 has 317 decameric sequences for which binding affinity data is available in the IEDB (67 sequences with $k_D \leq 100$ nM, 53 sequences with $100 \text{ nM} < k_D \leq 1000$ nM, 52 sequences with $1000 \text{ nM} < k_D \leq 6500$ nM and 145 sequences with $k_D > 6500$ nM). Only 16 very strongly binding peptides made up the training dataset for generation of the Ca RMSF thresholds. HLA-B*27-05 bound to the 10-mer peptide KRWILGLNK (PDB ID: 4G9D) was used as the template. A balanced held-out test dataset was created by randomly selecting 50 peptides from each binding class with no overlap between sequences in the training and test set.

Case 3: Crystal structure template not available and large training dataset exists

HLA-A*23-01 bound to 9-mer peptides: Allele HLA-A*23-01 bound to 9-mer peptides is an example for which a template crystal structure is not available. HLA-A*23-01 has 1915 nonameric sequences for which binding affinity data is available in the IEDB (275 sequences with $k_D \leq 100$ nM, 206 sequences with $100 \text{ nM} < k_D \leq 1000$ nM, 276 sequences with $1000 \text{ nM} < k_D \leq 6500$ nM and 1158 sequences with $k_D > 6500$ nM). 100 very strongly binding peptides were selected as the training set to generate Ca RMSF thresholds. Selection of the template for alleles with no available crystal structures was done by selecting the closest homolog that had a p-MHC crystal structure available and with shared significant peptide binding motifs as reported in IEDB or other sources. HLA-A*24-02 has a p-MHC crystal structure available and shares >90 % sequence identity with HLA-A*23-01. Furthermore, both HLA-A*24-02 and HLA-A*23-01 share similar binding motifs with residues W and Y preferred at anchor position 2 and F preferred at anchor position 9. Thus, HLA-A*24-02 bound to the 9-mer peptide VYGFVRACL (PDB ID: 2BCK) was used as the template for modeling peptide binding to HLA-A*23-01. A balanced held-out test dataset was created by randomly selecting 90 peptides from each binding class with no overlap between sequences in the training and test set.

Case 4: Crystal structure template not available and only a small training dataset exists—Alleles HLA-A*23-01 bound to 10-mer peptides and HLA-A*32-01 bound to 9-mer peptides are two instances for which a template crystal structure is not available and fewer than 1000 peptide binding observations exist in the IEDB.

HLA-A*23-01 bound to 10-mer peptides: HLA-A*23-01 has 580 decameric sequences for which binding affinity data is available in the IEDB (119 sequences with $k_D \leq 100$ nM, 102 sequences with $100 \text{ nM} < k_D \leq 1000$ nM, 66 sequences with $1000 \text{ nM} < k_D \leq 6500$ nM and

293 sequences with $k_D > 6500$ nM). Only 19 very strongly binding peptides were selected as the training set to generate Ca RMSF thresholds. HLA-A*24-02 bound to the 10-mer peptide RYPLTFGWCF (PDB ID: 5HGH) was used as the template. A balanced held-out test dataset was created by randomly selecting 66 peptides from each binding class with no overlap between sequences in the training and test set.

HLA-A*32-01 bound to 9-mer peptides: HLA-A*32-01 has 831 nonameric sequences for which binding affinity data is available in the IEDB (299 sequences with $k_D \leq 100$ nM, 135 sequences with $100 \text{ nM} < k_D \leq 1000$ nM, 96 sequences with $1000 \text{ nM} < k_D \leq 6500$ nM and 301 sequences with $k_D > 6500$ nM). 74 very strongly binding peptides were selected as the training set to generate Ca RMSF thresholds. HLA-A*30-03 bound to the 9-mer peptide QIMYNYPAM (PDB ID: 6J29) was used as the template. HLA-A*32-01 and HLA-A*30-03 share $>90\%$ sequence identity. A balanced held-out test dataset was created by randomly selecting 100 peptides from each binding class except from the weak binding class for which all 96 peptides were selected. There was no overlap between sequences in the training and test set.

Testing on datasets of peptides with known immunogenicity

Bjerregaard et al. curated dataset (PMID: 29187854)—The curated dataset by Bjerregaard et al. (PMID: 29187854) consists of 290 neopeptides that are HLA-A*02-01 binders but their experimental binding affinities are not known. However, in this case, owing to the abundance of binding observations for HLA-A*02-01, NetMHCpan is expected to produce accurate binding affinity estimates and these can be used to compare with the present method. Out of the 290 peptides, 239, 49 and 2 neopeptides were predicted by NetMHCpan to belong to the very strong ($k_D \leq 100$ nM), strong ($100 \text{ nM} < k_D \leq 1000$ nM) and weak ($1000 \text{ nM} < k_D \leq 6500$ nM) binding range respectively. 15/239 and 2/49 peptides in the NetMHCpan predicted very strong ($k_D \leq 100$ nM) and strong binding range ($100 \text{ nM} < k_D \leq 1000$ nM) were experimentally validated to elicit a T cell response and thus the dataset contains 17 T cell epitopes.

Koaloğlu-Yalçın et al. curated dataset (PMID: 30377561)—The 195 HLA-A*02-01-nonamer complexes in the curated Koaloğlu-Yalçın et al. (PMID: 30377561) dataset had 78 experimentally validated T cell epitopes and 117 random peptides. Out of the 78 T cell epitopes, 20, 40, and 18 were mutated cancer epitopes, non-mutated cancer epitopes, and viral epitopes, respectively. Once again, for this dataset, even though the T cell reactivities for the 195 HLA-A*02-01-nonamer complexes are known, the quantitative experimental binding affinities of the p:MHC complexes are not known. Therefore, in place of experimental binding affinities, the distribution of the peptides in the different binding ranges as predicted by NetMHCpan is analyzed.

Mutated cancer epitope dataset—18, 1, and 1 of the 20 mutated cancer epitopes were predicted by NetMHCpan to belong to the very strong, strong, and weak binding ranges, respectively.

Non-mutated cancer epitopes—16, 13, 6, and 5 of the 40 non-mutated cancer epitopes were predicted by NetMHCpan to belong to the very strong, strong, weak, and non-binding ranges, respectively.

Non-mutated non-cancer epitopes or viral epitopes—11, 2, 3, and 2 of the 18 viral epitopes were predicted by NetMHCpan to belong to the very strong, strong, weak, and non-binding ranges, respectively.

Random peptides—2, 6, 6, and 103 of the 117 random peptides were predicted by NetMHCpan to belong to the very strong, strong, weak, and non-binding ranges, respectively.

Statistical analysis

Results are expressed as means $\pm$ SEM or ranges, as appropriate. The chi-squared test was performed to compare the positive predictive values obtained by the sequence-based method and the combined sequence and structure-based method. Significance was assumed at $p < 0.05$. The predictive performance was evaluated in terms of the following metrics:

$$\text{Sensitivity} = \frac{TP}{(TP + FN)} \quad \text{Specificity} = \frac{TN}{(TN + FP)} \quad \text{PPV or precision} = \frac{TP}{(TP + FP)}$$

True labels for peptides were determined using the threshold $k_D = 100$ nM unless otherwise stated.

Results

Conservation of conformation in the X-ray crystallographic poses of binding peptides.

We first examined the conformational variability of antigens in the MHC binding groove by comparing the X-ray structures of the murine MHC class I protein (H-2D^b) in complex with 37 non-redundant peptides. The experimental dissociation constants for twelve of these peptides are deposited in the IEDB and are in the nanomolar range. 36 of the 37 peptides are predicted by the sequence-based method NetMHCpan 4.0 to bind very strongly to MHC (Table I).

The 37 peptides assume extended conformations in the groove and show low structural deviations from each other (Fig. 1, A and B). Superposition of the bound peptides onto the bound peptide in a high-resolution reference crystal structure (PDB ID: 1JPG) led to a low root-mean-square deviation (RMSD) of the C $_{\alpha}$ atoms of 0.7 Å. Therefore, the peptide conformation in the 37 crystal structures is largely conserved.

Conservation of conformation in computationally docked poses of strong-binding peptides

To assess the quality of our predictions we categorized peptide:MHC binding affinities in the following four classes:

1)	very strong binders		k_D	100 nM
2)	strong binders	100 nM	$< k_D$	1000 nM
3)	weak binders	1000 nM	$< k_D$	6500 nM
4)	non-binders	6500 nM	$< k_D$	43,500 nM

An important aim in T cell-dependent vaccine design is to construct vaccines containing as high a proportion of very strong MHC binders as possible to increase the likelihood of including actual T cell epitopes. In an exploratory analysis, by following the modeling protocol that uses the programs MODELLER and Rosetta FlexPepDock as described in the Methods section, we modeled the binding to H-2D^b of 100 peptides randomly selected from the IEDB comprising 50 very strong binders (k_D 100 nM) and 50 non-binders (6,500 nM $< k_D$ 43,500 nM). We found that the top-ranked poses of non-binders were often noticeably displaced out of the binding groove whereas those of strong binders had more conserved conformations that resemble peptide conformations in H-2D^b crystal structures (Fig. 1, C and D). Furthermore, the root-mean-square fluctuations (RMSFs) of the C_α atoms among the models constructed for each peptide tended to be lower for the very strong binders than for the non-binders (Fig. 1 E).

Next, following the protocol described in the Methods section, we modeled the large, targeted dataset of 1,271 p-MHC complexes with peptides of known binding affinities to H-2D^b. We selected the top-scoring docked model for each p-MHC complex and partitioned the complexes into the four binding classes based on their experimental binding affinities (i.e., very strong, strong, weak, and non-binding). To quantify the conformational variability within each binding class, we aligned the structures of all the p-MHC complexes in each class to the reference crystal structure and calculated the C_α RMSF of the peptides at each position in the binding groove (i.e., residues 1-9). We performed the same analysis for the set of 37 non-redundant peptide-H-2D^b co-crystal structures. The 3D structures of the peptide pose of the top-ranked models of all docked peptide sequences in each binding class are shown in Fig. S1, A – D. The analysis revealed a consistently increasing trend in C_α RMSF with decreasing binding affinity (Fig. 2 A). The peptide conformations in the crystal structures and the docked models of very strong binders display the smallest C_α RMSF values, while the non-binders exhibit the largest values. For the residues located at and between the anchor positions, 5 and 9, the magnitude of the C_α RMSFs for the very strong binders is even lower than for the set of crystal structures.

To quantify the difference in the C_α RMSF values of each binding class from those of the class of strongest binding peptides, we calculated the difference between the average C_α RMSF values for each of the four dissociation constant ranges and the average in the k_D 100 nM range, i.e. $\langle C_{\alpha} \text{ RMSF} \rangle_i - \langle C_{\alpha} \text{ RMSF} \rangle_{1-100 \text{ nM}}$, where i is the binding class i.e., k_D 100 nM, $100 \text{ nM} < k_D$ 1000 nM, $1000 \text{ nM} < k_D$ 6500 nM, $6500 \text{ nM} < k_D$ 43,500 nM. For the nine residue positions, this difference mostly increases consistently as the binding affinity decreases (Fig. 2 B).

Discriminating very strong binders based on per-residue C_α RMSF

The above calculations show that, on average, the calculated poses of strong binders more closely resemble the binding pose of peptides in crystal structures and are less conformationally variable than the less well-binding peptides. Using this result, in what follows we assume that class I MHC H-2D^b restricts the conformational freedom of strongly binding peptides in the binding groove. Thus, if the docked conformations of any 9-mer peptide closely resemble each other, then we assume that peptide is likely to bind strongly to MHC. We test this hypothesis by assessing whether a low C_α RMSF can discriminate peptides in the very strong binding class from those in the other binding classes.

Out of the 1000-model ensemble of structures generated by Rosetta FlexPepDock for each of the 1271 peptides docked to the target H-2D^b, the 50 models with the lowest Rosetta energies were selected for calculation of the C_α RMSF at each of the nine residue positions. The per-residue C_α RMSFs show an increasing trend with increasing experimental dissociation constant at each position (Fig. 3 A). In contrast, the side-chain RMSFs showed a weaker correlation than the C_α RMSFs with binding affinities and were not included in the predictive criteria (Fig. S1 E). We note that these trends do not depend on the number of models selected, as the trends were the same when the top 200 models were selected (data not shown). The ensemble of structures generated is not comprehensive in the statistical mechanical sense, since the positions of receptor atoms are fixed during refinement and no Boltzmann averaging of energies was performed. Using a subset of the ensemble of structures generated is much more computationally efficient.

The average per-residue C_α RMSF values in the very strongly binding class of peptides ($k_D \leq 100$ nM) in Fig. 3 A were selected as thresholds to distinguish between binders and non-binders (Table II). For every peptide in the four ranges, we counted the number of residues for which the thresholds specified in Table II were met. Fig. 3 B shows the fraction of peptides that are predicted to bind within each of the four classes when any 7, 8 or 9 peptide residues satisfy the thresholds. Clearly, greater proportions of peptides within the very strong ($k_D \leq 100$ nM) and strong binding ($100 \text{ nM} < k_D < 1000$ nM) classes meet a majority of the cutoffs than do the peptides in the other three classes. For the set of 1,271 peptides, application of a very stringent rule for selecting binding peptides, i.e., that the RMSF threshold must be satisfied for all 9 residues, makes it 10 times more likely that a peptide predicted to bind strongly indeed belongs to the class with experimental $k_D \leq 100$ nM than to the class with $k_D > 6,500$ nM. Thus, meeting the thresholds for all 9 residue positions was selected as the criterion for a peptide binding strongly to the MHC for further analysis and comparison.

Combining sequence and structure-based approaches to eliminate false positives and increase vaccine specificity

We used the sequence-based NetMHCpan 4.0 method to predict the qualitative binding affinities of the peptides in the dataset to H-2D^b. We defined “binders” as those peptides flagged as “SB” (strong binders) by the default settings of NetMHCpan 4.0, while all others were considered “non-binders” including those that were flagged as “WB” (weak binders).

We note that the NetMHCpan 4.0 classification is not the same as the four-range classification we used above.

Prediction of binding for the 1,271 peptides to H-2D^b using NetMHCpan 4.0 revealed high statistical sensitivity, correctly predicting as “binders” 98% of the peptides in our defined class of experimentally very strongly binding peptides ($k_D \leq 100$ nM). However, for some alleles a limitation of NetMHCpan4.0 is low statistical specificity, i.e., a tendency to overpredict binding strengths of weak and non-binding peptides. Indeed, for H-2D^b 45% of the peptides in the experimental weak-binding range ($1,000 \text{ nM} < k_D \leq 6,500 \text{ nM}$) and 12% of the peptides in the experimental non-binding ($6500 \text{ nM} < k_D \leq 43,500$) ranges were predicted to be binders (Fig. 4 A, Fig. S1 F). Use of this approach to select vaccine candidates might therefore be expected to lead to a relatively high likelihood that weak binders are included, which would be unproductive. There is, therefore, significant opportunity for improving the specificity of the predictions (39). Because only a small minority of the many possible random nonapeptides are strong binders to MHC, a specificity improvement should translate into an improvement in positive predictive value (i.e., the fraction or percentage of true positives among the predicted positives).

A major aim in the design of T cell response-inducing vaccines is to select a collection of peptide antigens that each has a high likelihood of binding very strongly to MHC as all T cell epitopes are good MHC binders (67). Ideally, therefore, for a p-MHC binding prediction approach to be most useful in antigen identification and vaccine development, it should have in the statistical sense both high sensitivity ($TP/(TP + FN)$), where TP = true positives and FN = false negatives, and high specificity ($TN/(TN + FP)$), where TN = true negatives and FP = false positives (39). We found that the sequential use of NetMHCpan 4.0 followed by our structure-based approach removes most of the false positives found using NetMHCpan4.0 alone. In this approach, peptides are first screened using NetMHCpan 4.0 and those predicted to bind are then filtered through structure-based modeling, in which the per-residue RMSF of the ensemble of top-scoring docked poses of peptides is compared to pre-determined RMSF thresholds. Peptides that meet the 9-residue RMSF thresholds defined above are defined as “binders”. Screening using NetMHCpan 4.0 and then applying the per-residue RMSF threshold criterion from Table II resulted in 98 peptides that are predicted to bind out of the 1,271 tested. This approach resulted in the numbers of false positives predicted to be binders that are actually in the experimentally weak ($1,000 \text{ nM} < k_D \leq 6,500 \text{ nM}$) and non-binding ranges ($k_D > 6,500 \text{ nM}$) being reduced from 45% and 12% with NetMHCpan to 6% and 0.5%, respectively (Fig. 4 A, Fig. S1 F). Since the number of true weak binders and non-binders is expected to be much greater than the number of true strongly binding peptides, these improvements in specificity should translate into improvements in positive predictive value ($TP/(TP + FP)$), which is of paramount importance in vaccine design.

We then compared the statistical performance of the combined method with NetMHCpan4.0 for a binary classification of peptides into binders and non-binders (Fig. 4 B). For the sake of comparison between the sequence-based method NetMHCpan and the combined approach, true labels of the peptides as binders or non-binders were determined based on a 500 nM experimental dissociation constant cutoff. The main drawback of applying the structure-

based filter is a high miss rate, i.e., low sensitivity, which corresponds in practice to a considerable number of true binders being left out. However, applying the filter leads to greater specificity and overall to a higher positive predictive value, increasing the likelihood that peptides picked as binders actually have high binding affinity. Indeed, the application of the combined method reduces the average dissociation constant of predicted binders by about 700 nM (Table III). Diagnostic tests evaluating the performance of the sequence-based, structure-based and combined methods are provided in the Supporting Information (Fig. S1 G, Table S2).

Fig. 4 C shows the probability of a predicted binder falling into any of the four dissociation constant ranges for the predictions made by NetMHCpan 4.0 and the combined method. In the ideal scenario, for a binary classification biased to pick only the strongest binding peptides, the probability would be 1 in the first class ($k_D \leq 100$ nM) and 0 in the next three classes. When only peptides that have $k_D \leq 100$ nM are to be selected for inclusion in a vaccine, the positive predictive value increases from 40% with NetMHCpan 4.0 alone to 52% with the combined method ($p = 0.027$, chi-squared test). At the same time, the probability of finding a false positive that belongs in the non-binding range ($6,500$ nM $< k_D \leq 43,500$ nM) with the combined method is 2% compared to 6% with the sequence-based method alone. The combined sequence and structure-based p-MHC binding prediction workflow and the required software packages are schematized in the visual abstract.

Application of the combined sequence and structure-based method to human MHC I alleles with varying amount of training data and alleles that lack a template crystal structure

To assess the broader applicability of our combined sequence and structure-based p:MHC binding prediction approach that was applied as proof of concept to the murine H-2D^b allele, we selected five human MHC alleles (HLA-A*02-01, HLA-B*44-03, HLA-B*27-05, HLA-A*23-01, HLA-A*32-01) that represent varying levels of modeling and prediction difficulty. We modeled straightforward cases for which a template crystal structure and a large training dataset of very strongly binding peptides were available as well as challenging cases for which no crystal structure template and only a few very strongly binding ($k_D \leq 100$ nM) peptide sequences were available for training the structure-based model. In addition to nonameric peptides, we also included some cases with decameric peptides to test the structure-based approach on peptides of different lengths. Furthermore, curated datasets of experimentally verified immunogenic and non-immunogenic HLA-A*02-01-specific peptides were also subjected to the structure-based approach to understand to what extent the low sensitivity of the structure-based method can affect the selection of immunogenic epitopes. A description of the training and validation sets along with the templates that were used for the various cases tested are provided in the Methods section.

The Ca RMSF thresholds obtained for the structure-based model for different alleles are given in Table IV. For the balanced held-out test sets, similar to the result for the murine H-2D^b allele, the structure-based method on its own selects a greater percentage of peptides in the very strongly binding class as binders compared to peptides in any other binding class, especially the weakly and non-binding classes (Fig. 5 A–F). This outcome is consistent not

only for straightforward cases where the template crystal structure and sufficient training data are available, as in the case of HLA-A*02-01-nonameric complexes, but also for the more challenging cases where no template crystal structure exists, such as the HLA-A*23-01-decameric complexes where as few as 19 very strongly binding sequences were used for training. For all cases tested with the human allele, the pan-allele sequence-based method, NetMHCpan exhibits high sensitivity and outperforms the structure-based method, i.e., it picks almost all peptides from the very strong binding class and very few from the non-binding class. However, we cannot overlook the fact that the peptides tested here were a part of the training data for NetMHCpan but were non-overlapping with the training set for the structural-based method. Moreover, there is an average 8% gain in the positive predictive value (i.e., the probability that a predicted binder will in fact bind with $K_D \leq 100$ nM) for the combined approach, i.e., when peptides are first filtered using the sequence-based approach and then the structure-based approach is applied to the filtered peptides (Fig. 6). Further, an advantage of the structure-based method is that, regardless of whether a large training data is available, it can be applied independently by simply classifying the binding capability of the peptides based on their Ca RMSF ranking.

Testing on datasets of peptides with known immunogenicity

We also applied the combined sequence and structure-based approach to predicting the binding of 485 peptides of known immunogenicity in the curated datasets from Bjerregaard et al. (68) and Koşaloğlu-Yalçın et al. (69). In addition, we evaluated whether by effectively selecting the highest affinity MHC binders by the combined approach we could increase the likelihood over the standalone sequence-based approach of picking an actual immunogenic peptide for inclusion in a vaccine. We also assessed the effect of the relatively high miss rate of the combined approach in filtering out known epitopes and the reproducibility of the structure-based and combined sequence and structure-based method in selecting the highest affinity binders for the two datasets tested. The Bjerregaard dataset (PMID: 29187854) contains human immunogenic and non-immunogenic neopeptides (peptides derived from somatic mutations) originating from single amino acid substitutions for which T cell reactivity has been experimentally tested. The Koşaloğlu-Yalçın dataset (PMID: 30377561) contains known, naturally processed, immunogenic neoepitopes and random peptides. From these datasets, we selected HLA-A*02-01 allele-restricted peptides because the data for this allele were most abundant and thus relevant for statistical validation of our approach. The peptides in these datasets have experimentally validated T cell reactivity but quantitative HLA binding affinities, which is what NetMHCpan predicts, are not available. However, owing to the abundance of binding observations for HLA-A*02-01, NetMHCpan has excellent performance for this particular allele (AUC > 90%) and thus NetMHCpan predicted binding affinities can be used in lieu of experimental binding affinities. An overview of the distribution of the peptides in the different binding ranges as predicted by NetMHCpan is provided in Table V and a description of both datasets is provided in the Methods section.

Bjerregaard et al. curated dataset (PMID: 29187854)

We modeled the dataset of 290 HLA-A*02-01-nonamer complexes curated by Bjerregaard et al. (PMID: 29187854) using the protocol described in the Methods section. For each of

the 290 peptide sequences that were docked to HLA-A*02-01 we extracted the corresponding top-ranked 50-model ensemble, calculated the per-residue C α RMSF at each of the nine residue positions. To predict binding abilities, we then compared it to the C α RMSF thresholds established from the training with HLA-A*02-01.

Again, consistent with our proof-of-principle case, the standalone structure-based method selects a greater percentage of very strong binders than peptides in any other class. Using the criteria of meeting seven of nine RMSF thresholds, the method has a sensitivity of 41% (99/239) for HLA-A*02-01-restricted neopeptides in the very strong binding class (Fig. 7 A), which is consistent with the sensitivity observed for very strong binding peptides in the test sets for HLA-A*02-01 (Fig. 5 A). Although the total number of T cell epitopes contributed to only ~5% of the peptides in the dataset, six of the seventeen experimentally verified T cell epitopes in the Bjerregaard dataset were predicted as binders by the combined sequence and structure-based approach. Therefore, while the combined sequence and structure-based approach is not specifically designed to select T cell epitopes it does select them with a sensitivity that is similar to the very strong MHC binders (6/17 = 35%).

Among other criteria, such as the difference in binding affinities between mutated and non-mutated counterparts, immunogenic epitope candidates are regularly selected from the pool of peptides that have the highest binding affinities. In such a scenario, the fraction of epitopes retained only by the sequence-based method (15/239 = 6.27%) is about the same as that retained by the combined sequence and structure-based method (6/99 = 6.06%) for this dataset. Thus, the odds of picking an immunogenic peptide for this dataset pertaining to HLA-A*02-01 remained unchanged if the standalone sequence-based or the combined sequence and structure-based approach was used.

Koalo lu-Yalçın et al. curated dataset (PMID: 30377561)

From the distribution of epitopes in the various binding classes, we again found that T cell epitopes are predominantly strong MHC binders as 90% of the cancer epitopes were predicted by NetMHCpan to belong to very strong binding class (k_D 100 nM) and a majority of all T cell epitopes in this dataset (78%) were predicted by NetMHCpan to bind at least strongly ($1 \leq k_D \leq 1000$ nM). Similar to the sensitivity observed in selecting very strong binders in the test datasets of the HLA-A*02-01 allele and in the Bjerregaard dataset, ~35-45% of the mutated cancer epitopes, non-mutated cancer epitopes, and viral epitopes within the very strongly binding class predicted by NetMHCpan were selected as binders by the standalone structure-based method whereas peptides in the weak binding class were rarely selected as binders (Fig. 7 B–E). This finding shows that the standalone structure-based method has reproducible sensitivity and specificity in selecting very strong binders while filtering out weak and non-binders. The most striking evidence of this reproducibility is that, even though only 2/117 peptides in the random peptides dataset were predicted to be very strong binders by NetMHCpan, the structure-based methods picked 50% of the peptides, i.e., 1 of the 2 peptides in the very strong binding class to bind while it picked < 3% of the 103 random peptides that had NetMHCpan-predicted binding affinities of $k_D > 6500$ nM.

With the above examples, we have demonstrated through our combined sequence and structure-based approach that the very strong binders, irrespective of their immunogenicity, will be selected with a ~40% hit rate while a non-binder will be picked with a hit rate < ~3-5%. Since a T-cell epitope is more likely to be a very strongly binding peptide and since the PPV of the combined sequence and structure-based method of a picked peptide being a very strong binder is significantly higher than when only the sequence or the structure-based method alone is used, the chance of a predicted binder being an actual T cell epitope may potentially be higher when the combined approach is used.

Discussion

Structural modeling and computer simulation of p-MHC complexes can in principle determine accurate binding affinities while also providing first-principles information on binding energy contributions. However, these calculations are often computationally expensive and have not been shown to be more accurate than sequence-based AI algorithms, which therefore have been of more practical use. However, sequence-based methods have drawbacks. Perhaps the most serious of these is that the accuracy of binding affinity predictions is largely dependent on the size and diversity of the dataset used for training, i.e., the dataset requires binding affinity measurements within all affinity ranges for a large number of diverse peptides for the specific MHC allele considered (70). AI models that are trained on such large datasets have relatively high accuracy. However, some alleles have insufficient numbers of binding affinity measurements to enable training of the model. Moreover, while binding affinity measurements for nonameric peptides are abundant for most alleles the measurements for 8- to 15-mer peptides are not uniformly available, making training on these non-canonical lengths more challenging. The above factors can affect the accuracy of predictions (71). Furthermore, to conclusively establish the success of predictive models, testing the models for their predictions on large independent datasets would be required (37). However, most machine learning-based tools report cross-validation accuracies rather than accuracies on held-out datasets, and almost all observations in the IEDB dataset are potentially included in the training data for these tools (32, 37, 70, 72). Also, for some MHC alleles these methods yield a relatively large fraction of false positives.

By analyzing ensembles of conformations generated by the flexible docking of nonapeptides to the mouse MHC H-2D^b the computational approach explored here allows efficient differentiation between binding and non-binding peptides. Bound peptide conformations generated by FlexPepDock for high-affinity peptides are on average more similar to each other (and similar to the conformation of the bound peptides in crystal structures) than for peptides with low binding affinities. The Rosetta REF2015 energies of complexes involving low-affinity binders as determined by docking were similar to that of tight-binding peptides (Fig. S2), and thus not diagnostic of high-affinity binding. From a thermodynamic perspective, additional conformational flexibility would be expected to make an entropic contribution in favor of binding. However, based on what is observed in crystal structures and in this work, we hypothesize that the MHC protein uses its conformational free energy to enhance the binding of peptides that bind in a particular conformation, and that such complexes are thus primed to bind with higher affinity in the ultimate ternary complex with the T cell receptor. The observed low conformational flexibility of peptides in high-affinity

pMHC complexes is analogous to the case of DNA polymerization in which the DNA polymerase holds the deoxynucleotide triphosphate substrate in the conformation required for Watson-Crick base-pairing, increasing the free energy discrimination between Watson-Crick and non-Watson-Crick base pairs. Exclusion of water from the polymerase active site further increases the discrimination against incorrect base pairing by increasing the free energy gained from hydrogen bonding (73).

As a proof of concept, we first tested the structure-based approach on 1,271 peptide sequences for which the experimental binding affinities to H-2D^b are known. For this dataset, the results indicate that using the structural modeling approach in a standalone manner eliminates almost all the false positives in the weakest and non-binding range that are selected to bind by the sequence-based method while also significantly increasing the probability that a predicted high-affinity binder actually binds in the desired 1-100 nM range. When combined with a sequence-based method, the positive predictive value of the method increases from 40% (sequence-based method alone) to 52% ($p = 0.027$). The average dissociation constant, K_D , of the peptides selected as binders decreases significantly as well. Similarly, in the second stage of testing, when the combined sequence and structure-based approach was tested on human alleles we obtained an average ~8% increase in PPV over the standalone sequence-based method. The test set we used has equal numbers of strong, weak, and non-binders. In practice, the numbers of weak or non-binding nonapeptides are far greater, thus potentially further enhancing the improvement in positive predictive value available by improving specificity compared to what we have observed in this study.

In contrast to machine-learning approaches, the structure-based method introduced here is not limited by the availability of large datasets (e.g., our training dataset for the structure-based method consisted of only 40, 19 and 16 peptide sequences for HLA-B*44-03 with 9-mers, HLA-B*27-05 with 10-mers, and HLA-A*23-01 with 10-mers, respectively) and can be applied to peptides of any profile length and to any MHC allele, by modeling peptides that are known to bind very strongly to that allele and then determining the per-residue RMSF thresholds that discriminate peptides that bind from those that do not. If a p-MHC crystal structure exists it can be used as the template or, for alleles lacking crystal structures, homology models can be generated from the closest sequence homolog for which a structure has been determined as has been shown here in test cases for HLA-A*23-01 and HLA-A*32-01. A characteristic of the approach described here is that, we estimate that ~100 peptide sequences with very high binding affinity for the particular allele should be known to allow for training of the structure-based model. However, in the absence of any prior experimental binding affinity data, even that condition should be able to be overcome simply by ranking the binding strength of the peptide sequences based on their per-residue RMSFs obtained following docking. This strategy is important because it alleviates the problem of predicting binding affinities for peptides to MHC alleles for which few binding data exist, such that machine learning based on sequence-binding relationships would be impossible.

For alleles for which sequence-based predictive tools exist, the effective combination of sequence-based screening followed by structural modeling, optimization and comparison with predetermined thresholds of very strong binders has clear potential for rapidly and

precisely distinguishing between high-affinity and low-affinity binders. The principal drawback of the proposed combined method is a high miss rate (i.e., a large number of false negatives). In this regard we note that alternative methods of RMSF calculation that are also quick and perhaps reduce the miss rate could be used. Also, the thresholds in Table II and IV might be further optimizable.

Notwithstanding, for T cell-based vaccine design the combined method in its present form would be expected to be useful when applied to tumor or viral genomes for which machine learning (e.g., NetMHCpan) predicts a relatively large number of binding antigens. Filtering using structural modeling may then significantly reduce the number of predicted binding antigens, but the remaining collection is likely to both have higher specificity (i.e., $TN/(TN + FP)$) and positive predictive value (i.e., $TP/(TP + FP)$). The resulting vaccine candidates might therefore be expected to be more likely to contain a higher proportion of very strong binders ($K_D < 100$ nM) and possibly actual T cell epitopes than when using only the sequence-based method.

We have demonstrated by testing on human and murine MHC alleles that the positive predictive value (PPV) for predicting tight binders improves with the combined sequence and structure-based approach in comparison to the sequence-based method alone. Although we expect this approach to lead to an enrichment of the immunogenic epitopes amongst predicted binders, this outcome could not be demonstrated for existing datasets of documented immunogenic epitopes. The PPV for predicting immunogenicity was about the same as the sequence-based method, i.e., the fraction of immunogenic epitopes predicted to be binders to the total number of predicted binders by both the combined approach and by standalone sequence-based NetMHCpan was ~6%. However, we note that the experimental data needed for definitive testing of the enrichment of immunogenic epitopes in a set of predicted binders are currently unavailable. For this effort, immunogenicity data would be needed for alleles other than HLA-A*-02-01 allele, which is the best-performing allele for the sequence-based method. In other words, data would be needed for alleles for which sequence-based methods perform less well. Further, the Bjerregaard and Yalcin datasets are unbalanced, and the combined method would need to be tested on a dataset that also includes non-binders. When the required experimental immunogenic data become available the combined method will be able to be tested more thoroughly. It may become possible to improve the fraction of known epitopes in the set of predicted binders by training the algorithm against immunogenicity directly or by also including the ternary complex with TCR.

Sequence-based methods are popular in part owing to the large number of peptides for which peptide-MHC binding affinity predictions can be made within a matter of minutes. However, the present structure-based method is also quick. Docking and producing 1000 models of each peptide to the MHC with Rosetta FlexPepDock takes roughly 90 minutes on a single node 2.33-GHz Intel Xeon 32-core processor. Thus, roughly 30,000 tumor or viral genomes could be screened in one day on a ~100 petaflop supercomputer such as Summit at Oak Ridge National Laboratory. Therefore, if whole exome sequencing of virus or tumors can be similarly rapidly performed, it may be possible to execute the present combined method, or an improved derivative of it, on a massive scale.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was sponsored by the Laboratory Directed Research and Development Program at Oak Ridge National Laboratory (ORNL), which is managed by UT-Battelle, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC05-00OR22725. This work used resources of the Compute and Data Environment for Science (CADES) at ORNL and of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility. LMW was supported by NIH NCI grants R01-CA50633 and NCI P30-CA51008.

References

1. Gell PGH, and Benacerraf B. 1959 —II.: Studies on Hypersensitivity Delayed Hypersensitivity to Denatured Proteins in Guinea Pigs. *Immunology* 2: 64. [PubMed: 13640681]
2. Vyas JM, Van der Veen AG, and Ploegh HL. 2008 The known unknowns of antigen processing and presentation. *Nature Reviews Immunology* 8: 607.
3. Hewitt EW 2003 The MHC class I antigen presentation pathway: strategies for viral immune evasion. *Immunology* 110: 163–169. [PubMed: 14511229]
4. Zinkernagel R, and Doherty P. 1974 Restriction of in vitro T cell-mediated cytotoxicity in lymphocytic choriomeningitis within a syngeneic or semiallogeneic system.
5. Terasaki PI 2007 A brief history of HLA. *Immunologic research* 38: 139–148. [PubMed: 17917021]
6. Klein L, Kyewski B, Allen PM, and Hogquist KA. 2014 Positive and negative selection of the T cell repertoire: what thymocytes see (and don't see). *Nature Reviews Immunology* 14: 377.
7. Yewdell JW 2006 Confronting complexity: real-world immunodominance in antiviral CD8+ T cell responses. *Immunity* 25: 533–543. [PubMed: 17046682]
8. Lafuente EM, and Reche PA. 2009 Prediction of MHC-peptide binding: a systematic and comprehensive overview. *Current pharmaceutical design* 15: 3209–3220. [PubMed: 19860671]
9. Jensen PE 2007 Recent advances in antigen processing and presentation. *Nature Immunology* 8: 1041. [PubMed: 17878914]
10. Sette A, Vitiello A, Rehman B, Fowler P, Nayarsina R, Kast WM, Melief CJM, Oseroff C, Yuan L, Ruppert J, Sidney J, Del Guercio MF, Southwood S, Kubo RT, Chesnut RW, Grey HM, and Chisari FV. 1994 The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes. *Journal of Immunology* 153: 5586–5592.
11. Rensing ME, Sette A, Brandt RMP, Ruppert J, Wentworth PA, Hartman M, Oseroff C, Grey HM, Melief CJM, and Kast WM. 1995 Human CTL epitopes encoded by human papillomavirus type 16 E6 and E7 identified through in vivo and in vitro immunogenicity studies of HLA-A*0201-binding peptides. *Journal of Immunology* 154: 5934–5943.
12. Chen W, Khilko S, Fecondo J, Margulies DH, and Mc Cluskey J. 1994 Determinant selection of major histocompatibility complex class I-restricted antigenic peptides is explained by class I-peptide affinity and is strongly influenced by nondominant anchor residues. *Journal of Experimental Medicine* 180: 1471–1483. [PubMed: 7523572]
13. Sette A, Sidney J, del Guercio MF, Southwood S, Ruppert J, Dahlberg C, Grey HM, and Kubo RT. 1994 Peptide binding to the most frequent HLA-A class I alleles measured by quantitative molecular binding assays. *Molecular Immunology* 31: 813–822. [PubMed: 8047072]
14. Engels B, Engelhard Victor H., Sidney J, Sette A, Binder David C., Liu Rebecca B., Kranz David M., Meredith Stephen C., Rowley Donald A., and Schreiber H. 2013 Relapse or Eradication of Cancer Is Predicted by Peptide-Major Histocompatibility Complex Affinity. *Cancer Cell* 23: 516–526. [PubMed: 23597565]
15. Kammertoens T, and Blankenstein T. 2013 It's the peptide-MHC affinity, stupid. *Cancer Cell* 23: 429–431. [PubMed: 23597560]

16. Reche PA, Keskin DB, Hussey RE, Ancuta P, Gabuzda D, and Reinherz EL. 2006 Elicitation from virus-naïve individuals of cytotoxic T lymphocytes directed against conserved HIV-1 epitopes. *Medical Immunology* 5: 1. [PubMed: 16674822]
17. Comber JD, and Philip R. 2014 MHC class I antigen presentation and implications for developing a new generation of therapeutic vaccines. *Therapeutic advances in vaccines* 2: 77–89. [PubMed: 24790732]
18. Vitiello A, and Zanetti M. 2017 Neoantigen prediction and the need for validation. *Nature biotechnology* 35: 815.
19. Karasaki T, Nagayama K, Kuwano H, Nitadori J. i., Sato M, Anraku M, Hosoi A, Matsushita H, Takazawa M, and Ohara O. 2017 Prediction and prioritization of neoantigens: integration of RNA sequencing data with whole-exome sequencing. *Cancer science* 108: 170–177. [PubMed: 27960040]
20. Hundal J, Carreno BM, Petti AA, Linette GP, Griffith OL, Mardis ER, and Griffith M. 2016 pVAC-Seq: A genome-guided in silico approach to identifying tumor neoantigens. *Genome Medicine* 8: 11. [PubMed: 26825632]
21. Bjerregaard A-M, Nielsen M, Hadrup SR, Szallasi Z, and Eklund AC. 2017 MuPeXI: prediction of neo-epitopes from tumor sequencing data. *Cancer Immunology, Immunotherapy* 66: 1123–1130. [PubMed: 28429069]
22. Matsushita H, Vesely MD, Koboldt DC, Rickert CG, Uppaluri R, Magrini VJ, Arthur CD, White JM, Chen Y-S, and Shea LK. 2012 Cancer exome analysis reveals a T-cell-dependent mechanism of cancer immunoediting. *Nature* 482: 400. [PubMed: 22318521]
23. Rizvi NA, Hellmann MD, Snyder A, Kvistborg P, Makarov V, Havel JJ, Lee W, Yuan J, Wong P, Ho TS, Miller ML, Rekhtman N, Moreira AL, Ibrahim F, Bruggeman C, Gasmi B, Zappasodi R, Maeda Y, Sander C, Garon EB, Merghoub T, Wolchok JD, Schumacher TN, and Chan TA. 2015 Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* 348: 124. [PubMed: 25765070]
24. McGranahan N, Furness AJS, Rosenthal R, Ramskov S, Lyngaa R, Saini SK, Jamal-Hanjani M, Wilson GA, Birkbak NJ, Hiley CT, Watkins TBK, Shafi S, Murugaesu N, Mitter R, Akarca AU, Linares J, Marafioti T, Henry JY, Van Allen EM, Miao D, Schilling B, Schadendorf D, Garraway LA, Makarov V, Rizvi NA, Snyder A, Hellmann MD, Merghoub T, Wolchok JD, Shukla SA, Wu CJ, Peggs KS, Chan TA, Hadrup SR, Quezada SA, and Swanton C. 2016 Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science* 351: 1463. [PubMed: 26940869]
25. Snyder A, Makarov V, Merghoub T, Yuan J, Zaretsky JM, Desrichard A, Walsh LA, Postow MA, Wong P, Ho TS, Hollmann TJ, Bruggeman C, Kannan K, Li Y, Elipenahli C, Liu C, Harbison CT, Wang L, Ribas A, Wolchok JD, and Chan TA. 2014 Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma. *New England Journal of Medicine* 371: 2189–2199. [PubMed: 25409260]
26. Hugo W, Zaretsky JM, Sun L, Song C, Moreno BH, Hu-Lieskovan S, Berent-Maoz B, Pang J, Chmielowski B, Cherry G, Seja E, Lomeli S, Kong X, Kelley MC, Sosman JA, Johnson DB, Ribas A, and Lo RS. 2016 Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. *Cell* 165: 35–44. [PubMed: 26997480]
27. Ott PA, Hu Z, Keskin DB, Shukla SA, Sun J, Bozym DJ, Zhang W, Luoma A, Giobbie-Hurder A, Peter L, Chen C, Olive O, Carter TA, Li S, Lieb DJ, Eisenhaure T, Gjini E, Stevens J, Lane WJ, Javeri I, Nelliappan K, Salazar AM, Daley H, Seaman M, Buchbinder EI, Yoon CH, Harden M, Lennon N, Gabriel S, Rodig SJ, Barouch DH, Aster JC, Getz G, Wucherpfennig K, Neuberger D, Ritz J, Lander ES, Fritsch EF, Hacohen N, and Wu CJ. 2017 An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature* 547: 217–221. [PubMed: 28678778]
28. Lundegaard C, Lamberth K, Harndahl M, Buus S, Lund O, and Nielsen M. 2008 NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8–11. *Nucleic acids research* 36: W509–W512. [PubMed: 18463140]
29. Andreatta M, and Nielsen M. 2015 Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics* 32: 511–517. [PubMed: 26515819]

30. Nielsen M, Lundegaard C, Blicher T, Lamberth K, Harndahl M, Justesen S, Røder G, Peters B, Sette A, and Lund O. 2007 NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and-B locus protein of known sequence. *PloS one* 2: e796. [PubMed: 17726526]
31. Nielsen M, and Andreatta M. 2016 NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. *Genome medicine* 8: 33. [PubMed: 27029192]
32. O'Donnell TJ, Rubinsteyn A, Bonsack M, Riemer AB, Laserson U, and Hammerbacher J. 2018 MHCflurry: Open-Source Class I MHC Binding Affinity Prediction. *Cell Systems* 7: 129–132.e124. [PubMed: 29960884]
33. Vita R, Overton JA, Greenbaum JA, Ponomarenko J, Clark JD, Cantrell JR, Wheeler DK, Gabbard JL, Hix D, Sette A, and Peters B. 2015 The immune epitope database (IEDB) 3.0. *Nucleic acids research* 43: D405–D412. [PubMed: 25300482]
34. Rammensee HG, Bachmann J, Emmerich NPN, Bachor OA, and Stevanovi S. 1999 SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics* 50: 213–219. [PubMed: 10602881]
35. Schuler MM, Nastke M-D, and Stevanovi S. 2007 SYFPEITHI In *Immunoinformatics*. Springer 75–93.
36. Sette A, Peters B, Metushi IG, Greenbaum JA, Sidney J, Nielsen M, Lund O, Trolle T, and Kim Y. 2015 Automated benchmarking of peptide-MHC class I binding predictions. *Bioinformatics* 31: 2174–2181. [PubMed: 25717196]
37. Zhao W, and Sher X. 2018 Systematically benchmarking peptide-MHC binding predictors: From synthetic to naturally processed epitopes. *PLOS Computational Biology* 14: e1006457. [PubMed: 30408041]
38. Paul S, Weiskopf D, Angelo MA, Sidney J, Peters B, and Sette A. 2013 HLA class I alleles are associated with peptide-binding repertoires of different size, affinity, and immunogenicity. *The Journal of Immunology* 191: 5831–5839. [PubMed: 24190657]
39. van Buuren MM, Calis JJA, and Schumacher TNM. 2014 High sensitivity of cancer exome-based CD8 T cell neo-antigen identification. *OncoImmunology* 3: e28836. [PubMed: 25083320]
40. Bonsack M, Hoppe S, Winter J, Tichy D, Zeller C, Küpper MD, Schitter EC, Blatnik R, and Riemer AB. 2019 Performance Evaluation of MHC Class-I Binding Prediction Tools Based on an Experimentally Validated MHC–Peptide Binding Data Set. *Cancer Immunology Research* 7: 719–736. [PubMed: 30902818]
41. Trolle T, Metushi IG, Greenbaum JA, Kim Y, Sidney J, Lund O, Sette A, Peters B, and Nielsen M. 2015 Automated benchmarking of peptide-MHC class I binding predictions. *Bioinformatics* 31: 2174–2181. [PubMed: 25717196]
42. Yanover C, and Bradley P. 2011 Large-scale characterization of peptide-MHC binding landscapes with structural simulations. *Proceedings of the National Academy of Sciences* 108: 6981.
43. Doytchinova IA, and Flower DR. 2001 Toward the quantitative prediction of T-cell epitopes: CoMFA and CoMSIA studies of peptides with affinity for the class I MHC molecule HLA-A*0201. *Journal of Medicinal Chemistry* 44: 3572–3581. [PubMed: 11606121]
44. Wan S, Knapp B, Wright DW, Deane CM, and Coveney PV. 2015 Rapid, Precise, and Reproducible Prediction of Peptide–MHC Binding Affinities from Molecular Dynamics That Correlate Well with Experiment. *Journal of Chemical Theory and Computation* 11: 3346–3356. [PubMed: 26575768]
45. Schueler-Furman O, Altuvia Y, Sette A, and Margalit H. 2000 Structure-based prediction of binding peptides to MHC class I molecules: application to a broad range of MHC alleles. *Protein science : a publication of the Protein Society* 9: 1838–1846. [PubMed: 11045629]
46. Altuvia Y, and Margalit H. 2004 A structure-based approach for prediction of MHC-binding peptides. *Methods* 34: 454–459. [PubMed: 15542371]
47. Ochoa R, Laio A, and Cossio P. 2019 Predicting the Affinity of Peptides to Major Histocompatibility Complex Class II by Scoring Molecular Dynamics Simulations. *Journal of Chemical Information and Modeling* 59: 3464–3473. [PubMed: 31290667]
48. Duan F, Duitama J, Al Seesi S, Ayres CM, Corcelli SA, Pawashe AP, Blanchard T, McMahon D, Sidney J, and Sette A. 2014 Genomic and bioinformatic profiling of mutational neoepitopes

reveals new rules to predict anticancer immunogenicity. *Journal of Experimental Medicine* 211: 2231–2248. [PubMed: 25245761]

49. Park M-S, Park SY, Miller KR, Collins EJ, and Lee HY. 2013 Accurate structure prediction of peptide–MHC complexes for identifying highly immunogenic antigens. *Molecular Immunology* 56: 81–90. [PubMed: 23688437]
50. Schneider M, Rosam M, Glaser M, Patronov A, Shah H, Back KC, Daake MA, Buchner J, and Antes I. 2016 BiPPred: Combined sequence-and structure-based prediction of peptide binding to the Hsp70 chaperone BiP. *Proteins: Structure, Function, and Bioinformatics* 84: 1390–1407.
51. Antes I, Siu SW, and Lengauer T. 2006 DynaPred: a structure and sequence based method for the prediction of MHC class I binding peptide sequences and conformations. *Bioinformatics* 22: e16–e24. [PubMed: 16873467]
52. Aranha MP, Spooner C, Demerdash O, Czejdo B, Smith JC, and J. C. J. B. e. B. A.-G. S. Mitchell. 2020 Prediction of peptide binding to MHC using machine learning with sequence and structure-based feature sets. 1864: 129535.
53. Eswar N, Webb B, Marti-Renom MA, Madhusudhan M, Eramian D, Shen M. y. Pieper U, and Sali A. 2006 Comparative protein structure modeling using Modeller. *Current protocols in bioinformatics* 15: 5.6. 1–5.6. 30.
54. Martí-Renom MA, Stuart AC, Fiser A, Sánchez R, Melo F, and Šali A. 2000 Comparative protein structure modeling of genes and genomes. *Annual review of biophysics and biomolecular structure* 29: 291–325.
55. Šali A, and Blundell TL. 1993 Comparative protein modelling by satisfaction of spatial restraints. *Journal of molecular biology* 234: 779–815. [PubMed: 8254673]
56. Shen M. y., and Sali A. 2006 Statistical potential for assessment and prediction of protein structures. *Protein Science* 15: 2507–2524. [PubMed: 17075131]
57. Raveh B, London N, and Schueler-Furman O. 2010 Sub-angstrom modeling of complexes between flexible peptides and globular proteins. *Proteins: Structure, Function, and Bioinformatics* 78: 2029–2040.
58. Kaufmann KW, Lemmon GH, DeLuca SL, Sheehan JH, and Meiler J. 2010 Practically useful: what the Rosetta protein modeling suite can do for you. *Biochemistry* 49: 2987–2998. [PubMed: 20235548]
59. Raveh B, London N, Zimmerman L, and Schueler-Furman O. 2011 Rosetta FlexPepDock ab-initio: simultaneous folding, docking and refinement of peptides onto their receptors. *PloS one* 6: e18934. [PubMed: 21572516]
60. Liu T, Pan X, Chao L, Tan W, Qu S, Yang L, Wang B, and Mei H. 2014 Subangstrom accuracy in pHLA-I modeling by Rosetta FlexPepDock refinement protocol. *Journal of chemical information and modeling* 54: 2233–2242. [PubMed: 25050981]
61. Alam N, and Schueler-Furman O. 2017 Modeling peptide-protein structure and binding using Monte Carlo sampling approaches: Rosetta FlexPepDock and FlexPepBind. In *Modeling Peptide-Protein Interactions*. Springer 139–169.
62. Alford RF, Leaver-Fay A, Jeliaskov JR, O'Meara MJ, DiMaio FP, Park H, Shapovalov MV, Renfrew PD, Mulligan VK, and Kappel K. 2017 The Rosetta all-atom energy function for macromolecular modeling and design. *Journal of chemical theory and computation* 13: 3031–3048. [PubMed: 28430426]
63. London N, Raveh B, Movshovitz-Attias D, and Schueler-Furman O. 2010 Can self-inhibitory peptides be derived from the interfaces of globular protein-protein interactions? *Proteins* 78: 3140–3149. [PubMed: 20607702]
64. London N, Lamphear CL, Hougland JL, Fierke CA, and Schueler-Furman O. 2011 Identification of a novel class of farnesylation targets by structure-based modeling of binding specificity. *PLoS computational biology* 7: e1002170. [PubMed: 21998565]
65. Alam N, Zimmerman L, Wolfson NA, Joseph CG, Fierke CA, and Schueler-Furman O. 2016. Structure-Based Identification of HDAC8 Non-histone Substrates. *Structure (London, England : 1993)* 24: 458–468.

66. London N, Raveh B, Cohen E, Fathi G, and Schueler-Furman O. 2011 Rosetta FlexPepDock web server—high resolution modeling of peptide–protein interactions. *Nucleic acids research* 39: W249–W253. [PubMed: 21622962]
67. Patronov A, and Doytchinova I. 2013 T-cell epitope vaccine design by immunoinformatics. *Open Biol* 3: 120139–120139. [PubMed: 23303307]
68. Bjerregaard AM, Nielsen M, Jurtz V, Barra CM, Hadrup SR, Szallasi Z, and Eklund AC. 2017 An Analysis of Natural T Cell Responses to Predicted Tumor Neoepitopes. *Front Immunol* 8: 1566. [PubMed: 29187854]
69. Ko alo lu-Yalçın Z, Lanka M, Frentzen A, Logandha Ramamoorthy Premalal A, Sidney J, Vaughan K, Greenbaum J, Robbins P, Gartner J, Sette A, and Peters B. 2018 Predicting T cell recognition of MHC class I restricted neoepitopes. *OncoImmunology* 7: e1492508. [PubMed: 30377561]
70. Kim Y, Sidney J, Buus S, Sette A, Nielsen M, and Peters B. 2014 Dataset size and composition impact the reliability of performance benchmarks for peptide-MHC binding predictions. *BMC Bioinformatics* 15: 241–241. [PubMed: 25017736]
71. Luo H, Ye H, Ng HW, Shi L, Tong W, Mendrick DL, and Hong H. 2015 Machine Learning Methods for Predicting HLA-Peptide Binding Activity. *Bioinformatics and Biology Insights* 9s3: BBI.S29466.
72. Trolle T, Metushi IG, Greenbaum JA, Kim Y, Sidney J, Lund O, Sette A, Peters B, and Nielsen M. 2015 Automated benchmarking of peptide-MHC class I binding predictions. *Bioinformatics* 31: 2174–2181. [PubMed: 25717196]
73. Beckman RA, and Loeb LA. 1993 Multi-stage proofreading in DNA replication. *Quarterly reviews of biophysics* 26: 225–331. [PubMed: 8022969]

Key points

- Computational docking and conformational analyses was used to predict peptide:MHC binding.
- Structure-based method was applicable even for alleles with limited training data.
- Combining sequence and structure-based methods improved p:MHC binding prediction PPV.

Figure 1.

Superposition of X-ray crystal structures of (A) 39 nonameric peptide-H-2D^b complexes and (B) peptides only. Side chains are shown in B and anchor residues 3, 5, and 9 are labeled. Colors: MHC, blue; β -2 microglobulin, gray; peptides, orange. PDB entry 1JPG was chosen as the reference for alignment and RMSD calculation as it had the highest resolution (2.2 Å) of the 39 available crystal structures. Comparison of peptide conformations in (C) top-ranked FlexPepDock models of 50 strong binding complexes, and (D) top-ranked FlexPepDock models of 50 non-binding complexes. The MHC is not shown. (E) Average

C_α RMSF distributions for the modeled ensembles of very strong and non-binding 9-mer peptides. Values were determined after removing overall roto-translational motion by best-fit superposition of the peptide conformations in the 50-model ensemble for each individual peptide to the top-ranked model of that ensemble.

Figure 2.

Position-specific C_α RMSF for the set of peptide poses in each of the four binding classes and in the set of crystal structures after superposition of the docked p-MHC complex onto the reference crystal structure (PDB entry 1JPG). Each of the four sets consists of the top ranked pose of every peptide in that binding class. (A) C_α RMSF at each peptide residue position for peptides in each binding affinity class and in crystal structures. (B) Difference between the mean C_α RMSF in each class and the mean C_α RMSF in the class of very strong binders ($k_D < 100$ nM).

Figure 3.

(A) Average C α RMSF at each residue position for each binding affinity class. The nomenclature used for the peptide residues consists of the letter P followed by a number indicating its position in the sequence going from the N- to the C-terminus. Prior to the calculation of the C α RMSF, the peptide poses of the top 50 models for each peptide sequence were superposed onto the peptide pose in the top-ranked model. Error bars represent a 95% confidence interval. Dotted lines connect the average C α RMSFs in each class. The dashed gray box encloses the thresholds selected. (B) Percentage of peptides within each of the experimental dissociation constant ranges that are predicted to bind with RMSF thresholds satisfied at 7, 8 and 9 residue positions.

Figure 4.

(A) Number of peptides predicted to be binders that are found in each of the four experimental dissociation constant ranges. Sequence-based = predicted to bind by NetMHCpan4.0. Structure-based = meeting the thresholds of Table II in the present modeling. Also shown is the sequential application of the sequence- and structure-based methods. (B) Performance of the sequence-based NetMHCpan4.0 and the combined method (NetMHCpan4.0 plus the present structure-based approach). TP, TN, FP, FN determined based on a 500-nM dissociation constant cutoff. Sensitivity = (TP/(TP+FN)), specificity = (TN/(TN+FP)), and the positive predictive value (PPV) = (TP/(TP+FP)). The improvement in PPV is statistically significant at $p = 0.027$, chi-squared test. PPV depends not only on

sensitivity and specificity, but the fractions of true negatives and true positives in the underlying population. Our test dataset with equal numbers of true negative and true positives may underestimate the improvement in PPV compared to realistic dataset with many true negatives. (C) Fraction of binders that fall into each of the four ranges of experimental dissociation constants as predicted by NetMHCpan 4.0 alone and by combining NetMHCpan 4.0 with the structure-based approach. The dotted box indicates the positive predictive value (PPV) for distinguishing binders from non-binders, assuming an experimental dissociation constant cutoff of $K_D < 100$ nM. Cumulative probabilities are shown as dashed lines.

Figure 5.

Performance of the standalone structure-based approach for different human alleles in terms of percentage of peptides predicted as binders in each of the four experimental dissociation constant ranges.

Figure 6.

(A-F) Fraction of predicted binders that fall into each of the four ranges of experimental dissociation constants as predicted by NetMHCpan 4.0 alone and by combining NetMHCpan 4.0 with the structure-based approach. The dotted box indicates the positive predictive value (PPV) for distinguishing binders from non-binders, assuming an experimental dissociation constant cutoff of $k_D < 100$ nM. (G) The difference in the positive predictive value (PPV) at threshold ($k_D < 100$ nM) obtained from the standalone sequence-

based NetMHCpan (PPV_sequence), and from the combined sequence and structure-based approach (PPV_combined).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 7.

Percentage of peptides that are predicted as binders by the structure-based method alone in the different NetMHCpan predicted dissociation constant ranges for (A) the Bjerregaard et al. curated dataset and for (B) the Koşaloğlu-Yalçın et al. curated mutated cancer epitopes dataset, (C) non-mutated cancer epitopes dataset (D) Viral epitopes dataset and (E) National cancer institute random peptides dataset.

Table I:

Data on X-ray crystal structures and the corresponding docked p-MHC complexes along with experimental and predicted dissociation constants (k_D). Experimental dissociation constants are from the Immune Epitope Database. The predicted dissociation constants and qualitative predictions obtained using NetMHCpan 4.0, which classifies peptides as Strong binders (SB), Weak binders (WB)), or Non-binders (NB), are listed.

Peptide sequence	ID	Experimental k_D (nM)	NetMHCpan4.0 predicted k_D (nM)	NetMHCpan4.0 qualitative prediction	Overall RMSD (Å) ^a	Peptide RMSD (Å) ^b	Peptide Cα RMSD (Å) ^c
FAPGNYPAL	1ce6	5.4	43.7	SB	1.92	2.05	1.44
AAVYNFATM	1ffo		7.6	SB	0.64	1.75	0.91
SAVYNFATM	1ffp		6.3	SB	0.82	1.29	1.02
SSVVGVMWYL	1inq	539.0	416.8	SB	0.57	2.78	1.28
FQPQNGQFI	1jpg	5.4	239.1	SB	0.82	1.94	1.23
SSVIGVMWYL	1juf	1400.0	726.6	SB	0.49	2.83	1.06
FAPSNYPAL	1qlf		15.6	SB	2.04	2.23	1.81
KAVYNFATM	1s7u	11.5	9.8	SB	1.12	2.13	1.02
KAVYNLATM	1s7v	22.2	11.1	SB	2.01	2.75	1.63
KALYNFATM	1s7w		7.1	SB	1.12	2.35	1.18
KAVFNFATM	1s7x		9.6	SB	1.12	1.63	0.69
KALYNYAPI	1zhh		5.4	SB	1.73	2.88	1.91
LSLRNPILV	3buy	4.4	119.8	SB	1.19	2.29	1.85
KVPRNQDWL	3cc5	58.4	154.9	SB	0.78	2.86	2.20
EGPRNQDWL	3ch1	454.9	1423.9	SB	1.53	3.17	2.44
ASNENAETM	3cpl		12.4	SB	0.73	1.73	0.84
ASAENMETM	3ftg		9.0	SB	1.72	1.53	1.13
KAVANFATM	3quk	1.0	5.3	SB	1.40	1.82	0.80
KAVSNFATM	3qul		6.1	SB	1.14	1.98	0.70
KAPANFATM	3tbs		5.6	SB	1.32	1.72	0.95
KAPSNFATM	3tgt		6.6	SB	0.98	3.26	2.68
KGPANFATM	3tbv		16.7	SB	1.10	1.55	0.82
KGPSNFATM	3tbw		21.7	SB	1.24	2.31	0.95
KAPFNFATM	3tby		9.6	SB	1.78	1.95	1.10
YQLENYCGL	3ws3		93.9	SB	1.82	2.20	1.93
YAIENYLEL	3ws6		4.0	SB	1.12	2.01	1.27
ASNENIETM	4huu		9.9	SB	1.01	1.76	0.89
ASNENWETM	4huv		9.5	SB	0.75	1.52	1.19
ASNENTETM	4huw		13.8	SB	0.65	1.42	0.84
ASNENMETM	4hux	16.8	9.5	SB	0.64	1.99	0.79
ASNEHMETM	4l8b		304.2	SB	0.45	1.83	0.83
ASDENMETM	4l8c		66.5	SB	0.83	1.69	1.10
ASNEDMETM	4l8d		626.8	SB	0.91	3.17	2.08
MCLRMTAVM	5e8n		2989	WB	1.45	2.68	1.53
WLVTNLSYL	5jwd	154.2	226.1	SB	0.77	2.58	1.15

Peptide sequence	ID	Experimental k_D (nM)	NetMHCpan4.0 predicted k_D (nM)	NetMHCpan4.0 qualitative prediction	Overall RMSD (Å) ^a	Peptide RMSD (Å) ^b	Peptide C α RMSD (Å) ^c
SQLLNAKYL	5wli		29.3	SB	1.61	1.82	1.11
KAPYNFATM	5til		10.2	SB		template	
Average					1.15 (0.45)^c	2.15 (0.54)^d	1.3 (0.51)^e

^aRMSD of p-MHC complex of the top-ranked docking model from the corresponding crystal structure after global superposition.

^bMSD of peptide in top-ranked docking model from peptide in the crystal structure after global superposition.

^cRMSD of peptide C α atoms in top-ranked docking model from peptide C α atoms in the crystal structure after global superposition.

^{d, e, f} standard deviations are shown in parentheses.

Table II.

Position-specific C_α RMSF thresholds for binary classification of binders and non-binders to the H-2D^b allele. The peptide residue numbering is from the N- to C-terminus.

Peptide residue	C _α RMSF threshold (Å)
P1	0.90
P2	0.50
P3	0.60
P4	0.60
P5	0.44
P6	0.43
P7	0.38
P8	0.34
P9	0.50

Table III.

Average dissociation constant of peptides predicted as strong binders by sequence-based, structure-based, and combined methods. The prevalence of binders in the dataset is 0.39.

Method	Number of peptides predicted as binders	Average experimental dissociation constant (K _D) of predicted binders (nM)
NetMHCpan 4.0	731	1,707
Structure-based (this work)	107	1,958
Combined (NetMHCpan 4.0 + Structure-based)	98	1,051

Table IV:

The C α RMSF thresholds obtained for the structure-based model for different human MHC I alleles

Peptide residue position	C α RMSF thresholds calculated (Å)					
	HLA-A*02-01 with 9 mers	HLA-B*44-03 with 9 mers	HLA-B*27-05 with 10 mers	HLA-A*23-01 with 9 mers	HLA-A*23-01 with 10 mers	HLA-A*32-01 with 9 mers
1	0.85	0.94	0.98	0.93	1.14	1.00
2	0.58	0.64	0.69	0.62	0.77	0.67
3	0.67	0.74	0.77	0.68	0.87	0.65
4	0.71	0.88	0.95	0.81	0.92	0.64
5	0.91	0.88	0.98	0.79	0.91	0.69
6	0.91	0.74	0.76	0.72	0.78	0.83
7	0.79	0.75	0.86	0.67	0.87	0.88
8	0.69	0.70	0.74	0.61	0.86	0.81
9	0.92	0.87	0.83	0.86	0.88	0.97
10			1.03		1.18	

Table V:

NetMHCpan 4.0 predicted distribution of the Bjerregaard and Kosalogulu-Yalcin datasets's HLA-A*02-01 restricted peptides in different binding classes. The percentage of peptides in a given class for each individual dataset is in parenthesis.

NetMHCpan 4.0 predicted dissociation constant range	<u>Bjerregaard dataset</u>	<u>Kosalogulu-Yalcin Dataset</u>			
		Mutated cancer epitopes	Non-mutated cancer epitopes	Viral epitopes	Random peptides
Very Strong binders ($k_D < 100$ nM)	239 (82.4%)	18 (90.0%)	16 (40.0%)	11 (61.1%)	2 (1.7%)
Strong binders ($100 < k_D \leq 1000$ nM)	49 (16.9%)	1 (5.0%)	13 (32.5%)	2 (11.1%)	6 (5.1%)
Weak binders ($1000 < k_D \leq 6500$ nM)	2 (0.7%)	1 (5.0%)	6 (15.0%)	3 (16.7%)	6 (5.1%)
Non-binders ($k_D > 6500$ nM)	0 (0.0%)	0 (0.0%)	5 (12.5%)	2 (11.1%)	103 (88.0%)
		20 (100%)	40 (100%)	18 (100%)	117 (100%)
Total observations	290 (100%)		195 (100%)		