



## Prediction of peptide binding to MHC using machine learning with sequence and structure-based feature sets

Michelle P. Aranha<sup>a,b,1</sup>, Catherine Spooner<sup>c,1</sup>, Omar Demerdash<sup>b,d</sup>, Bogdan Czejdo<sup>c</sup>, Jeremy C. Smith<sup>a,b</sup>, Julie C. Mitchell<sup>d,\*</sup>

<sup>a</sup> Department of Biochemistry and Cellular and Molecular Biology, University of Tennessee, Knoxville, TN 37996, United States of America

<sup>b</sup> University of Tennessee/Oak Ridge National Laboratory Center for Molecular Biophysics, Oak Ridge National Laboratory, Oak Ridge, TN 37830, United States of America

<sup>c</sup> Department of Mathematics and Computer Science, Fayetteville State University, Fayetteville, NC 28301, United States of America

<sup>d</sup> Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, United States of America

### ARTICLE INFO

#### Keywords

MHC-peptide  
Binding affinity  
Machine learning

### ABSTRACT

Selecting peptides that bind strongly to the major histocompatibility complex (MHC) for inclusion in a vaccine has therapeutic potential for infections and tumors. Machine learning models trained on sequence data exist for peptide:MHC (p:MHC) binding predictions. Here, we train support vector machine classifier (SVMC) models on physicochemical sequence-based and structure-based descriptor sets to predict peptide binding to a well-studied model mouse MHC I allele, H-2D<sup>b</sup>. Recursive feature elimination and two-way forward feature selection were also performed. Although low on sensitivity compared to the current state-of-the-art algorithms, models based on physicochemical descriptor sets achieve specificity and precision comparable to the most popular sequence-based algorithms. The best-performing model is a hybrid descriptor set containing both sequence-based and structure-based descriptors. Interestingly, close to half of the physicochemical sequence-based descriptors remaining in the hybrid model were properties of the anchor positions, residues 5 and 9 in the peptide sequence. In contrast, residues flanking position 5 make little to no residue-specific contribution to the binding affinity prediction. The results suggest that machine-learned models incorporating both sequence-based descriptors and structural data may provide information on specific physicochemical properties determining binding affinities.

### 1. Introduction

In immunotherapy treatments, putative antigenic peptides or neoantigens are administered as a part of a vaccine. Some of these peptides are recognized by the T cell receptors on cytotoxic or helper T cells which in turn induces a cellular response against the invading pathogen or cancerous tumors. Such neoantigen-based vaccines have been found to induce a positive clinical response in melanoma patients [1,2]. Binding of antigenic peptides to MHC is a requirement for T cell receptor binding and efforts to develop increasingly accurate methods of identifying which peptides bind to MHC are underway [3,4]. Allele-specific and pan-specific sequence-based methods that use artificial neural network models trained on available datasets of peptides have proven hugely successful [3,4]. Various structure-based methods that train on structural characteristics of p-MHC also hold promise for improving prediction for cases where the small size of datasets makes training difficult [5–7].

The sequence-based methods utilize standard scoring matrices to encode the protein and peptide sequences and as such make interpretation of the physicochemical characteristics of binding difficult. To overcome this difficulty, the inclusion of interpretable physicochemical descriptor sets and structural models in training can help characterize the binding landscape. Several sequence-based physicochemical descriptor sets exist that have been previously used in characterizing or predicting the nature of protein-protein interactions [8]. In this study, we have tested the efficacy of these physicochemical descriptors in predicting p-MHC binding, using a well-characterized mouse MHC, H-2D<sup>b</sup>. Furthermore, predictive classification models based on structure-based descriptors that can be extracted from common docking protocols were also tested. Finally, we used feature selection strategies on the combined sequence and structure-based descriptor sets to build a model using the most optimal features.

\* Corresponding author.

E-mail address: [mitchelljc@ornl.gov](mailto:mitchelljc@ornl.gov) (J.C. Mitchell)

<sup>1</sup> Both authors contributed equally to this article.

## 2. Methods

### 2.1. Dataset and features

The experimentally known binding affinities of our dataset of 1278 peptide sequences to the mouse allele, H-2D<sup>b</sup>, were sourced from the immune epitope database (IEDB) website [9]. Based on bounds determined by previous peptide-MHC binding studies, a 500 nM cutoff in the experimental dissociation constant ( $k_D$ ) was selected as the threshold to define a binary classifier, i.e.,  $k_D \leq 500$  nM (binders) and  $k_D > 500$  nM (non-binders) [10]. Based on this threshold, the dataset contained 501 binders and 777 non-binders. The observations were randomized and divided into a training and an independent test set containing 80 and 20% of all observations respectively. The training dataset contained 398 positive and 624 negative binders, while the independent test set contained 103 positive and 153 negative binders. To address the issue of imbalance, during training, a penalized class-weighted SVM was used to improve the classification and to avoid the algorithm from focusing on the majority class. Finally, testing of the model on the independent test set allowed us to both evaluate the performance of the model and assess for any overfitting.

Our initial pool of features was based on both sequence and structure-based features. We describe the construction of this feature set below:

#### 2.1.1. Sequence-based features

The function of a peptide or protein depends on the physicochemical properties of its amino acid sequence such as hydrophobicity, alpha helix propensity, beta sheet propensity, bulkiness (ratio of side chain volume to length), charge and the frequency of occurrence in protein sequences. The “aaDescriptors” function of an R-based “Peptides” package contained eight physicochemical descriptor sets that were used to encode peptide sequences into numerical vectors [11]. It has recently been reported that the different descriptor sets indeed describe the AA space differently although there are commonalities in the way they are constructed [12]. A brief description of the descriptor sets used is provided below:

**BLOSUM indices** [13]: It represents more than 500 amino acid indices from the AAindex database by a set of uncorrelated scales. It is based on a VARIMAX analysis [14] of physicochemical properties which were subsequently converted to indices based on the BLOSUM62 substitution matrix.

**Factor Analysis Scales of Generalized Amino Acid Information (FAS-GAI)** [15]: Numerical representation of amino acids based on six factors associated with hydrophobicity (F1), alpha-helix and beta-turn propensities (F2), bulkiness (F3), compositional characteristics (F4), local flexibility (F5), and electronic properties (F6).

**Kidera Factors (KF)** [16]: These were originally derived by applying multivariate analysis to 188 physical properties of the 20 amino acids and using principal component analysis and factor analysis to reduce the dimensionality of the features. This function calculates the average of the ten Kidera factors for a protein sequence. KF1: Helix/bend preference, KF2: Side-chain size, KF3: Extended structure preference, KF4: Hydrophobicity, KF5: Double-bend preference, KF6: Partial specific volume, KF7: Flat extended preference, KF8: Occurrence in alpha region, KF9: pK value of carboxyl group, KF10: Surrounding hydrophobicity.

**MSWHIM** [17]: In this scale, 20 amino acids are represented by three components derived from principal component analysis of 3D electrostatic properties of residues. The first component discriminates between positive and negatively charged residues and between aromatic and bulkier aliphatic residues. The second component dis-

criminates Asp and Glu from other residues. The third component discriminates Arg and Lys from all other residues.

**Cruciani properties** [18]: This comprises a set of three principal properties (PP) describing the side-chains based on their polarity (PP1), size/hydrophobicity (PP2), and H-bonding capability (PP3).

**The ProtFP descriptor set** [12]: This descriptor set was constructed from a large initial selection of indices obtained from the AAindex database for all 20 naturally occurring amino acids.

**Vectors of Hydrophobic, Steric, and electronic properties (VHSE scales)** [19]: The VHSE was derived from principal component analysis (PCA) of 50 physicochemical variables of 20 coded amino acids that contained independent families of 18 hydrophobic properties, 17 steric properties, and 15 electronic properties, respectively. VHSE1 and VHSE2 relate to hydrophobic properties, VHSE3 and VHSE4 relate to steric properties, VHSE5 to VHSE8 relate to electronic properties.

**Z scales** [20]: Each Z scale represent an amino-acid property. Z1: hydrophobicity, Z2: Steric properties (Steric bulk/Polarizability), Z3: Electronic properties (Polarity / Charge), Z4 and Z5: They relate to electronegativity, heat of formation, electrophilicity, and hardness.

#### 2.1.2. Structure-based modeling and features

Initial 3D models were generated with the program MODELLER starting from an X-ray crystal structure (PDB ID: 5TIL) of H-2D<sup>b</sup> 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 [21]. These top-ranked peptide-MHC models were then refined with FlexPepDock [22,23] from the Rosetta modeling suite [24]. 1000 models per p-MHC complex were constructed with FlexPepDock and were ranked based on their Rosetta Ref2015 score [25]. Top 50 models of each p-MHC complex were selected for the calculation of structure-based features.

Two structure-based feature sets were constructed from docking-specific Rosetta energy terms [23] (hereafter, referred to as Rosetta features) and from in-house scoring function developed by Demerdash and Mitchell [26] (hereafter, referred to as DM features). Features from the DM structure-based descriptor sets are briefly described in the APPENDIX. The structure-based descriptors include energetic terms associated with electrostatics, dispersion, hydrogen bond, cation-pi interactions, statistical potential terms, distances and number of interface atoms and root mean square fluctuations of backbone and side-chain residues. Feature persistence in terms of the averages and standard deviations of the Rosetta component scores for each of the p-MHC conformational ensemble was also included in the feature set.

### 2.2. Package

The classification was implemented using the SVM library and SVC function in Scikit Learn.

### 2.3. Feature selection

Varying the soft margin constant, C allows one to adjust the penalty for misclassification, i.e., the higher the C value the greater is the penalty for misclassification and vice versa. To understand which features were consistently selected at low and high C values, we performed recursive feature elimination (RFE) with a linear kernel SVM at low C (1, 1.25, 1.5 and 1.75) and high C (2, 3, 5, 10, 25) values on a combined sequence-based descriptor set that contained vectors from all eight sequence-based descriptor sets and on the two structure-based descriptor sets. The features ranked 1 in both low and high C runs of the RFE were selected as features to be included

in the optimal feature set. The optimal feature set thus contained sequence and structure-based descriptors on which sequential two-way forward feature selection was performed to ascertain importance of the selected features [27]. A “randomized-search” on  $C$  and  $\gamma$  using cross-validation with a non-linear radial basis function kernel in SVMC gave the optimized hyper-parameters (give the optimized values for  $C$  and  $\gamma$ ) for forward feature selection. To ensure that the order of the features added during forward feature selection was consistent, we repeated the experiment five times.

#### 2.4. Performance measures

The performance of model was evaluated in terms of the sensitivity, specificity, accuracy and F1 scores and the area under the receiver-operator characteristic (ROC) curve (AUC) with leave-one-out fivefold cross-validation and also on the independent test set.

### 3. Results

#### 3.1. Benchmarking sequence-based descriptors for predicting p-MHC

We investigated the performance of SVM models generated using each of eight sequence-based descriptor sets and compared it to the current state-of-the-art, NetMHCpan 4.0, which for our dataset prediction achieved a sensitivity, specificity, and precision of 0.93, 0.66 and 0.64 respectively. We first built the models without recursive feature elimination by taking into account all the features in each descriptor set and next, by using features left after recursive feature elimination which was performed to reduce the feature space required to classify peptides correctly. The descriptor performance with all features and with features left after recursive feature elimination evaluated on the independent test set using six metrics is shown in Table 1. The total number of features encoding any nonameric peptide varies between 27 and 90 and the accuracy varies between 0.65 and 0.73 across different descriptor sets. The largest reduction of the descriptor set following recursive feature elimination was for the Kidera set (reduced from 90 descriptors to 22) and FASGAI vectors (reduced from 54 to 21) (Fig. 1). There is negligible loss in accuracy after recursive feature elimination, thus the reduced descriptor set can be used without loss of information. MSWHIM, which uses the fewest features in its descriptor set, also performs the worst on most metrics. VHSE gives the best overall AUC score (0.73) using all 72 features (8 features per residue) and also with the reduced feature set of 61 features. The best model would have to have high specificity with high sensitivity, as this model would lead to the fewest false positives while simultaneously detecting the majority of the binders. We found that after recursive feature elimination VHSE, BLOSUM and ProtFP which each have between 60 and 70 descriptors, have the highest specificities and sensitivities ( $\geq 0.70$ ) and models built using these descriptors sets would thus have the ability to correctly reject non-binders while retaining most of the binders. While quite low on sensitivity, the models achieve specificity and precision comparable with NetMHCpan 4.0.

#### 3.2. Generating the optimal hybrid sequence and structure-based feature space using recursive feature elimination

Next, we considered the two structure-based descriptor sets individually and combined the feature vectors from the different sequence-based descriptor sets. Using a linear SVM kernel, for each of these three descriptors sets we conducted a recursive feature elimination at different soft margin constant,  $C$  values. The number of features left after RFE for every  $C$  is shown in Table 2. We found 34, 9 and 8 features from the sequence-based, Rosetta and DM feature sets respectively persisted across the different  $C$  values. These persis-

tent features characterize the peptide binding affinities with the sparsest descriptor set and we use these 51 features as a hybrid descriptor set. Many of the sequence-based features or scales are mixtures of several physical properties; however, wherever possible a physical interpretation of the features selected after feature elimination is given in Table 3. The BLOSUMindices feature set contributes the highest number of features (10 features) followed by Kidera factors (7 features), VHSE (4 features) and Z-scales (4 features).

The performance of the 51-feature hybrid descriptor set along with the 34-feature sequence-based descriptor set, 9-feature Rosetta structure-based descriptor set, and 8-feature structure-based DM feature set is shown in Fig. 2 and Table 4. The hybrid feature space has the best overall accuracy, of 0.71, compared to the individual sequence and structure-based descriptor sets. The sequence-based descriptor set had high recall but low specificity, while the Rosetta features structure-based set had high specificity but quite low recall. Precision attained by the individual descriptor sets was low. However, the precision of the model based on the hybrid feature set was comparable to the precision achieved using the benchmark NetMHCpan 4.0.

Fig. 3 shows the number of times features at specific residue positions of the nonameric peptides are included in the final optimal subset of 51 sequence and structure-based features. Out of the 51 features, 20 features are specific to positions 5 and 9 which have previously been identified as anchor residues for binding peptides by studies focused on motif-based searches [28]. Features at positions 2 and 8 were also frequently chosen while features at positions 1,3 and 4 occurred less frequently. Features specific to residues at positions 6 and 7 do not make individual contributions to the model and are hence identified as those that make the least difference to the p-MHC binding prediction. Based on the sequence-based features selected at the anchors, the size, hydrophobicity, the secondary structure preference, and the electronic properties appeared very important. An analysis of the degree of sequence conservation for binders ( $k_D \leq 500$  nM) and non-binders ( $k_D > 500$  nM) using the WebLogo program [29] is shown in Figs. 4A and B. Sequence restrictions at residue positions 5 and 9 are observed along with some restrictions at positions 2 and 3 for the peptides binding to H-2D<sup>b</sup> (Fig. 4A). Peptides binding to H-2D<sup>b</sup> exhibit a preference for Asn at positions 5 and Ile, Leu and Met at position 9. A lower degree of sequence conservation at residue position, 5 and 9 is observed in non-binders as compared to binders (Fig. 4B). In terms of structure-based properties, the average solvent accessible area at residue 1, the root-mean-square fluctuations of all atoms and the side chains at residues 2 and 5 respectively were found to be important along with the hydrogen bond distance of the residue at 9 from the receptor.

Even though the 51 features characterize physicochemical properties, such as hydrophobicity, charge, size the features from the different descriptor sets appear largely uncorrelated (Fig. 5) i.e. the feature set is minimally redundant. Principal component analysis revealed that further reduction of the 51 features would not yield appreciable dimensionality reduction as it would take  $> 25$  components to capture 80% of the variance (Fig. 6). Selecting fewer dimensions would mean discarding a significant part of the variance. To completely describe the peptides, preserving this diversity of the features was required.

#### 3.3. Forward feature selection to ascertain feature importance

We sought to further rank the importance of the 51 features using forward feature selection. As the 2-way forward feature selection strategy is non-deterministic, the forward feature selection was run 5 times and the AUC obtained versus the number of cycles is shown in Fig. 7. In each cycle, a pair of features are appended to the base fea-

**Table 1**  
Performance of sequence-based descriptor sets evaluated on the independent test set.

Using all features								After recursive feature elimination						
Descriptor set	# of features	Accuracy	Precision	Sensitivity	Specificity	F-score	AUC	# of features	Accuracy	Precision	Sensitivity	Specificity	F-score	AUC
MSWHIM	27	0.64	0.54	0.73	0.58	0.62	0.65	8	0.62	0.52	0.68	0.58	0.59	0.63
Cruciani	27	0.64	0.54	0.69	0.61	0.61	0.65	20	0.64	0.54	0.65	0.63	0.59	0.64
Kidera	90	0.71	0.64	0.67	0.75	0.65	0.71	22	0.68	0.58	0.71	0.66	0.64	0.68
FASGAI	54	0.66	0.56	0.65	0.66	0.60	0.66	21	0.67	0.57	0.73	0.63	0.64	0.68
z scales	45	0.68	0.59	0.67	0.69	0.63	0.68	23	0.69	0.59	0.76	0.64	0.66	0.70
ProtFP	72	0.70	0.61	0.73	0.69	0.66	0.71	70	0.71	0.62	0.74	0.70	0.68	0.72
BLOSUM	90	0.72	0.65	0.66	0.76	0.66	0.71	67	0.73	0.65	0.69	0.75	0.67	0.72
VHSE	72	0.73	0.64	0.74	0.72	0.68	0.73	61	0.73	0.62	0.76	0.71	0.69	0.73

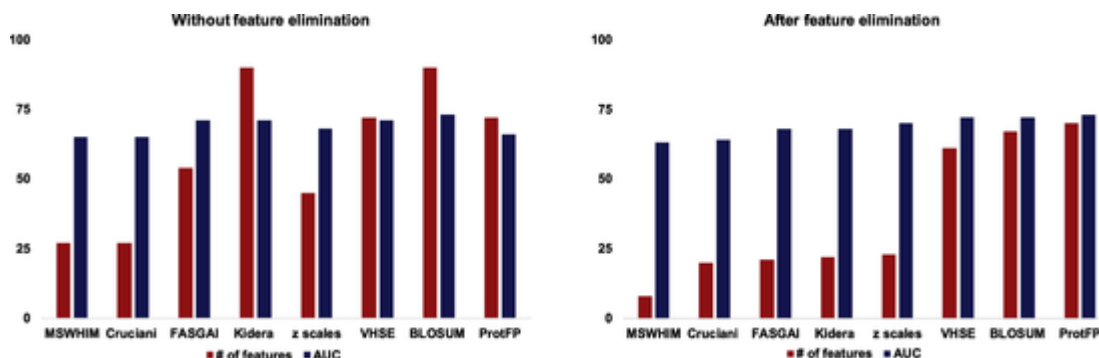


Fig. 1. Number of features and the performance in terms of AUC of the different sequence-based descriptor sets before and after recursive feature elimination.

Table 2

Number of features selected across different  $C$  values.

Feature set	Number of features selected for low and high $C$										Top ranked features across all $C$
	Low $C$					High $C$					
	1.00	1.25	1.50	1.75	2	3	5	10	25		
Sequence-based	154	113	118	134	238	89	97	121	85	34	
Structure-based Rosetta features	71	69	65	71	71	62	10	12	103	9	
Structure-based DM features	57	103	109	81	65	59	18	133	103	8	
Hybrid features (sequence & structure-based)	Total number of features									51	

ture pair. In all five experiments, the sequence-based features VHSE2.2 and BLOSUM9.5 were selected. The AUC score plateaued after 5–6 cycles i.e. after appending 10–12 features in all five runs. We collected the features that were appended in the first five cycles of each of the five experiments and analyzed the frequency of occurrence of these features (Fig. 8A). This resulted in a collection of a total of 23 features of which three features – BLOSUM9.5, PC and VHSE2.2 were appended within the first five cycles consistently across all five experiments. Features that persisted in most of the five experiments in the first five cycles were considered to be more important than the others.

In the same manner, the frequency with which features appear in the first 10 cycles in all 5 experiments is shown in Fig. 8B. This resulted in 45 features. We examined the influence of the recurring top-ranked feature on the model performance. To do so we first developed a model based on features that recur the greatest number of times and each time we appended the next set of features that appear the greatest number of times. This was continued until the entire feature set was consumed.

Table 5 presents the performance of the models built using features that recur a different number of times in total in the five experimental runs within the first five or ten cycles of the forward feature selection strategy. For instance, VHSE2.2, BLOSUM9.5, and PC appear within the first five cycles in all five runs of the forward feature selection and the model derived using just these three features has a test set AUC of 0.64 while its sensitivity and specificity is 0.76 and 0.52 respectively. Going from a 3-feature to a 23-feature model gives only a modest gain in AUC of 6%. To obtain the best performance we concluded it was important to retain all 51 features of the hybrid feature set.

#### 4. Conclusion

In this study, we examine physicochemical and structural descriptors that influence peptide binding to the mouse MHC I allele, H-2D<sup>b</sup>. Additionally, we provide a benchmarking of the commonly

available sequence-based physicochemical descriptors for a mouse MHC, H-2D<sup>b</sup> using 1278 peptides. We found that descriptor sets ProtFP, VHSE and BLOSUM scales were more accurate than the other descriptor sets tested.

Recursive feature elimination identified 17 structural features as contributing most towards the predictive model. Rosetta interface energy scores and peptide energy scores from the Rosetta feature set, that were previously identified by FlexPepBind [30–33] as important structural descriptors in binding prediction, were also identified by our model. Also, geometrical features such as RMSFs among models of the ensemble, along with the solvent accessible surface area and the hydrogen bond distance at the anchor sites were selected to be included in the final feature set.

The specificity and precision of the binding predictions obtained from models using these three descriptor sets were on par with the state-of-the-art sequence-based tool, NetMHCpan, suggesting that the method might be useful for filtering samples with a large number of potential antigens. In contrast, the method appears to perform worse than NetMHCpan in terms of sensitivity.

Of particular interest is the finding that the best-performing model is trained using majority of the features describing anchor positions, with close to half of the descriptors describing properties of the anchor positions, residues 5 and 9 in the sequence, consistent with the role of these residues in anchoring the interaction. The electronic properties or the charge of anchor residues at position 5 and the hydrophobicity of anchor site 9 and auxiliary anchor 2 were identified as important determinants of binding. Also, of interest is the finding that, even though residues at positions 4 and 6 flank the anchor residue at position 5, they make little to no residue-specific contribution to the binding affinity prediction of the model. The results therefore suggest that machine-learned models incorporating both sequence-based descriptors and structural data might be logically interpretable and provide specific physicochemical properties determining binding affinities. Linking these results to simulation-de-

**Table 3**

Hybrid feature space of sequence and structural descriptors. Structure-based features from Rosetta and DM descriptor sets are in bold and in italics respectively.

Residue position	Feature	Interpretation
1	<i>AvgPolASA_Pep1</i>	Accessible surface area of polar atoms of residue
2	BLOSUM8.2 BLOSUM9.2	$\alpha$ -helix propensity Frequency of appearance of amino acid in proteins (composition), $\alpha$ -helix propensity, bulkiness
	KF10.2	Surrounding hydrophobicity (Kidera Factors)
	VHSE2.2	Hydrophobic properties
	ProtFP7.2 <i>AvgAllAtomRSA_Pep2</i>	All atom RMSD of residue
3	BLOSUM4.3 BLOSUM8.3 KF7.3	Bulkiness, charge $\alpha$ -helix propensity Flat extended preference
4	BLOSUM8.4	$\alpha$ -helix propensity
5	BLOSUM5.5  BLOSUM9.5	Charge, $\alpha$ -helix propensity Frequency of appearance of amino acid in proteins (composition), $\alpha$ -helix propensity, bulkiness,
	F6.5	Electronic properties/ charge
	KF10.5	Surrounding hydrophobicity (Kidera Factors)
	KF3.5	$\beta$ structure preference related
	MSWHIM1.5	Discriminates between positive and negative charged residues and between aromatic and bulkier aliphatic
	ProtFP5.5	Obtained from PCA, physico-chemical relevance not specified
	VHSE6.5 VHSE8.5 Z3.5	Electronic properties Electronic properties Electronic properties (Polarity / Charge)
	Z5.5	electronegativity, heat of formation, electrophilicity and hardness
	<i>AvgSideChRSA_Pep5</i> BLOSUM8.8 F1.8	RMSD of side chain $\alpha$ -helix propensity Hydrophobicity index (FASGAI)
8		

**Table 3 (Continued)**

Residue position	Feature	Interpretation
	KF6.8	Partial specific volume
	MSWHIM1.8	Discriminates between positive and negative charged residues and between aromatic and bulkier aliphatic
	ProtFP5.8	Obtained from PCA, physico-chemical relevance not specified
9	BLOSUM2.9 BLOSUM6.9	Bulkiness Frequency of appearance of amino acid in proteins (composition)
	KF1.9	Helix/bend preference (Kidera Factors)
	KF5.9	Double-bend preference
	PP1.9	Polarity (CrucianiProperties)
	VHSE1.9	Hydrophobic properties
	Z2.9	Steric properties (Steric bulk/ Polarizability)
	Z3.9	Electronic properties (Polarity / Charge)
	ProtFP8.9 <i>AvgHbondDistPep9</i>	Distance of donor-acceptor
	<b>Ave_I_hb</b>	Ensemble average of number of hydrogen bonds across the interface
	<b>Ave_I_sc</b>	Ensemble average of Interface energy
	<b>Ave_frac_iatoms_less_1A</b>	Fraction of interface atoms that are <1 Å from binding partner
	<b>Ave_frac_iatoms_less_1A_bb</b>	Fraction of backbone interface atoms that are <1 Å from binding partner
	<b>Ave_frac_iatoms_less_2A</b>	Fraction of interface atoms that are <2 Å from binding partner
	<b>Sdev_I_hb</b>	Standard deviation of the number of hydrogen bonds across the interface in the ensemble
	<b>Sdev_fa_dun</b>	Internal energy of sidechain rotamers as derived from Dunbrack's statistics.

Table 3 (Continued)

Residue position	Feature	Interpretation
	Sdev_pep_sc_noref	Standard deviation of the peptide score (sum over energy contributed by the peptide to the total score; consists of the the internal peptide energy and the interface energy) without the amino acid dependent energy terms (Eaa) that were optimized to generate designs with natural amino acid content in the ensemble
	Sdev_rmsSC_CAPRI_if	Standard deviation of the RMSD between output model and the native structure, over peptide and receptor interface residues, where interface is defined similarly to rms (ALL/BB/CA)_if in the ensemble
	EL_sternberg	Coulombic electrostatic w/ Sternberg's distance-dependent dielectric
	PC	Heuristic pi-cation
	Pamino	Morse potential-based pi-cation/pi-amine
	StDevMainChASA_ABPeP	Standard deviation in accessible surface area.

rived work would be an interesting future activity, as would extension of the studies to other common MHC alleles.

#### Uncited reference

[43]

#### Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

#### Acknowledgements

This work was supported by LDRD funding from the US Department of Energy.

#### Appendix A

##### A.1. DM structure-based features

For each peptide-MHC complex, an ensemble of structures was generated using the protein structural modeling suite Rosetta [34], specifically, the flexible peptide docking module FlexPepDock [35]. For each set of complexes, the average and standard deviation of a set of features was calculated, including energetics, hydrogen bonding, solvent-accessible surface area (SASA), and the atomic Cartesian coordinates (root mean-squared deviation; RMSD). The non-energetic terms were calculated for the total complex, the protein chains individually, and for the individual residues in the peptide. Energetic terms were calculated for the total complex only. There were 492 features in total.

**Energetics.** Twenty-seven energetic terms that model electrostatics, van der Waals, desolvation, hydrogen bonding, shape complementarity, interface surface area, and aromatic interactions were calculated using in-house code [26,36]. These are described as follows.

**van der Waals energy.** Van der Waals energy was calculated using a softened 12–6 Lennard-Jones potential. In this regime, the repulsive and attractive contributions to the energy are denoted as  $L_J^{rep}$  and  $L_J^{attr}$ , respectively. The softening was performed as follows. First, the term  $Q$  is defined as follows:

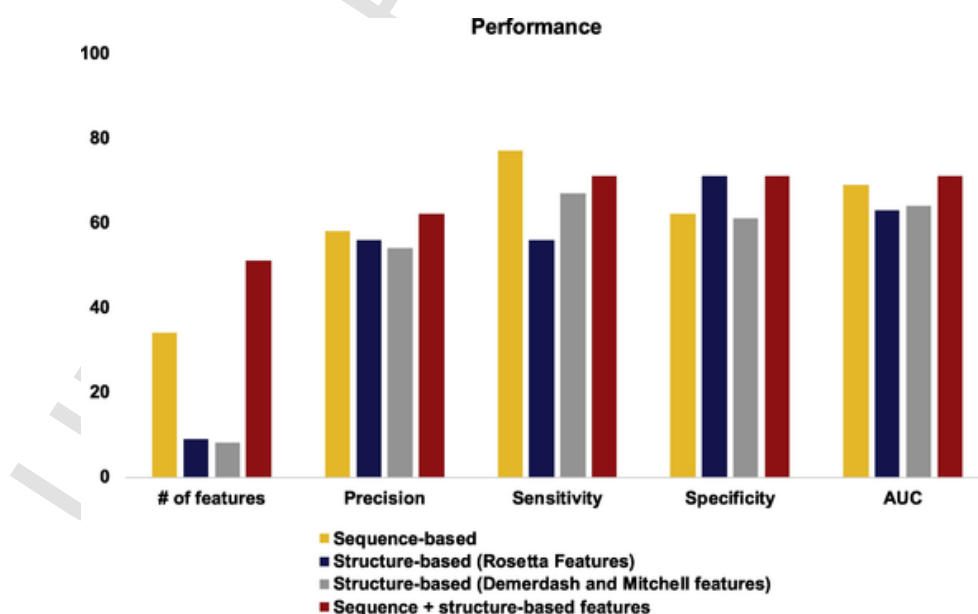


Fig. 2. Performance of sequence, structure and combined feature set models on independent test set.

**Table 4**  
Comparison of the performance of models based on the persistent features selected after recursive feature elimination using different  $C$  values.

Type of model	Training Set CV LOO performance				Test Set Performance							
	Accuracy	Precision	Recall	Specificity	F-score	AUC	Accuracy	Precision	Recall	Specificity	F-score	AUC
Sequence-based feature set (34 features)	0.72	0.61	0.79	0.67	0.69	0.73	0.68	0.58	0.77	0.62	0.66	0.69
Structure-based feature set (8 Rosetta features)	0.69	0.59	0.66	0.71	0.62	0.68	0.65	0.56	0.56	0.71	0.56	0.63
Structure-based feature set (9 DM features)	0.67	0.55	0.75	0.61	0.64	0.68	0.64	0.54	0.67	0.61	0.60	0.64
Hybrid feature set (51 features)	0.73	0.63	0.77	0.71	0.69	0.74	0.71	0.62	0.71	0.71	0.66	0.71

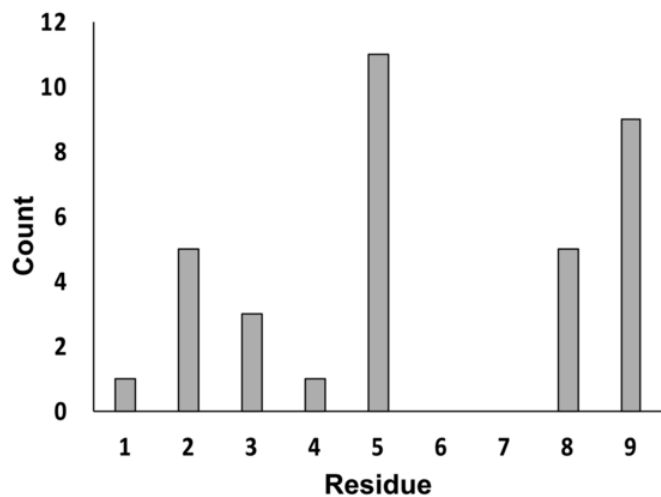


Fig. 3. Number of features encoding each residue position in the hybrid feature space.

$$Q = \epsilon_{ij} \left[ \left( \frac{r_0}{r_{ij}} \right)^{12} - 2 \left( \frac{r_0}{r_{ij}} \right)^6 \right] \quad (1)$$

where  $r_{ij}$  is the interatomic distance between an atom in MHC (receptor) and an atom in the neoantigen peptide (ligand),  $r_0$  is the equilibrium distance, and  $\epsilon_{ij}$  is the well depth. When  $Q < 1.0$ , the energy is purely attractive:

$$\begin{aligned} L_{J^{attr}} &= Q \\ L_{J^{rep}} &= 0 \end{aligned} \quad (2)$$

when  $Q \geq 1.0$  and  $r_{ij} \geq r_0/2$ , a modest penalty is incurred in the attractive term:

$$C_{tot} = \sum_i C_i = \sum_i (\lambda_{iR} - \lambda_0) \cdot (\lambda_{iL} - \lambda_0) \quad (3)$$

Finally, for significant collisions there is a large penalty in the repulsive term and a smaller penalty in the attractive term as follows:

$$\begin{aligned} L_{J^{attr}} &= \epsilon_{ij} \\ L_{J^{rep}} &= \epsilon_{ij} \left[ \frac{1}{64^2} \left( \frac{r_0}{r_{ij}} \right)^{12} - \frac{1}{32} \left( \frac{r_0}{r_{ij}} \right)^6 \right] \end{aligned} \quad (4)$$

**Knowledge-based potential.** The Boltzmann statistics-based, pairwise, atom contact based potential of Zhang et al. was used.

**Interface size.** Interface size was obtained as a proxy for desolvation and was calculated simply by counting the number of protein residues in the interface.

**Shape complementarity.** Shape complementarity at the protein ligand interface was calculated using the following function [37]:

$$C_{tot} = \sum_i C_i = \sum_i (\lambda_{iR} - \lambda_0) \cdot (\lambda_{iL} - \lambda_0) \quad (5)$$

In the above,  $C_{tot}$  is the total shape complementarity at the interface,  $C_i$  is the shape complementarity calculated at the  $i$ th interface atom of the receptor, and  $\lambda_{iR}$  and  $\lambda_{iL}$  are fractal density dimensions calculated at the  $i$ th interface atom of the receptor with respect to

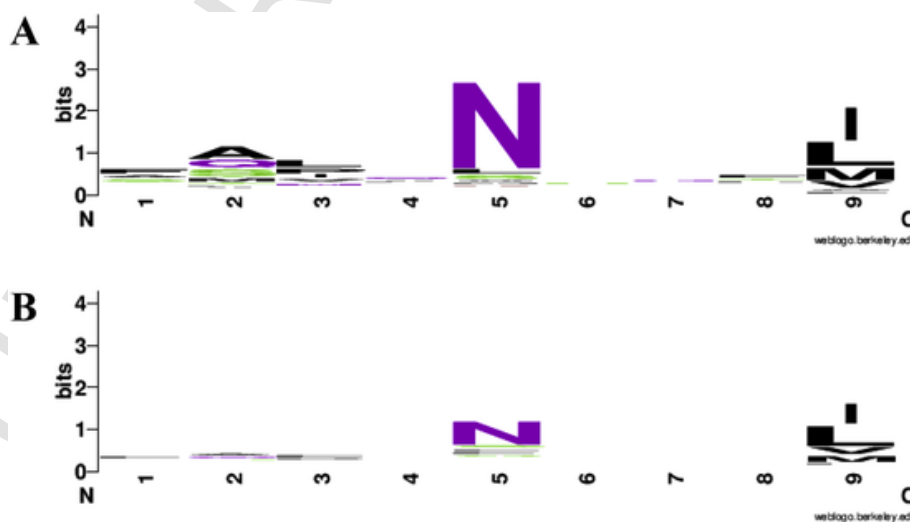


Fig. 4. A: Sequence logo of peptides that bind to H-2D<sup>b</sup> ( $k_D \leq 500$  nM). B: Sequence logo of peptides that do not bind to H-2D<sup>b</sup> ( $k_D > 500$  nM).

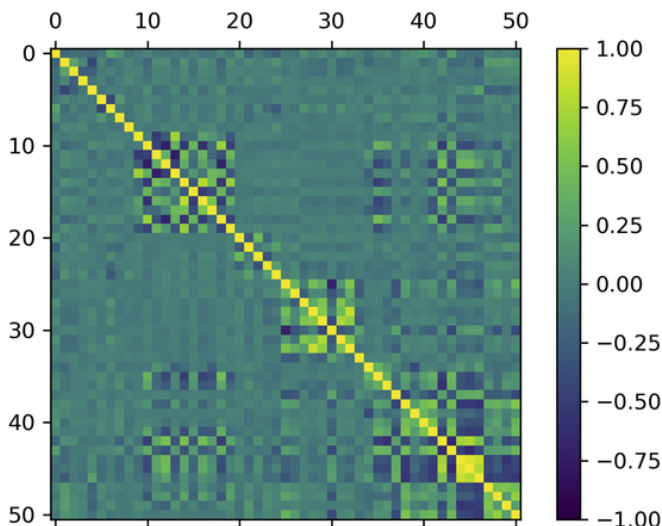


Fig. 5. Correlation plot between features of the hybrid descriptor set.

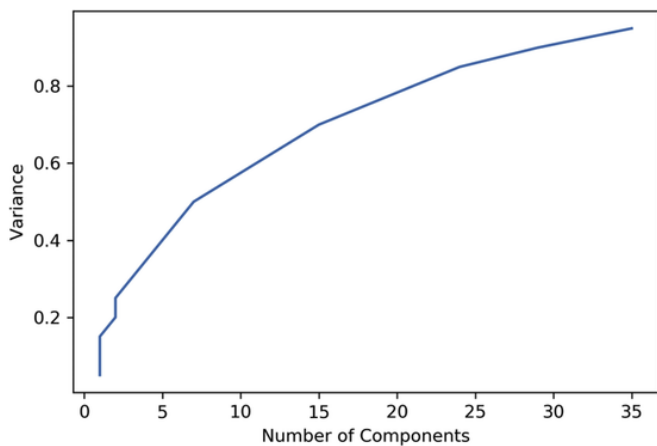


Fig. 6. Variance captured as a function of number of principal components.

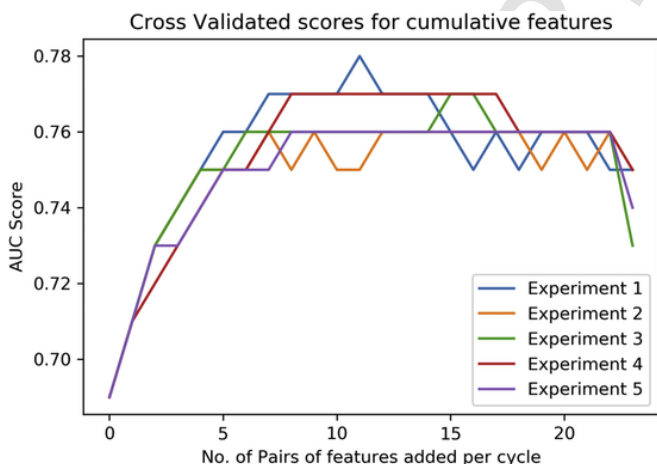


Fig. 7. Performance as a function of number of cycles. The feature set grows by 2 in each cycle.

the receptor and ligand, respectively. The fractal density dimension,  $\lambda_i$ , is the exponent in a power law relating number of atoms,  $N$ , to the radial distance  $r$  from a reference point centered at an interfacial receptor atom as follows:

$$N = r^{\lambda_i} \quad (6)$$

For a locally flat surface,  $\lambda_i$  would be 3.0, and this would be the corresponding value for the reference value of the fractal density dimension,  $\lambda_0$ . However, for atomic number densities in typical proteins, the observed reference value is less than the 3.0 [38]; hence, in this work, a value of 2.75 is used for  $\lambda_0$ . For each receptor interface atom  $i$ ,  $\lambda_i$  is calculated from a least-squares fit of  $\log(N)$  vs.  $\log(r)$ .

**Electrostatic energy.** A number of unscreened or screened electrostatic models were used to represent vacuum electrostatics and/or polar desolvation [26]. These all can be described by the following general form:

$$E_{elec} = \sum_i^{N_{ligand}} \sum_j^{N_{protein}} 331.8 \frac{q_i q_j}{\epsilon(r_{ij}) r_{ij}} \quad (7)$$

The distinguishing feature among the models was the functional form of the screening term  $\epsilon(r_{ij})$ , which corresponds to the prescription for polar desolvation. In the first model, a constant value of  $\epsilon(r_{ij}) = 10$  was used to approximate a value for the protein surface between vacuum and water. In the second model, a simple distance-dependent dielectric was used where  $\epsilon(r_{ij}) = r_{ij}$  [39]. For the next set of models, different functional forms for a sigmoidal or quasi-sigmoidal distance-dependent dielectric were used. The first of these was due to Warshel and co-workers [40,41]:

$$\begin{aligned} \epsilon(r_{ij}) &= 16.55 r_{ij} < 3\text{\AA} \\ \epsilon(r_{ij}) &= 1 + 60 \left(1 - e^{-\frac{r_{ij}}{10}}\right) r_{ij} \geq 3\text{\AA} \end{aligned}$$

The following quasi-sigmoidal distance-dependent dielectric was developed by Sternberg and co-workers [42]:

$$\begin{aligned} \epsilon(r_{ij}) &= 4 r_{ij} \leq 6\text{\AA} \\ \epsilon(r_{ij}) &= 38 r_{ij} - 224 \text{ } 6\text{\AA} < r_{ij} < 8\text{\AA} \\ \epsilon(r_{ij}) &= 80 r_{ij} \geq 8\text{\AA} \end{aligned} \quad (8)$$

The third sigmoidal dielectric due to Ramstein and Lavery [43] is as follows:

$$\begin{aligned} \epsilon(r_{ij}) &= 78 - \frac{(78-1)}{2} \left( (0.16 r_{ij})^2 + 2 \cdot 0.16 r_{ij} \right. \\ &\quad \left. + 2 \right) e^{-0.16 r_{ij}} \end{aligned} \quad (9)$$

Lastly, a sigmoidal dielectric due to Hingerty and co-workers [44] was used as follows:

$$\epsilon(r_{ij}) = 78 - 77 \left( \frac{r}{2.5} \right)^2 \frac{e^{\frac{r}{2.5}}}{\left( e^{\frac{r}{2.5}} - 1 \right)^2} \quad (10)$$

**Hydrogen bonding energy.** Four different hydrogen bonding potentials were used. The first two of these used a softened 12–10 potential with an angular dependency term  $f(\theta, \phi)$  as follows:

$$\begin{aligned} E_{hb} &= \sum_i^{N_{ligand}} \sum_j^{N_{protein}} D_0 \left[ 5 \left( \frac{r_0}{r_{ij}} \right)^{12} - 6 \left( \frac{r_0}{r_{ij}} \right)^{10} \right] f(\theta, \phi) r_{ij} > r_0 + 0.1 \\ E_{hb} &= \sum_i^{N_{ligand}} \sum_j^{N_{protein}} D_0 \left[ 5 \left( \frac{r_0}{r_{ij}+1} \right)^{12} - 6 \left( \frac{r_0}{r_{ij}+1} \right)^{10} \right] f(\theta, \phi) r_0 - 0.5 \leq r_{ij} \leq \\ E_{hb} &= \sum_i^{N_{ligand}} \sum_j^{N_{protein}} \frac{D_0}{64^2} \left[ \frac{5}{64^2} \left( \frac{r_0}{r_{ij}} \right)^{12} - \frac{3}{2} \left( \frac{r_0}{r_{ij}} \right)^{10} \right] f(\theta, \phi) r_{ij} < r_0 - 0. \end{aligned}$$

here  $r_{ij}$  is the donor-acceptor distance. In the first of the 12–10 potentials [45], the well depth  $D_0 = 9.5$ , the equilibrium donor-acceptor distance  $r_0 = 2.75$ , and the angular dependency term was as

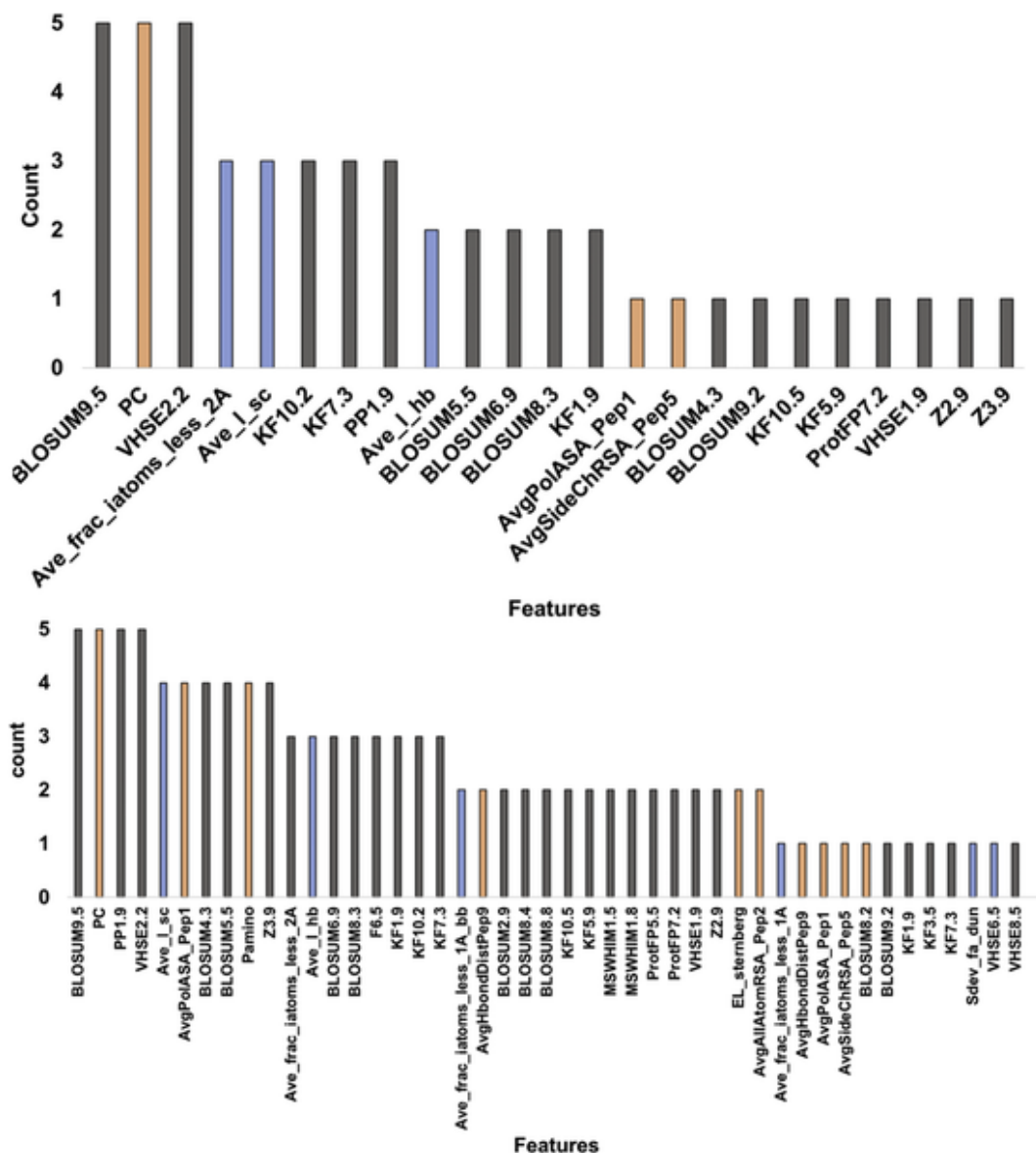


Fig. 8. A) Histogram showing the frequency of appearance of features in the first 5 cycles in the 5 independent runs of forward feature selection. B) Histogram showing the frequency of appearance of features in the first 10 cycles in the 5 runs independent runs of forward feature selection. The bars of the sequence-based, structure-based Rosetta and DM features are shown in gray, blue and orange respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

follows:

$$f(\theta, \phi) = f(\theta) = \cos^2\theta \quad (12)$$

In the above,  $\theta$  is the donor-hydrogen-acceptor angle and  $\phi$  is the angle between the hydrogen, acceptor, and the atom covalently bound to the acceptor. In the second of the 12–10 potentials from the in-house code, a more elaborate angular dependency term was used along with slightly different values of the well depth,  $D_0$  (8.0), and equilibrium distance,  $r_0$  (2.8), [46]. Here the angular dependency was as follows, differing according to the hybridization state of the donor and acceptor:

$$\begin{aligned} sp^3 \text{ donor} - sp^3 \text{ acceptor} : f(\theta, \phi) &= \cos^2\theta \cos^2(\phi - 109.5) \\ sp^3 \text{ or } sp^2 \text{ donor} - sp^2 \text{ acceptor} : f(\theta, \phi) &= \cos^2\theta \cos^2\phi \\ sp^2 \text{ donor} - sp^3 \text{ acceptor} : f(\theta, \phi) &= f(\theta) = \cos^4\theta \end{aligned} \quad (13)$$

The third hydrogen bonding potential from the in-house code was a Morse potential with angular dependency [47]:

$$E_{hb} = \sum_i^{N_{ligand}} \sum_j^{N_{protein}} D_0 \left(1 - e^{-\alpha(r_{ij} - r_0)}\right)^2 f(\theta, \phi) \quad (14)$$

Here, the well depth  $D_0$  was 7.785 and  $r_0 = 1.912$  is the equilibrium distance between the hydrogen and the acceptor,  $r_{ij}$  is the distance between the hydrogen and acceptor,  $\alpha$  is a parameter controlling the curvature of the potential well (set to 1.234). The angular dependency was as follows, maintaining the definitions of  $\theta$  and  $\phi$  as above:

$$\begin{aligned} f(\theta, \phi) &= \sum_{i=1}^3 (a_i \cos^i\theta + b_i \sin^i\theta) \sum_{j=1}^3 (c_j \cos^j\phi + d_j \sin^j\phi) \\ a_1 &= -0.106, b_1 = 0.671, c_1 = 1.494, d_1 = -1.494 \\ a_2 &= -2.953, b_2 = 2.976, c_2 = -0.059, d_2 = 2.906 \\ a_3 &= 1.494, b_3 = -0.224, c_3 = -0.482, d_3 = -1.918 \end{aligned} \quad (15)$$

**Table 5**

Performance of models based on features that persisted in different number of runs of the two-way forward feature selection. N is the number of two-way forward feature selection runs in which the features persisted in first 5 or first 10 cycles.

N	Type of model	Training set CV LOO performance						Test set performance					
		Accuracy	Precision	Recall	Specificity	F-score	AUC	Accuracy	Precision	Recall	Specificity	F-score	AUC
Features collected after first five cycles													
5	3-feature	0.67	0.55	0.80	0.58	0.65	0.69	0.62	0.52	0.76	0.52	0.61	0.64
≥3	8-feature	0.72	0.61	0.78	0.68	0.68	0.73	0.67	0.57	0.71	0.64	0.63	0.67
≥2	13-feature	0.73	0.62	0.79	0.69	0.69	0.74	0.64	0.54	0.73	0.59	0.62	0.66
≥1	23-feature	0.74	0.63	0.77	0.71	0.69	0.74	0.67	0.57	0.72	0.64	0.64	0.68
Features collected after first ten cycles													
5	4-feature	0.68	0.57	0.80	0.61	0.67	0.71	0.63	0.53	0.74	0.56	0.62	0.65
≥4	10-feature	0.72	0.60	0.81	0.66	0.69	0.73	0.62	0.52	0.72	0.56	0.610	0.64
≥3	18-feature	0.73	0.62	0.80	0.69	0.70	0.75	0.66	0.56	0.70	0.63	0.62	0.66
≥2	32-feature	0.73	0.62	0.80	0.69	0.70	0.75	0.70	0.60	0.78	0.65	0.68	0.71
≥1	45-feature	0.74	0.63	0.82	0.69	0.71	0.75	0.70	0.59	0.77	0.65	0.67	0.71

Lastly, a 6–4 potential with angular dependency was implemented as follows [39]:

$$E_{hb} = \sum_i^{N_{ligand}} \sum_j^{N_{protein}} D_0 \left[ \left( \frac{r_0}{r_{ij}} \right)^6 - \left( \frac{r_0}{r_{ij}} \right)^4 \right] f(\theta, \phi) \quad (16)$$

where

$$f(\theta, \phi) = \cos^2 \theta \cos^2 \phi$$

In this potential, there are well-depths and equilibrium distances specific to the participating donor and acceptor atoms as follows:

$$\begin{aligned} \text{Nitrogen donor, oxygen acceptor} : D_0 = 2.9, r_0 = 4.9 \\ \text{Nitrogen donor, nitrogen acceptor} : D_0 = 3.0, r_0 = 5.0 \\ \text{Oxygen donor, nitrogen acceptor} : D_0 = 2.85, r_0 = 4.85 \\ \text{Oxygen donor, oxygen acceptor} : D_0 = 2.75, r_0 = 4.75 \end{aligned} \quad (17)$$

**$\pi$ - $\pi$  stacking energy.** Two potentials were used to model the energy due to  $\pi$ - $\pi$  stacking between aromatic rings. The first of these was a Morse potential with angular dependency originally proposed by Yuki et al. [48] and modified by us to include softening at short interatomic distances [26]. This potential consists of a so-called parallel component and a perpendicular component, referring to the orientation of the aromatic rings with respect to each other:

$$\begin{aligned} E_{\pi-\pi}^{perpendicular} &= \sum_i^{N_{prot}} \sum_j^{N_{lig}} D_0 \left\{ \left[ 1 - e^{-a(r_{ij}-r_0)} \right]^2 - 1 \right\} \left[ 1 - (\vec{n}_i - \vec{n}_j)^2 \right] \cos^2 \theta \\ E_{\pi-\pi}^{perpendicular} &= \sum_i^{N_{prot}} \sum_j^{N_{lig}} D_0 \left\{ \left[ 1 - e^{-a(4.6-r_0)} \right]^2 - 1 \right\} \left[ 1 - (\vec{n}_i - \vec{n}_j)^2 \right] \cos^2 \theta \\ E_{\pi-\pi}^{perpendicular} &= \sum_i^{N_{prot}} \sum_j^{N_{lig}} D_0 \left\{ \left[ 1 - e^{-0.5a(r_{ij}-r_0)} \right]^2 - 1 \right\} \left[ 1 - (\vec{n}_i - \vec{n}_j)^2 \right] \cos^2 \theta \\ E_{\pi-\pi}^{parallel} &= \sum_i^{N_{prot}} \sum_j^{N_{lig}} D_0 \left\{ \left[ 1 - e^{-a(r_{ij}-r_0)} \right]^2 - 1 \right\} (\vec{n}_i - \vec{n}_j)^2 \cos^2 \theta \\ E_{\pi-\pi}^{parallel} &= \sum_i^{N_{prot}} \sum_j^{N_{lig}} D_0 \left\{ \left[ 1 - e^{-a(3.6-r_0)} \right]^2 - 1 \right\} (\vec{n}_i - \vec{n}_j)^2 \cos^2 \theta \\ E_{\pi-\pi}^{parallel} &= \sum_i^{N_{prot}} \sum_j^{N_{lig}} D_0 \left\{ \left[ 1 - e^{-0.5a(r_{ij}-r_0)} \right]^2 - 1 \right\} (\vec{n}_i - \vec{n}_j)^2 \cos^2 \theta \end{aligned}$$

here,  $r_{ij}$  is the distance between the centers of the aromatic rings,  $\vec{n}$  is the unit normal vector with respect to the plane of the aromatic ring,  $\theta$  is the angle formed by the normal vector of one ring and a vector originating from the center of the first ring and pointing towards the center of the second ring,  $D_0$  is the well depth,  $a$  prescribes the curvature of the potential well, and  $r_0$  and  $\theta_0$  are the

equilibrium values of  $r_{ij}$  and  $\theta_0$ , respectively. The values of the parameters are as follows:

$$\begin{aligned} E_{\pi-\pi}^{perpendicular} D_0 = 2.41, a = 1.29, r_0 = 5.05, \theta_0 = 0.0^\circ \\ E_{\pi-\pi}^{parallel} D_0 = 2.57, a = 1.40, r_0 = 3.95, \theta_0 = 27.0^\circ \end{aligned} \quad (19)$$

A second heuristic potential from the in-house code was as follows:

$$E_{\pi-\pi} = \sum_i^{N_{prot}} \sum_j^{N_{lig}} \frac{-0.05 \cdot (8.0 - r_{ij})}{4.0} \quad (20)$$

In the above,  $r_{ij}$  is the distance between any pair of atoms, each from either an aromatic ring in the protein or ligand.

**$\pi$ -cation/polar energy.** Interactions between aromatic rings and nitrogen-containing cationic groups, and two types of polar, non-charged groups (amine and hydroxyl groups) were calculated using three functional forms. The first of these was a Morse potential adapted from Yuki et al. [48] as follows:

$$\begin{aligned} E_{\pi-cat/polar} &= \sum_i^{N_{prot}} \sum_j^{N_{lig}} D_0 \left\{ \left[ 1 - e^{-a(r_{ij}-r_0)} \right]^2 - 1 \right\} \cos^2(\theta - \theta_0) \quad r_{ij} > r_{cut} \\ E_{\pi-cat/polar} &= \sum_i^{N_{prot}} \sum_j^{N_{lig}} D_0 \left\{ \left[ 1 - e^{-a(r_{cut}-r_0)} \right]^2 - 1 \right\} \cos^2(\theta - \theta_0) \quad r_{cut} \leq r_{ij} \\ E_{\pi-cat/polar} &= \sum_i^{N_{prot}} \sum_j^{N_{lig}} D_0 \left\{ \left[ 1 - e^{-0.5a(r_{ij}-r_0)} \right]^2 - 1 \right\} \cos^2(\theta - \theta_0) \quad r_{ij} \leq r_{cut} \end{aligned}$$

Parameters for this potential are as described above for the corresponding Morse potential for  $\pi$ - $\pi$  stacking interactions, except here  $r_{ij}$  is the distance from the center of the aromatic ring and the heavy atom of the amino, cation, or hydroxyl group ( $r_0$  is the corresponding equilibrium distance), and  $\theta$  is the angle formed by the normal vector of the aromatic ring and a vector originating from the center of that ring and pointing towards the cationic or polar heavy atom. The parameters for the various interactions are as follows:

$$\begin{aligned} E_{\pi-cat} D_0 = 3.74, a = 1.25, r_0 = 3.61, r_{cut1} = 3.2, r_{cut2} = 2.4, \theta_0 = 0.0^\circ \\ E_{\pi-hydroxyl} D_0 = 2.79, a = 1.19, r_0 = 3.46, r_{cut1} = 2.4, \theta_0 = 0.0^\circ \end{aligned}$$

The second potential from the in-house code was heuristic and evaluated for  $\pi$ -cation and  $\pi$ -amine interactions:

$$E_{\pi-cat/amine} = \sum_i^{N_{prot}} \sum_j^{N_{lig}} \frac{-0.025 \cdot (8.0 - r_{ij})}{4.0} \quad (23)$$

Here  $r_{ij}$  is the distance between the nitrogen atom and any of the constituent atoms of the aromatic ring.

Lastly, from the in-house code a Lennard-Jones-type 12–4 potential proposed by Minoux and Chipot [49] was evaluated for  $\pi$ -cation and  $\pi$ -amine interactions and is as follows:

$$E_{\pi\text{-cat/amine}} = \sum_i \sum_j \frac{N_{pmi} N_{ij}}{r_{ij}^{12}} - \frac{144.355}{r_{ij}^4} \quad (24)$$

Here, as in the heuristic potential for  $\pi$ -cation and  $\pi$ -amine interactions,  $r_{ij}$  is the distance between the nitrogen atom and any of the constituent atoms of the aromatic ring.

**Hydrogen bonding.** Total hydrogen bonds, number of donors/acceptors, distance of donor to acceptor, and donor-hydrogen-acceptor angle were calculated using PDB2PQR [50].

**Solvent-accessible surface area.** Absolute and relative SASA were calculated for all atoms, sidechain atoms, mainchain atoms, polar atoms, and non-polar atoms using NACCESS [51]. The relative SASA is calculated relative to the alanine tri-peptide.

**Root mean-squared deviation.** RMSD of the atomic coordinates were calculated for all atoms, the sidechain atoms, backbone atoms, and the alpha-carbon atoms. RMSD was calculated using the one of the structures of the ensemble as the reference structure.

## References

- [1] M. Delorenzi, T. Speed, An HMM model for coiled-coil domains and a comparison with PSSM-based predictions, *Bioinformatics* 18 (2002) 617–625.
- [2] T.N. Schumacher, R.D. Schreiber, Neoantigens in cancer immunotherapy, *Science* 348 (2015) 69–74.
- [3] V. Jurtz, S. Paul, M. Andreatta, P. Marcantili, B. Peters, M. Nielsen, NetMHCpan-4.0: improved peptide-MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data, *J. Immunol. (Baltimore, Md.)* 1950 (199) (2017) 3360–3368.
- [4] T.J. O'Donnell, A. Rubinsteyn, M. Bonsack, A.B. Riemer, U. Laserson, J. Hammerbacher, MHCflurry: open-source class I MHC binding affinity prediction, *Cell Syst.* 7 (2018) (129–132.e124)s.
- [5] F. Duan, J. Duitama, S. Al Seesi, C.M. Ayres, S.A. Corcelli, A.P. Pawashe, T. Blanchard, D. McMahon, J. Sidney, A. Sette, Genomic and bioinformatic profiling of mutational neopeptides reveals new rules to predict anticancer immunogenicity, *J. Exp. Med.* 211 (2014) 2231–2248.
- [6] V. Zoete, M. Irving, M. Ferber, M. Cuendet, O. Michielin, Structure-based, rational design of T cell receptors, *Front. Immunol.* 4 (2013) 268.
- [7] M.M. Rigo, D.A. Antunes, M.V. De Freitas, M.F. de Almeida Mendes, L. Meira, M. Sinigaglia, G.F. Vieira, DockTope: a web-based tool for automated pMHC-I modeling, *Sci. Rep.* 5 (2015) 18413.
- [8] G.J. van Westen, R.F. Swier, J.K. Wegner, A.P. IJzerman, H.W. van Vlijmen, A. Bender, Benchmarking of protein descriptor sets in proteochemometric modeling (part 1): comparative study of 13 amino acid descriptor sets, *J. Cheminform.* 5 (2013) 41.
- [9] R. Vita, S. Mahajan, J.A. Overton, S.K. Dhanda, S. Martini, J.R. Cantrell, D.K. Wheeler, A. Sette, B. Peters, The immune epitope database (IEDB): 2018 update, *Nucleic Acids Res.* 47 (2018) D339–D343.
- [10] A. Sette, A. Vitiello, B. Reheman, P. Fowler, R. Nayarsina, W.M. Kast, C.J.M. Melief, C. Oseroff, L. Yuan, J. Ruppert, J. Sidney, M.F. Del Guercio, S. Southwood, R.T. Kubo, R.W. Chesnut, H.M. Grey, F.V. Chisari, The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes, *J. Immunol.* 153 (1994) 5586–5592.
- [11] D. Osorio, P. Rondón-Villarrea, R. Torres, Peptides: a package for data mining of antimicrobial peptides, *R J.* 7 (2015).
- [12] G.J. van Westen, R.F. Swier, I. Cortes-Ciriano, J.K. Wegner, J.P. Overington, A.P. IJzerman, H.W. van Vlijmen, A. Bender, Benchmarking of protein descriptor sets in proteochemometric modeling (part 2): modeling performance of 13 amino acid descriptor sets, *J. Cheminform.* 5 (2013) 42.
- [13] A.G. Georgiev, Interpretable numerical descriptors of amino acid space, *J. Comput. Biol.* 16 (2009) 703–723.
- [14] H.F. Kaiser, The varimax criterion for analytic rotation in factor analysis, *Psychometrika* 23 (1958) 187–200.
- [15] G. Liang, Z. Li, Factor analysis scale of generalized amino acid information as the source of a new set of descriptors for elucidating the structure and activity relationships of cationic antimicrobial peptides, *QSAR Combinatorial Sci.* 26 (2007) 754–763.
- [16] A. Kidera, Y. Konishi, M. Oka, T. Ooi, H.A. Scheraga, Statistical analysis of the physical properties of the 20 naturally occurring amino acids, *J. Protein Chem.* 4 (1985) 23–55.
- [17] A. Zaliani, E. Gancia, MS-WHIM scores for amino acids: a new 3D-description for peptide QSAR and QSPR studies, *J. Chem. Inf. Comput. Sci.* 39 (1999) 525–533.
- [18] G. Cruciani, M. Baroni, E. Carosati, M. Clementi, R. Valigi, S. Clementi, Peptide studies by means of principal properties of amino acids derived from MIF descriptors, *J. Chemom.* 18 (2004) 146–155.
- [19] H. Mei, Z.H. Liao, Y. Zhou, S.Z. Li, A new set of amino acid descriptors and its application in peptide QSARs, *Peptide Sci. Orig. Res. Biomol.* 80 (2005) 775–786.
- [20] M. Sandberg, L. Eriksson, J. Jonsson, M. Sjöström, S. Wold, New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids, *J. Med. Chem.* 41 (1998) 2481–2491.
- [21] M. Shen, A. Sali, Statistical potential for assessment and prediction of protein structures, *Protein Sci.* 15 (2006) 2507–2524.
- [22] B. Raveh, N. London, O. Schueler-Furman, Sub-angstrom modeling of complexes between flexible peptides and globular proteins, *Proteins: structure, Funct. Bioinform.* 78 (2010) 2029–2040.
- [23] B. Raveh, N. London, O. Schueler-Furman, Sub-angstrom modeling of complexes between flexible peptides and globular proteins, *Prot. Struct. Funct. BioinformXXX* 78 (2010) 2029–2040.
- [24] K.W. Kaufmann, G.H. Lemmon, S.L. DeLuca, J.H. Sheehan, J. Meiler, Practically useful: what the Rosetta protein modeling suite can do for you, *Biochemistry* 49 (2010) 2987–2998.
- [25] R.F. Alford, A. Leaver-Fay, J.R. Jeliazkov, M.J. O'Meara, F.P. DiMaio, H. Park, M.V. Shapovalov, P.D. Renfrew, V.K. Mulligan, K. Kappel, The Rosetta all-atom energy function for macromolecular modeling and design, *J. Chem. Theory Comput.* 13 (2017) 3031–3048.
- [26] O.N.A. Demerdash, J.C. Mitchell, Using physical potentials and learned models to distinguish native binding interfaces from de novo designed interfaces that do not bind, *Prot. Struct. Funct. Bioinform.* 81 (2013) 1919–1930.
- [27] F. Seeger, A. Little, Y. Chen, T. Woolf, H. Cheng, J.C. Mitchell, Feature design for protein interface hotspots using KFC2 and Rosetta, in: E. Gasparovic, C. Domeniconi (Eds.), *Research in Data Science*, Springer International Publishing, Cham, 2019, pp. 177–197.
- [28] H.-G. Rammensee, J. Bachmann, N.P.N. Emmerich, O.A. Bachor, S. Stevanović, SYFPEITHI: database for MHC ligands and peptide motifs, *Immunogenetics* 50 (1999) 213–219.
- [29] G.E. Crooks, G. Hon, J.-M. Chandonia, S.E. Brenner, WebLogo: a sequence logo generator, *Genome Res.* 14 (2004) 1188–1190.
- [30] N. London, C.L. Lamphear, J.L. Hougland, C.A. Fierke, O. Schueler-Furman, Identification of a novel class of farnesylation targets by structure-based modeling of binding specificity, *PLoS Comput. Biol.* 7 (2011) e1002170.
- [31] N. London, S. Gullá, A.E. Keating, O. Schueler-Furman, In silico and in vitro elucidation of BH3 binding specificity toward Bcl-2, *Biochemistry* 51 (2012) 5841–5850.
- [32] N. Alam, L. Zimmerman, N.A. Wolfson, C.G. Joseph, C.A. Fierke, O. Schueler-Furman, Structure-based identification of HDAC8 non-histone substrates, *Structure (Lond., Engl.)* 24 (2016) (1993) 458–468.
- [33] N. Alam, O. Schueler-Furman, Modeling peptide-protein structure and binding using Monte Carlo sampling approaches: Rosetta FlexPepDock and FlexPepBind, in: O. Schueler-Furman, N. London (Eds.), *Modeling Peptide-Protein Interactions: Methods and Protocols*, Springer New York, New York, NY, 2017, pp. 139–169.
- [34] C.A. Rohl, C.E.M. Strauss, K.M.S. Misura, D. Baker, Protein structure prediction using rosetta, *Methods Enzymol.* 383 (2004) (66–+).
- [35] B. Raveh, N. London, O. Schueler-Furman, Sub-angstrom modeling of complexes between flexible peptides and globular proteins, *Prot. Struct. Funct. Bioinform.* 78 (2010) 2029–2040.
- [36] S.J. Fleishman, T.A. Whitehead, E.M. Strauch, J.E. Corn, S.B. Qin, H.X. Zhou, J.C. Mitchell, O.N.A. Demerdash, M. Takeda-Shitaka, G. Terashi, I.H. Moal, X.F. Li, P.A. Bates, M. Zacharias, H. Park, J.S. Ko, H. Lee, C. Seok, T. Bourquard, J. Bernauer, A. Poupon, J. Aze, S. Soner, S.K. Ovali, P. Ozbek, N. Ben Tal, T. Haliloglu, H. Hwang, T. Vreven, B.G. Pierce, Z.P. Weng, L. Perez-Cano, C. Pons, J. Fernandez-Recio, F. Jiang, F. Yang, X.Q. Gong, L.B. Cao, X.J. Xu, B. Liu, P.W. Wang, C.H. Li, C.X. Wang, C.H. Robert, M. Guharoy, S.Y. Liu, Y.Y. Huang, L. Li, D.C. Guo, Y. Chen, Y. Xiao, N. London, Z. Itzhaki, O. Schueler-Furman, Y. Inbar, V. Potapov, M. Cohen, G. Schreiber, Y. Tsuchiya, E. Kanamori, D.M. Standley, H. Nakamura, K. Kinoshita, C.M. Driggers, R.G. Hall, J.L. Morgan, V.L. Hsu, J. Zhan, Y.D. Yang, Y.Q. Zhou, P.L. Kastiris, A.M.J.J. Bonvin, W.Y. Zhang, C.J. Camacho, K.P. Kilambi, A. Sircar, J.J. Gray, M. Ohue, N. Uchikoga, Y. Matsuzaki, T. Ishida, Y. Akiyama, R. Khashan, S. Bush, D. Fouches, A. Tropsha, J. Esquivel-Rodriguez, D. Kihara, P.B. Stranges, R. Jacak, B. Kuhlman, S.Y. Huang, X.Q. Zou, S.J. Wodak, J. Janin, D. Baker, Community-wide assessment of protein-interface modeling suggests improvements to design methodology, *J. Mol. Biol.* 414 (2011) 289–302.
- [37] J.C. Mitchell, R. Kerr, L.F. Ten Eyck, Rapid atomic density methods for molecular shape characterization, *J. Mol. Graph. Model* 19 (2001) (325–+).
- [38] L.A. Kuhn, M.A. Siani, M.E. Pique, C.L. Fisher, E.D. Getzoff, J.A. Tainer, The interdependence of protein surface-topography and bound water-molecules revealed by surface accessibility and fractal density measures, *J. Mol. Biol.* 228 (1992) 13–22.
- [39] B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan, M. Karplus, Charmm - a program for macromolecular energy, minimization, and dynamics calculations, *J. Comput. Chem.* 4 (1983) 187–217.
- [40] A. Warshel, S.T. Russell, Calculations of electrostatic interactions in biological-systems and in solutions, *Q. Rev. Biophys.* 17 (1984) 283–422.
- [41] A. Warshel, S.T. Russell, A.K. Churg, Macroscopic models for studies of electrostatic interactions in proteins - limitations and applicability, *Proc. Natl. Acad. Sci.-Biol.* 81 (1984) 4785–4789.
- [42] H.A. Gabb, R.M. Jackson, M.J.E. Sternberg, Modelling protein docking using shape complementarity, electrostatics and biochemical information, *J. Mol. Biol.* 272 (1997) 106–120.

- [43] J. Ramstein, R. Lavery, Energetic coupling between DNA bending and base pair opening, *Proc. Natl. Acad. Sci. U. S. A.* 85 (1988) 7231–7235.
- [44] B.E. Hingerty, R.H. Ritchie, T.L. Ferrell, J.E. Turner, Dielectric effects in bio-polymers - the theory of ionic saturation revisited, *Biopolymers* 24 (1985) 427–439.
- [45] S.L. Mayo, B.D. Olafson, W.A. Goddard, Dreiding - a generic force-field for molecular simulations, *J. Phys. Chem.-Us* 94 (1990) 8897–8909.
- [46] B.I. Dahiyat, D.B. Gordon, S.L. Mayo, Automated design of the surface positions of protein helices, *Protein Sci.* 6 (1997) 1333–1337.
- [47] K.I. Cho, K. Lee, K.H. Lee, D. Kim, D. Lee, Specificity of molecular interactions in transient protein-protein interaction interfaces, *Prot. Struct. Funct. Bioinform.* 65 (2006) 593–606.
- [48] H. Yuki, Y. Tanaka, M. Hata, H. Ishikawa, S. Neya, T. Hoshino, Implementation of pi-pi interactions in molecular dynamics simulation, *J. Comput. Chem.* 28 (2007) 1091–1099.
- [49] H. Minoux, C. Chipot, Cation-pi interactions in proteins: can simple models provide an accurate description?, *J. Am. Chem. Soc.* 121 (1999) 10366–10372.
- [50] T.J. Dolinsky, J.E. Nielsen, J.A. McCammon, N.A. Baker, PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations, *Nucleic Acids Res.* 32 (2004) W665–W667.
- [51] S.J. Hubbard, J.M. Thornton, NACCESS, Department of Biochemistry and Molecular Biology, University College London, 1993 (Computer Program).

UNCORRECTED PROOF