3pHLA-score improves structure-based peptide-HLA binding affinity prediction

Anja Conev[1], Didier Devaurs[2], Mauricio Menegatti Rigo[1], Dinler Amaral Antunes[3,*], and Lydia E Kavraki[1,*]

[1]Department of Computer Science, Rice University, Houston, 77005, USA
[2]MRC Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, EH4 2XU, UK
[3]Department of Biology and Biochemistry, University of Houston, Houston, 77004, USA
[*]antunesda@central.uh.edu, kavraki@rice.edu

ABSTRACT

Binding of peptides to Human Leukocyte Antigen (HLA) receptors is a prerequisite for triggering immune response. Estimating peptide-HLA (pHLA) binding is crucial for peptide vaccine target identification and epitope discovery pipelines. Computational methods for binding affinity prediction can accelerate these pipelines. Currently, most of those computational methods rely exclusively on sequence-based data, which leads to inherent limitations. Recent studies have shown that structure-based data can address some of these limitations. In this work we propose a novel machine learning (ML) structure-based protocol to predict binding affinity of peptides to HLA receptors. For that, we engineer the input features for ML models by decoupling energy contributions at different residue positions in peptides, which leads to our novel per-peptide-position protocol. Using Rosetta's ref2015 scoring function as a baseline we use this protocol to develop 3pHLA-score. Our per-peptide-position protocol outperforms the standard training protocol and leads to an increase from 0.82 to 0.99 of the area under the precision-recall curve. 3pHLA-score outperforms widely used scoring functions (AutoDock4, Vina, Dope, Vinardo, FoldX, GradDock) in a structural virtual screening task. Overall, this work brings structure-based methods one step closer to epitope discovery pipelines and could help advance the development of cancer and viral vaccines.

Introduction

Human Leukocite Antigen (HLA) class I molecules are an important part of human cellular immune response[1,2]. HLAs are involved in the intracellular antigen presentation pathway; they are responsible for the transport and display of peptide antigens for T-cell scrutiny[3,4]. Therefore, the possibility of exploiting the HLA role in this pathway to engineer immune responses has shown great promise[5], as highlighted by efforts on personalized peptide vaccine development[6]. When designing peptide vaccines, a pool of potential peptide targets is identified from a protein of interest. Targets are then filtered to identify those most likely to induce an immune response. This whole process is referred to as epitope discovery[7]. Discovered immunogenic epitopes are able to bind HLA receptors, create stable peptide-HLA (pHLA) complexes (Figure S1) and induce an immunological response[8]. Unfortunately, epitope discovery is made challenging by the high diversity of HLA molecules. This diversity is a reflection of the high number of HLA alleles: more than 24,000 HLA-I alleles have been identified to date[9]. Each allele codes for a specific HLA receptor (e.g., HLA-A0201, HLA-B0702) with different peptide binding preferences. Fast and accurate computational evaluation of pHLA binding can speed up the search for epitopes and is an important part of epitope discovery pipelines.

So far computational pHLA binding affinity prediction efforts have been largely dominated by sequence-based approaches[10–15]. While these methods provide good accuracy and are a part of many existing pipelines, they have some inherent drawbacks[16]. For instance, they rely on a predefined amino acid alphabet to represent the pHLA. Most existing tools have canonical amino acids in their alphabet[10–12] and are thus unable to process phosphorylated peptides, although these peptides can be displayed by HLAs[17]. While recent efforts[18] expand the alphabet to include phosphorylation, the problem of the predefined alphabet persists. The presence of other post-translational modifications or small molecules within the binding site cannot be taken into account by such approaches. In addition, sequence-based predictors are highly dependent on the quality and composition of the training set[19,20]. This represents an important limitation because of the aforementioned high diversity of HLA alleles[21]. All these challenges indicate that sequence-based methods alone can not identify all relevant epitopes, which motivates further exploration and development of complementary approaches[22].

Structure-based methods use three-dimensional arrangements (i.e., conformations) of receptors and ligands[23]. They are not

restricted to a predefined amino acid alphabet and can be used in docking or structural virtual screening tasks[24]. In the context of these tasks, structure-based scoring functions are used to approximate the free energy of a molecular system. Most scoring functions are generic and can be used to score any complex of interest (including pHLAs), but their performance is often system-dependent[25]. To tailor scoring functions to a specific protein family, machine learning (ML) efforts are emerging[26,27]. As reliable pHLA modeling tools arise[23,28], and more data become available, we see a potential for pHLA ML scoring functions and structure-based methods to enter epitope discovery pipelines and complement existing sequence-based methods.

Under the hood, most scoring functions (such as Rosetta's ref2015[29]) approximate independent energy terms for a molecular complex and rely on the assumption that binding affinity can be described as a weighted sum of these terms[30]. Standard ML training protocols use the same assumption. GradDock[31], for example, involves ref2015 standard energy terms and redefines their weights to better fit the HLA system while keeping the additive formulation. However, this additive functional form of classical (and ML-derived) scoring functions has been challenged in previous studies[32,33]. SIEVE-Score[34] recently considered binding site residues and exemplified the benefit of decomposing the energy terms associated with binding site residues for interaction-energy-based learning. The idea of assessing peptide binding affinity via a decomposition into peptide residues has also been applied in the context of other computational approaches with mixed results[35], such as quantitative structure-activity relationship (QSAR) studies involving amino acid descriptors[36].

In our approach, we decompose the energy terms of a pHLA complex into separate contributions for all residues at each position in the peptide; we then use these energy terms as input to train ML models for binding affinity prediction. We call this approach the per-peptide-position training protocol. Our rationale is that structural information that is important for pHLA binding prediction gets lost when standard scoring functions (involving the additive formulation) are applied to the pHLA complex. We use our per-peptide-position protocol in the context of the Rosetta framework[37], which leads to our novel 3pHLA-score. The main novelty of our work resides in the combination of two complementary ideas in an innovative fashion: 1) tuning the weights of Rosetta's scoring function to more accurately assess pHLA binding; 2) keeping the energy terms associated with peptide's residue positions separate to not lose information through aggregation.

We test evaluate the predictive power of our per-peptide-position protocol in a first set of experiments where we compare **3pHLA-score** with the baseline **ref2015-score** and the **standard-HLA-score** trained using the standard additive protocol. Our results show a clear lead of the per-peptide-position protocol over the standard training protocol and the default ref2015 scoring function. We then validate 3pHLA-score on two independent datasets and compare it to six widely used scoring functions: AutoDock4[38], Vina[39], Vinardo[40], GradDock[31], DOPE[41], FoldX[42]. 3pHLA-score outperforms the other scoring function in the virtual screening setting and shows the ability to generalize well on the independent datasets. This work provides a guideline for future development of ML structure-based scoring functions. Furthermore, it brings structure-based methods closer to epitope discovery pipelines, which could help advance the development of peptide vaccines.

Methods

In this work we train ML models on pHLA energy terms that are decomposed into specific contributions associated with each residue position within a peptide. We call this approach the per-peptide-position protocol and we apply it to Rosetta's ref2015 energy terms to build our 3pHLA-score. Hence, in order to explain our work, we need to first describe the ref2015-score. In addition, we describe a score that we call standard-pHLA-score which uses an intermediate protocol between ref2015 and 3pHLA-score, as it is trained for the pHLA system using the original ref2015 energy terms without decomposition.

Baseline ref2015-score

The 3D conformation of a given pHLA complex is stored in a PDB (Protein DataBank[43]) file containing coordinates of all the atoms in this molecular complex (Figure 1a). Rosetta's ref2015 scoring function feeds this all-atom information into pre-parametrized mathematical and physical models to calculate different energy terms[29]. These energy terms are based on predefined equations that model different chemical and physical aspects of a molecular system, such as electrostatics, hydrogen bonding and van der Waals interactions. The ref2015 scoring function contains 19 energy terms listed in Supplementary Table S1. The total energy of the input structure is approximated as a linear weighted sum of these energy terms. The default weights of ref2015 have been optimized on a wide range of scientific benchmarks to bring Rosetta calculations in agreement with small-molecule thermodynamic data and high-resolution structural features[29]. In this study, we approximate binding energy using ref2015-score with the equation[44]:

$$E_{binding} = E_{complex} - (E_{receptor} + E_{peptide}) \quad (1)$$

where $E_{complex}$ is the ref2015 energy of the whole complex, $E_{receptor}$ is the ref2015 energy of the HLA receptor alone and $E_{peptide}$ is the ref2015 energy of the peptide (Figure 1a).

Standard-pHLA-score

ML models can be used to refine scoring functions and tailor them to a specific system of interest. However, they do not have priors on physical and chemical properties of the molecular system. If all-atom coordinates are used as features, they can introduce noise which slows down the training and makes the learning process more difficult. This is why an initial step of transforming the structural information into compact features is needed. A standard protocol is to use the energy terms provided by traditional scoring functions as features (i.e., inputs to the models) and to tune their weights to fit a particular system[31,45]. We formulate the standard ref2015 features as a vector containing the 19 ref2015 energy terms. We train non-linear ML models (see Machine learning models subsection) using these standard features to develop the standard-pHLA-score (Figure 1b).

3pHLA-score

To develop 3pHLA-score, we go beyond the standard featurization. We decompose ref2015 energy terms into energy contributions associated with each residue position in the peptide, which we call per-peptide-position features. This protocol is inspired by the domain knowledge about the pHLA complex. Experimental findings on peptide anchors suggest that important information about the binding can be retrieved by zooming into the energy of the binding pocket at specific regions surrounding different positions in the peptide[46]. To extract the per-peptide-position features, we first scored the whole pHLA complex with Rosetta's ref2015 (as explained in the subsection above). Next, we applied PyRosetta's[47] *residue_total_energies_array* function. This function allows us to see how the structural energy of the complex breaks down into per-peptide-position contributions. The output of *residue_total_energies_array* is an array of energy terms (Table S1) for each peptide residue position, which we stack to form the input vector (see Supplementary Material subsection *Per-peptide-position feature vector*).

This vector is used as input to the non-linear ML models (see Machine learning models subsection) to create 3pHLA-score (Figure 1c).

Machine learning models

For standard-pHLA-score and 3pHLA-score we used the same dataset and settings to train our ML models - they only differ in the input features extracted from molecular structures.

We trained Random Forest Regression models[48] on a per-HLA-allele basis. For each featurization, we trained 28 models - one for each HLA allele in the dataset. We built regression trees using the CART algorithm[49] with the mean absolute error as the split criterion. To create ensembles of regression trees we used bootstrap aggregation. We scaled experimental binding affinities into the [0,1] range[11,12] (Equation S.3) and used them as prediction targets.

We compiled the training set by extracting 90% of binders and 90% of non-binders with equally distributed binding affinities out of Dataset 1 (see below). The rest of the data constitutes the test set, which was left out of the training and cross-validation phase. We stratified the training set into 5 folds (each with equal distribution of binding affinities) for hyperparameter tuning in a 5-fold cross-validation setting. Using randomized search and the 5-fold cross-validation we tuned the following parameters: number of trees, number of features per tree, maximum tree depth and minimum samples per leaf. After tuning, we evaluated the performance of the final models on the left-out test set.

Note that our main experiments describe the use of Random Forest Regression models for training the standard-pHLA-score and 3pHLA-score. However, we assessed other regression techniques: linear regression, support vector machine regression and partial least squares regression. We provide related results and discussion in the *Alternative ML regression techniques* subsection of the Supplementary Material.

Dataset 1

This dataset consists of 77,581 pHLA structures modeled by the APE-Gen modeling tool[28,50]. It involves 28 HLA alleles (13 HLA-A, 12 HLA-B and 3 HLA-C alleles). Peptides included in this dataset are all of the length 9 (9-mers). The experimental binding affinity of each pHLA complex was extracted from MHCFlurry[10], which used IEDB[51] as its main source of information. As mentioned above, Dataset 1 was split into non-overlapping training and test portions to separately train and evaluate 3pHLA-score and standard-ref2015-score.

Dataset 2

Dataset 2 is an evaluation dataset containing 100 strong binders experimentally identified and curated in related work[10] along with 2,000 additional pHLA decoys extracted from the NetMHC dataset[11]. Selected pHLA complexes have no overlap with the training set (which is a subset of Dataset 1) and were modeled with APE-Gen using the methodology proposed in the reference study[28]. Dataset 2 was composed to mimic an epitope discovery setting where a large pool of peptide targets is screened, but only a small portion of the targets are true binders.

Dataset 3

Dataset 3 is an evaluation set containing 11 pHLA complexes for the HLA-A0201 allele with different levels of known experimental binding affinity (strong [0-5] nM, medium [50-500] nM and weak [500-25,000] nM) for which there exist crystal structures in the PDB. Three out of 11 peptides are 10-mers while the others are 9-mers. We collected crystal structures for each of the pHLA complexes (note that there were multiple entries for some complex complexes, see Supplementary Table S4). Multiple biological assemblies sometimes with alternative side chain positions were extracted from each PDB file and treated as separate structures. This led to the inclusion of 77 structures in Dataset 3. Preprocessing of the crystals was done using PyMol[52] (to remove water molecules and hydrogen atoms) and pdbfixer[53] (to add missing atoms). Since crystal structures of complexes involving non-binder peptides do not exist, five additional structures of experimentally determined non-binding peptides[50] for the HLA-A0201 allele were modeled with Docktope[54] and added to Dataset 3. The complete dataset is outlined in Supplementary Table S4; it contains 82 structures of pHLA complexes involving 16 peptides and the HLA-A0201 receptor. These pHLA complexes do not appear in the training set (which is a subset of Dataset 1). Dataset 3 is a good test of the generalizability of 3pHLA-score because it strongly differs from the training dataset - structures are not modeled by APE-Gen and some involve peptides of length 10.

Comparison of scoring functions

Several evaluation metrics were used to compare the performance of scoring functions (see Supplementary Material section Evaluation Metrics). Because we focused on assessing how well the functions could reproduce peptide rankings in terms of HLA-binding affinity, we used Pearson's correlation coefficient r and Spearman's correlation coefficient ρ to evaluate the regression performance. To assess classification power, we used the Area Under the Receiver Operator Curve (AUROC) and the Area Under the Precision-Recall Curve (AUPRC). The performance of 3pHLA-score on Dataset 2 and Dataset 3 was compared to other widely used scoring functions which use different techniques (Table S5). When visualized, scores were scaled using max normalization to fit [0-1] range, but inverted such that values closer to 1 represent stronger binders for all investigated scoring functions, while values closer to 0 represent weaker binders.

Results

We investigate the benefits of our per-peptide-position protocol by assessing the predictive power of 3pHLA-score on the test portion of Dataset 1 (see Methods subsection Dataset 1). We then compare the performance of the 3pHLA-score to six other widely used scoring functions in two different settings using independent datasets: Dataset 2 and Dataset 3.

Per-peptide-position featurization shows superior predictive power

First, we compare the regression and classification power of the following scoring functions on the test portion of Dataset 1: ref2015-score, standard-pHLA-score, and 3pHLA-score. We are interested to see how well the rank of predicted binding affinities matches the rank of the true binding affinity values for tested pHLA complexes. On the other hand, with the classification metrics (AUROC, AUPRC), we want to test how well predicted binding affinities separate the known binders from non-binders. The regression power of the scoring functions is evaluated on the test set using Pearson's correlation coefficient r (Figure 2). 3pHLA-score outperformed both ref-2015 and standard-pHLA-score: while 3pHLA-score achieves an average Pearson's correlation of 0.75 on the test set, ref2015-score and standard-pHLA-score achieve a significantly lower correlation of 0.09 and 0.46, respectively (Table 1). Figure S3 shows in detail the correlation between predicted and experimental scores for the best and worst performing 3pHLA-score models.

The same pattern is observed for all individual HLA alleles across all investigated metrics (Figure S2, Table S2). Additionally, we provide the same analysis for standard-pHLA-score and 3pHLA-score that are trained using alternative ML regression techniques (Supplementary Material subsection *Alternative ML regression techniques*). 3pHLA-score consistently outperforms standard-pHLA-score across all ML regression techniques we assessed.

The predictive power of the per-peptide-position protocol varies depending on the choice of positions

We know that different residue positions in a peptide (i.e., peptide positions) have different contributions to HLA binding and T-cell recognition. While middle positions are usually more exposed and therefore involved in the recognition by T-cells, the anchor positions are usually buried in the HLA groove and play a more direct role in pHLA binding[55]. For this reason, we conducted an ablation study to investigate the influence of different peptide positions on the performance of 3pHLA-score. 3pHLA-score was trained with three different position sets: all nine positions, anchor positions (1, 2, 3, 8, 9) or middle positions (4, 5, 6, 7). We generate binding affinity predictions for the test set using these different versions of 3pHLA-score and we investigate how well the affinities are ranked compared to the true affinities as well as how well the predictions separate binders from non-binders. We observe that the choice of positions in 3pHLA-score has a substantial influence its performance on the test set according to Pearson's correlation (Figure 3 , Table S3) and other metrics (Figure S4). The performance of training with

anchor positions only and all nine positions is comparable, with r values higher than 0.8 for most HLA alleles. The r values drop below 0.7 when middle positions only are used. The only exception is the HLA-B0801 allele, for which closer inspection of the binding motif in IEDB (iedb.org/mhc/252) clearly indicates the importance of position 5 for peptide binding, as reflected in the HLA-B0801 predictor's performance.

3pHLA-score outperforms well-validated structure-based scoring functions in an epitope discovery setting

The goal in structure-based virtual screening for epitope discovery is to distinguish true binder peptides from non-binders, which can be seen as a classification problem. To evaluate 3pHLA-score in an epitope discovery scenario, we compare it to a variety of widely used structural scoring functions (Table S5) on a dataset containing 100 strong binders and 2,000 decoys across 16 HLA alleles (Dataset 2). Our results show that 3pHLA-score clearly outperforms other evaluated scoring functions in this virtual screening setting, with an average AUPRC of 0.71 compared to the second best scoring function (Vinardo) with AUPRC of 0.35 (Table 2). This is consistent with 3pHLA-score achieving higher values of both AUROC and AUPRC for all investigated HLA alleles individually (Tables S7, S8), and 3pHLA-score separating binders from non-binders more clearly than other scoring functions (Figure S5). It is also important to note that Dataset 2 was not used in the training phase. Therefore, this experiment also demonstrates the capacity of 3pHLA-score to generalize to new datasets.

In the context of epitope discovery, current pipelines use sequence-based scoring functions. Therefore, we evaluate how 3pHLA-score compares to sequence-based methods and present the details of this analysis in Supplementary Material. Overall, 3pHLA-score has comparable performance to selected sequence-based methods with average AUROC of 0.977 compared to the best achieved AUROC of 0.993 with MHCFlurry2.0[10]. Note that we do not know if MHCFlurry2.0 has had a part of our test dataset in their training, which might give it a slight advantage.

3pHLA-score can generalize to an independent dataset

We tested the ability of 3pHLA-score to generalize to other "types" of structural data with the independent Dataset 3. Dataset 1 and Dataset 2 contain structures that were all modeled by APE-Gen[28] with peptides of 9 residues in length (9-mers). With an independent dataset we can investigate the possible biases towards this modeling tool and explore how to generalize to the peptides of length 10. Dataset 3 contains experimentally resolved three-dimensional pHLA structures involving binders and non-binders modeled by Docktope[54]. Importantly, it contains 10-mer peptides. As 3pHLA-score was trained on 9-mers, the size of the input of the model is 9×19 (i.e., 9 peptide positions times 19 energy terms). To score 10-mers, we excluded the energy terms of the middle position (i.e., position 6) of the peptide. The rationale for this approach lies in the aforementioned experimental findings on peptide anchors[46].

Since Dataset 3 contains peptides with a wide range of experimental binding affinities (strong, medium, weak binders, and non-binders), two tasks were identified for the scoring functions: a regression and a classification task. For the regression task, scoring functions are expected to predict the correct peptide ranking in terms of binding affinities. In this context it is also interesting to analyze the range of scores predicted for a given peptide within different structures (i.e., same complex, but different crystallography experiments). The smaller the range, the more consistent a scoring function is for scoring a certain peptide. For the classification task, we label peptides with three different binding affinity thresholds: 50 nM (distinguishing strong binders from others), 500 nM (distinguishing strong and medium binders from others), and 25,000 nM (distinguishing binders from non-binders). The classification power of scoring functions was evaluated using AUROC and AUPRC.

The scaled scores aggregated across structures for each peptide are shown in Figure 4. The scaled score for each structure in Dataset 3 is shown in Figure S6. Pearson's correlation coefficient between experimental binding affinity and predicted scores is given in Table S6. While DOPE scoring function consistently outperforms others, 3pHLA-score shows competitive performance in this challenging setting and is a runner-up in most of the evaluated tasks. In the regression setting this fact is reflected by DOPE achieving a correlation of 0.62, while 3pHLA-score achieves a correlation of 0.56 with experimental affinity. However, neither of these correlations are strong. On the other hand, both DOPE and 3pHLA-score produce small variations of the score for different structures of the same peptide which is a desirable property for an epitope discovery task. With respect to the classification task, DOPE produced the best results according to AUROC and AUPRC for most of the thresholds analyzed (Table 3). The 3pHLA-score also occupied a position of relevance, having the best AUPRC value for the 500 nM threshold and the second best AUPRC values for the 50 nM and 25,000 nM thresholds. When considering AUROC values, Vina and Vinardo are the second best for the 50 nM and 500 nM thresholds; the 3pHLA-score was again the second best for the 25,000 nM threshold.

Discussion

Motivated by experimental findings of peptide anchors, we hypothesize that important information for training ML pHLA scoring functions is lost in standard training protocols. We try to recover this information using our novel per-peptide-position protocol and we apply it to develop the 3pHLA-score.

In the first set of experiments we show how energy decoupling of the per-peptide-position protocol (as applied to 3pHLA-score) significantly increases predictive power of models (Figure 2, Figure S2, Table 1). Furthermore, we show that the predictive power of 3pHLA-score is highly dependent on the choice of peptide positions to be decoupled (Figure 3, Figure S4).

Next, we provide extensive comparison of the 3pHLA-score against other widely used scoring functions. 3pHLA-score shows a clear superior performance to other scoring functions when tested in the epitope discovery setting where we perform structure-based virtual screening of true peptide-binders to HLA receptors (Table 2, Figure S5).

Note that the training of the 3pHLA-score could not have been done using only experimentally-determined crystal structures, due to the limited number of pHLA crystals available (i.e., less than 800 in the PDB). Therefore, we chose to use models produced by APE-Gen, which is potentially the only currently available pHLA-specific modeling tool with the capacity to model thousands of complexes (e.g., nearly 80,000 complexes modeled for Dataset 1). The choice of the modeling method, however, can introduce a bias in the training of the scoring function. To test that, we used an independent dataset (i.e., Dataset 3) containing crystal structures and models produced by a different tool DockTope. Note that DockTope uses a very different modeling protocol, based on fixed backbone templates. Despite involving different types of structures, our results still show a good overall performance of 3pHLA-score on Dataset 3, being competitive with other popular scoring functions. These results suggest that 3pHLA-score can be used with crystal structures and models produced by other tools, without additional training, although a broader survey with other tools for pHLA modeling and peptide-docking will be needed to further corroborate this point. Interestingly, in this experiment the most consistent predictions across different structures of the same complex, and the strongest correlation with experimental data, were observed for DOPE (Table 3, Figure 4). This surprising result might be directly linked to the nature of this dataset and the intended use of DOPE. DOPE scoring function is a statistical potential used to assess the global quality of homology models produced by Modeller[56]. This provides two advantages to DOPE in the experiment with Dataset 3. First, this dataset is mostly composed of crystal structures, and DOPE's global assessment was observed in our experiment to be more resilient to small differences between different conformations of the same complex. Second, DOPE is well suited to distinguish the non-binders, which were modeled with a docking-based approach, from the experimentally-determined crystal structures used for all other complexes. Our results show that the 3pHLA-score predictions could be generalized to both DockTope models and crystal structures, while the good performance of DOPE did not generalize to other datasets. For instance, 3pHLA-score outperformed DOPE and other scoring functions on Dataset 2 (Table 2, Figure S5). It is therefore the method that provides the most consistent results across the three different datasets.

The discovered potential of per-peptide-position energy terms for pHLA system opens up many additional opportunities that we discuss here. To build 3pHLA-score we trained separate models for each HLA allele. This limits the use of 3pHLA-score to a fixed set of HLA alleles that is found in the training dataset. However, a bigger pan-allele dataset can be acquired in the future and the same method could be applied to train a more general pan-allele model. APE-Gen, the tool used here to model pHLA structures, is currently limited to modeling the peptides containing only the 20 standard amino acids. Therefore, modeling phosphorylated peptides (or peptides with other post-translational modifications) and assessing the HLA-binding energies of these peptides with 3pHLA-score is another interesting challenge, which would greatly broaden the impact of our methods to ongoing efforts in epitope discovery[57]. 3pHLA-score was trained here with a single conformation per peptide, to predict HLA binding affinity in the context of structural virtual screening. Future studies could investigate the use and refinement of 3pHLA-score to the geometry prediction task (i.e., ranking different conformations of the same pHLA complex). For that task we would propose using the same per-peptide-position training protocol on a dataset that contains multiple conformations per peptide mapped to a corresponding experimentally determined crystal structures. The baseline scoring function for extracting the energy terms used here was ref2015. Therefore, it remains to be determined how the same training protocol would perform when applied to another existing scoring function which provides energy terms for specific regions of the model. This question is left for future work. As discussed above, our per-peptide-position protocol could provide more opportunities than exemplified by 3pHLA-score. The protocol can be applied beyond the ref2015 energy terms as well as beyond the pHLA system. For that reason we make a distinction between the 3pHLA-score and the per-peptide-position protocol.

Overall, our results confirm that important structural signal for binding prediction gets lost when the standard energy terms are calculated at the all-peptide-atom level. This could point to the fact that the additive nature of the standard all-atom energy terms is not appropriate for the pHLA system. Our work emphasizes how experimental findings can help engineer more powerful features and train ML models with better predictive power. This can serve as a guideline for future attempts of training custom ML scoring functions for different systems of interest. As more structural pHLA data become available, we hope that our findings will inspire future efforts in training structure-based pHLA binding predictors that could enter epitope discovery pipelines and complement sequence-based methods. 3pHLA-score has direct application to epitope discovery projects, which could help advance the development of vaccines against several types of cancer and viral infections.

References

1. Neefjes, J., Jongsma, M. L. M., Paul, P. & Bakke, O. Towards a systems understanding of MHC class i and MHC class II antigen presentation. *Nat. Rev. Immunol.* **11**, 823–836, DOI: 10.1038/nri3084 (2011).
2. Rock, K. L., Reits, E. & Neefjes, J. Present yourself! by MHC class i and MHC class II molecules. *Trends Immunol.* **37**, 724–737, DOI: 10.1016/j.it.2016.08.010 (2016).
3. Stevanović, S. Structural basis of immunogenicity. *Transpl. Immunol.* **10**, 133–136, DOI: 10.1016/s0966-3274(02)00059-x (2002).
4. James, K. D., Jenkinson, W. E. & Anderson, G. T-cell egress from the thymus: Should i stay or should i go? *J. Leukoc. Biol.* **104**, 275–284, DOI: 10.1002/jlb.1mr1217-496r (2018).
5. Grau, M., Walker, P. R. & Derouazi, M. Mechanistic insights into the efficacy of cell penetrating peptide-based cancer vaccines. *Cell. Mol. Life Sci.* **75**, 2887–2896, DOI: 10.1007/s00018-018-2785-0 (2018).
6. Lizée, G. *et al.* Harnessing the power of the immune system to target cancer. *Annu. Rev. Medicine* **64**, 71–90, DOI: 10.1146/annurev-med-112311-083918 (2013).
7. Dudek, N. L., Perlmutter, P., Aguilar, M.-I., Croft, N. P. & Purcell, A. W. Epitope discovery and their use in peptide based vaccines. *Curr. Pharm. Des.* **16**, 3149–3157, DOI: 10.2174/138161210793292447 (2010).
8. Joglekar, A. V. & Li, G. T cell antigen discovery. *Nat. Methods* **18**, 873–880, DOI: 10.1038/s41592-020-0867-z (2020).
9. Robinson, J. *et al.* The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res.* **43**, D423–D431, DOI: 10.1093/nar/gku1161 (2014).
10. O'Donnell, T. J., Rubinsteyn, A. & Laserson, U. MHCflurry 2.0: Improved pan-allele prediction of MHC class I-presented peptides by incorporating antigen processing. *Cell Syst.* **11**, 42–48, DOI: 10.1016/j.cels.2020.06.010 (2020).
11. Andreatta, M. & Nielsen, M. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics* **32**, 511–517, DOI: 10.1093/bioinformatics/btv639 (2015).
12. O'Donnell, T. J. *et al.* MHCflurry: Open-source class I MHC binding affinity prediction. *Cell Syst.* **7**, 129–132, DOI: 10.1016/j.cels.2018.05.014 (2018).
13. Zhang, H., Lund, O. & Nielsen, M. The PickPocket method for predicting binding specificities for receptors based on receptor pocket similarities: application to MHC-peptide binding. *Bioinformatics* **25**, 1293–1299, DOI: 10.1093/bioinformatics/btp137 (2009).
14. Vielhaben, J., Wenzel, M., Samek, W. & Strodthoff, N. USMPep: universal sequence models for major histocompatibility complex binding affinity prediction. *BMC Bioinforma.* **21**, DOI: 10.1186/s12859-020-03631-1 (2020).
15. Venkatesh, G., Grover, A., Srinivasaraghavan, G. & Rao, S. MHCAttnNet: predicting MHC-peptide bindings for MHC alleles classes I and II using an attention-based deep neural model. *Bioinformatics* **36**, i399–i406, DOI: 10.1093/bioinformatics/btaa479 (2020).
16. Zhao, W. & Sher, X. Systematically benchmarking peptide-MHC binding predictors: From synthetic to naturally processed epitopes. *PLOS Comput. Biol.* **14**, e1006457, DOI: 10.1371/journal.pcbi.1006457 (2018).
17. Alpízar, A. *et al.* A molecular basis for the presentation of phosphorylated peptides by HLA-b antigens. *Mol. & Cell. Proteomics* **16**, 181–193, DOI: 10.1074/mcp.m116.063800 (2017).
18. Refsgaard, C. T., Barra, C., Peng, X., Ternette, N. & Nielsen, M. NetMHCphosPan - pan-specific prediction of MHC class I antigen presentation of phosphorylated ligands. *ImmunoInformatics* **1-2**, 100005, DOI: 10.1016/j.immuno.2021.100005 (2021).
19. Koch, C. P., Pillong, M., Hiss, J. A. & Schneider, G. Computational resources for MHC ligand identification. *Mol. Informatics* **32**, 326–336, DOI: 10.1002/minf.201300042 (2013).
20. Young, S. S., Yuan, F. & Zhu, M. Chemical descriptors are more important than learning algorithms for modelling. *Mol. Informatics* **31**, 707–710, DOI: 10.1002/minf.201200031 (2012).
21. Liao, W. W. P. & Arthur, J. W. Predicting peptide binding affinities to MHC molecules using a modified semi-empirical scoring function. *PLoS ONE* **6**, e25055, DOI: 10.1371/journal.pone.0025055 (2011).
22. Antunes, D. A., Abella, J. R., Devaurs, D., Rigo, M. M. & Kavraki, L. E. Structure-based methods for binding mode and binding affinity prediction for peptide-MHC complexes. *Curr. Top. Medicinal Chem.* **18**, 2239–2255, DOI: 10.2174/1568026619666181224101744 (2019).

Appeared in Scientific Reports, 2022
https://doi.org/10.1038/s41598-022-14526-x

23. Aranha, M. P. *et al.* Combining three-dimensional modeling with artificial intelligence to increase specificity and precision in peptide–MHC binding predictions. *The J. Immunol.* **205**, 1962–1977, DOI: 10.4049/jimmunol.1900918 (2020).

24. Devaurs, D. *et al.* Using parallelized incremental meta-docking can solve the conformational sampling issue when docking large ligands to proteins. *BMC Mol. Cell Biol.* **20**, DOI: 10.1186/s12860-019-0218-z (2019).

25. Palacio-Rodríguez, K., Lans, I., Cavasotto, C. N. & Cossio, P. Exponential consensus ranking improves the outcome in docking and receptor ensemble docking. *Sci. Reports* **9**, DOI: 10.1038/s41598-019-41594-3 (2019).

26. Guedes, I. A. *et al.* New machine learning and physics-based scoring functions for drug discovery. *Sci. Reports* **11**, DOI: 10.1038/s41598-021-82410-1 (2021).

27. Ain, Q. U., Aleksandrova, A., Roessler, F. D. & Ballester, P. J. Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **5**, 405–424, DOI: 10.1002/wcms.1225 (2015).

28. Abella, J., Antunes, D., Clementi, C. & Kavraki, L. APE-gen: A fast method for generating ensembles of bound peptide-MHC conformations. *Molecules* **24**, 881, DOI: 10.3390/molecules24050881 (2019).

29. Alford, R. F. *et al.* The rosetta all-atom energy function for macromolecular modeling and design. *J. Chem. Theory Comput.* **13**, 3031–3048, DOI: 10.1021/acs.jctc.7b00125 (2017).

30. Schulz-Gasch, T. & Stahl, M. Scoring functions for protein–ligand interactions: a critical perspective. *Drug Discov. Today: Technol.* **1**, 231–239, DOI: 10.1016/j.ddtec.2004.08.004 (2004).

31. Kyeong, H. H., Choi, Y. & Kim, H. S. GradDock: rapid simulation and tailored ranking functions for peptide-MHC class I docking. *Bioinformatics* **34**, 469–476, DOI: 10.1093/bioinformatics/btx589 (2017).

32. Li, H., Leung, K.-S., Wong, M.-H. & Ballester, P. J. Improving AutoDock Vina using random forest: The growing accuracy of binding affinity prediction by the effective exploitation of larger data sets. *Mol. Informatics* **34**, 115–126, DOI: 10.1002/minf.201400132 (2015).

33. Afifi, K. & Al-Sadek, A. F. Improving classical scoring functions using random forest: The non-additivity of free energy terms' contributions in binding. *Chem. Biol. & Drug Des.* **92**, 1429–1434, DOI: 10.1111/cbdd.13206 (2018).

34. Yasuo, N. & Sekijima, M. Improved method of structure-based virtual screening via interaction-energy-based learning. *J. Chem. Inf. Model.* **59**, 1050–1061, DOI: 10.1021/acs.jcim.8b00673 (2019).

35. Zhou, P. *et al.* Systematic comparison and comprehensive evaluation of 80 amino acid descriptors in peptide QSAR modeling. *J. Chem. Inf. Model.* **61**, 1718–1731, DOI: 10.1021/acs.jcim.0c01370 (2021).

36. Guan, P., Doytchinova, I. A., Walshe, V. A., Borrow, P. & Flower, D. R. Analysis of peptide-protein binding using amino acid descriptors: Prediction and experimental verification for human histocompatibility complex HLA-A*0201. *J. Medicinal Chem.* **48**, 7418–7425, DOI: 10.1021/jm0505258 (2005).

37. Leaver-Fay, A. *et al.* Chapter nineteen - Rosetta3: An object-oriented software suite for the simulation and design of macromolecules. In Johnson, M. L. & Brand, L. (eds.) *Computer Methods, Part C*, vol. 487 of *Methods in Enzymology*, 545–574, DOI: https://doi.org/10.1016/B978-0-12-381270-4.00019-6 (Academic Press, 2011).

38. Morris, G. M. *et al.* AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.* **30**, 2785–2791, DOI: 10.1002/jcc.21256 (2009).

39. Trott, O. & Olson, A. J. AutoDock vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* NA–NA, DOI: 10.1002/jcc.21334 (2009).

40. Quiroga, R. & Villarreal, M. A. Vinardo: A scoring function based on autodock vina improves scoring, docking, and virtual screening. *PLOS ONE* **11**, e0155183, DOI: 10.1371/journal.pone.0155183 (2016).

41. Shen, M.-Y. & Sali, A. Statistical potential for assessment and prediction of protein structures. *Protein Sci.* **15**, 2507–2524, DOI: 10.1110/ps.062416606 (2006).

42. Schymkowitz, J. *et al.* The FoldX web server: an online force field. *Nucleic Acids Res.* **33**, W382–W388, DOI: 10.1093/nar/gki387 (2005).

43. Berman, H. M. The protein data bank. *Nucleic Acids Res.* **28**, 235–242, DOI: 10.1093/nar/28.1.235 (2000).

44. Borrman, T., Pierce, B. G., Vreven, T., Baker, B. M. & Weng, Z. High-throughput modeling and scoring of TCR-pMHC complexes to predict cross-reactive peptides. *Bioinformatics* **36**, 5377–5385, DOI: 10.1093/bioinformatics/btaa1050 (2020).

Appeared in Scientific Reports, 2022
https://doi.org/10.1038/s41598-022-14526-x

45. Ye, W.-L. *et al.* Improving docking-based virtual screening ability by integrating multiple energy auxiliary terms from molecular docking scoring. *J. Chem. Inf. Model.* **60**, 4216–4230, DOI: 10.1021/acs.jcim.9b00977 (2020).

46. Bouvier, M. & Wiley, D. Importance of peptide amino and carboxyl termini to the stability of MHC class I molecules. *Science* **265**, 398–402, DOI: 10.1126/science.8023162 (1994).

47. Chaudhury, S., Lyskov, S. & Gray, J. J. PyRosetta: a script-based interface for implementing molecular modeling algorithms using rosetta. *Bioinformatics* **26**, 689–691, DOI: 10.1093/bioinformatics/btq007 (2010).

48. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32, DOI: 10.1023/a:1010933404324 (2001).

49. Breiman, L., Friedman, J., Stone, C. J. & Olshen, R. *Classification and Regression Trees* (Chapman and Hall/CRC, 1984).

50. Abella, J. R., Antunes, D. A., Clementi, C. & Kavraki, L. E. Large-scale structure-based prediction of stable peptide binding to class I HLAs using random forests. *Front. Immunol.* **11**, DOI: 10.3389/fimmu.2020.01583 (2020).

51. Vita, R. *et al.* The immune epitope database (IEDB): 2018 update. *Nucleic Acids Res.* **47**, D339–D343, DOI: 10.1093/nar/gky1006 (2018).

52. Schrödinger, LLC. The PyMOL molecular graphics system, version 1.8 (2015).

53. Eastman, P. *et al.* Openmm 4: A reusable, extensible, hardware independent library for high performance molecular simulation. *J. Chem. Theory Comput.* **9**, 461–469, DOI: 10.1021/ct300857j (2013).

54. Rigo, M. M. *et al.* DockTope: a web-based tool for automated pMHC-i modelling. *Sci. Reports* **5**, DOI: 10.1038/srep18413 (2015).

55. Achour, A. Major histocompatibility complex: Interaction with peptides. *eLS* DOI: https://doi.org/10.1038/npg.els.0000922 (2001). https://onlinelibrary.wiley.com/doi/pdf/10.1038/npg.els.0000922.

56. Šali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815, DOI: 10.1006/jmbi.1993.1626 (1993).

57. Alpízar, A. *et al.* A Molecular Basis for the Presentation of Phosphorylated Peptides by HLA-B Antigens. *Mol Cell Proteomics* **16**, 181–193 (2017).

58. Reynisson, B., Alvarez, B., Paul, S., Peters, B. & Nielsen, M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res.* **48**, W449–W454, DOI: 10.1093/nar/gkaa379 (2020).

Acknowledgements

The authors would like to thank Dr. Jayvee Abella for inspiring this work and making the APE-Gen datasets available and Romanos Fasoulis for his help with running the sequence-based scoring as well as other colleagues from Kavraki Lab for many helpful discussions.

Work on this project by A.C. and L.E.K. have been supported in part by National Institutes of Health NIH [U01CA258512]. Other support included: University of Edinburgh and Medical Research Council [MC_UU_00009/2 to D.D.]; Computational Cancer Biology Training Program fellowship [RP170593 to M.M.R.]; University of Houston Funds and Rice University Funds.

Author contributions statement

A.C., D.D., M.M.R. and D.A.A. conceived the experiment(s), A.C. conducted the experiment(s) and wrote the manuscript, A.C., D.D., M.M.R., D.A.A. and L.E.K. analyzed the results. All authors reviewed the manuscript.

Availability of materials and data

Sequencing data was not generated in this study.
The code and the data used for running the experiments and training along with the scoring function and datasets is available in the repository: *https://github.com/anon528/1*.

Competing interests

The author(s) declare no competing interests.

Ethics declarations

No human or animal data samples were used in this study.

Tables

Table 1. Results of scoring functions obtained using different training protocols on the test set averaged across all HLA alleles for all four evaluated metrics (Pearson's correlation coefficient $|r|$, Spearman's correlation coefficient $|\rho|$, the Area Under the Receiver Operator Curve AUROC and the Area Under the Precision Recall Curve AUPRC) . The highest values and best performing values in each column are bolded.

	$\|r\|$	$\|\rho\|$	AUROC	AUPRC
3pHLA-score	**0.75**	**0.90**	**0.98**	**0.99**
standard-pHLA-score	0.46	0.50	0.80	0.82
ref2015-score	0.09	0.07	0.44	0.56

Table 2. AUROC and AUPRC values aggregated for the virtual screening experiment across HLA alleles. The highest values and best performing values in each column are bolded.

	AUROC	AUPRC
3pHLA-score	**0.977**	**0.712**
Vinardo[40]	0.898	0.354
Vina[39]	0.871	0.291
GradDock[31]	0.778	0.182
DOPE[41]	0.769	0.141
AutoDock4[38]	0.751	0.141
FoldX[42]	0.687	0.142

Table 3. Quantified power of scoring functions to discriminate between peptides of different binding strength on the Dataset 3.

	thr = 50 nM		thr = 500 nM		thr = 25,000 nM	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
DOPE	**1.0**	**1.0**	**0.84**	**0.69**	**1.0**	**0.97**
3pHLA-score	0.76	**0.85**	0.76	**0.72**	**0.99**	**0.91**
Vina	0.84	0.71	**0.81**	0.52	0.90	0.27
Vinardo	**0.85**	0.73	0.78	0.49	0.87	0.23
FoldX	0.83	0.74	0.70	0.41	0.83	0.19
AutoDock4	0.73	0.63	0.63	0.35	0.82	0.17
GradDock	0.42	0.48	0.48	0.33	0.16	0.04

* Top 2 performing values are bolded.
** thr: threshold of binding affinity used to label different classes.

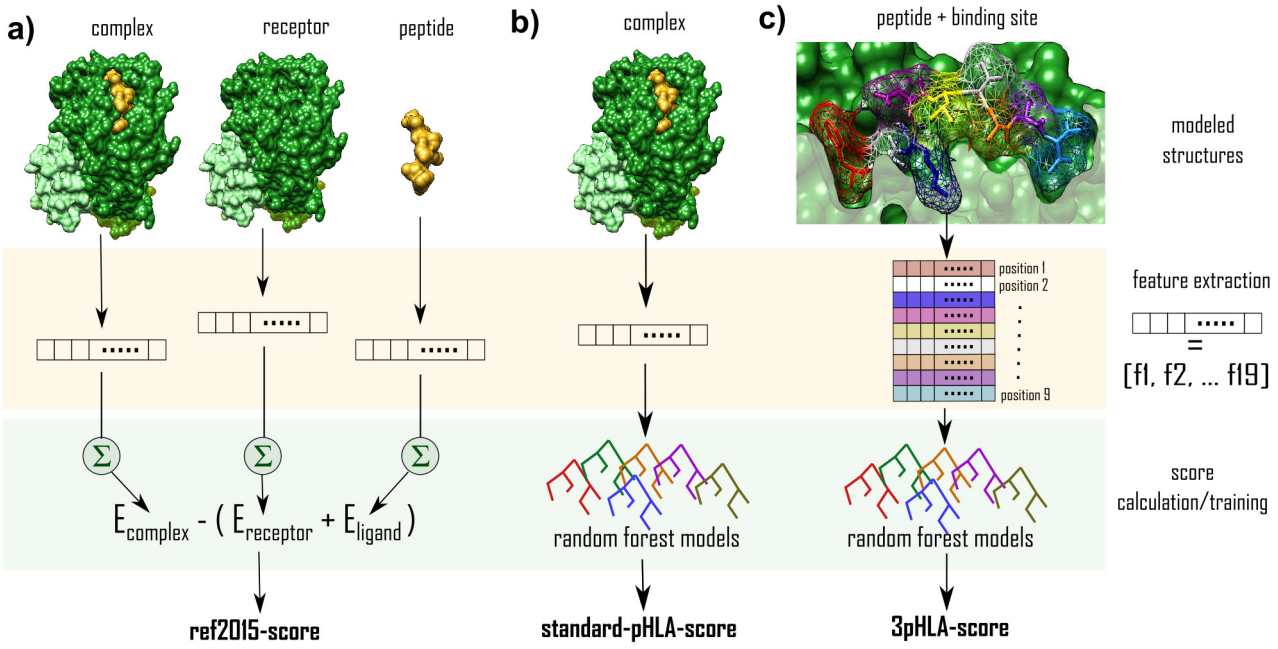

Figure 1. Description of three different protocols for approximating the binding affinity of a pHLA complex. Example input structures are visualized in the first row. The second row (orange stripe) shows the feature extraction phase of the scoring where ref2015 energy terms are extracted (Supplementary Table S1). The score calculation and training phase is indicated in row 3 (green stripe). **(a)** For ref2015-score, standard ref2015 energies are calculated for the complex, receptor, and ligand. They are then used to derive the binding energy with the Equation 1 **(b)** For standard-pHLA-score, standard features are extracted from the complex; scoring is done using trained random forest regression models. **(c)** For the 3pHLA-score, per-peptide-position features are extracted from the structure of the complex; scoring is done using trained random forest regression models.

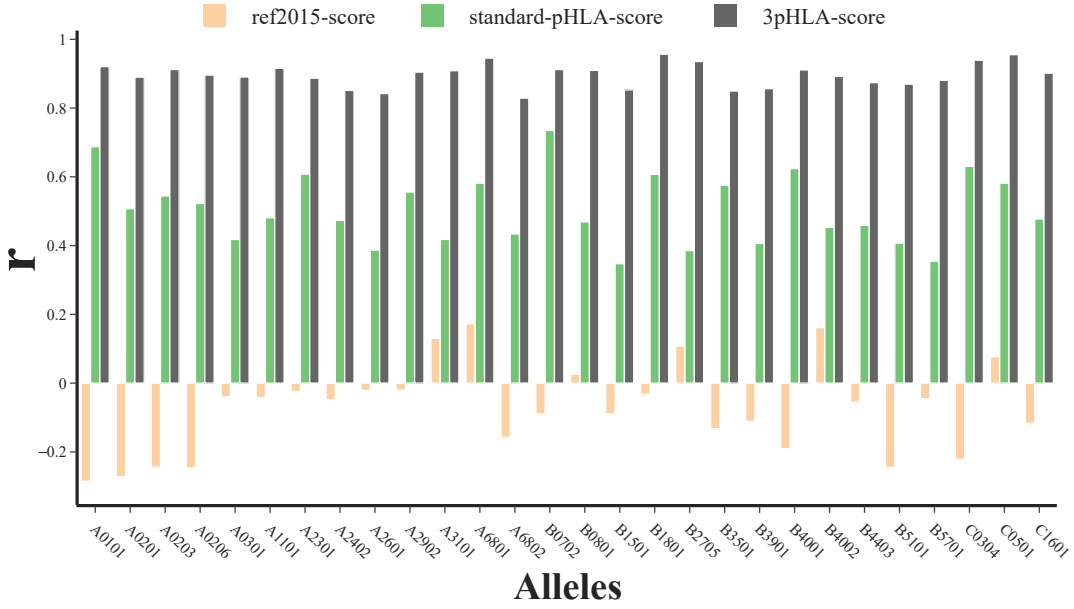

Figure 2. The predictive power of ref2015-score, standard-pHLA-score, and 3pHLA-score is evaluated and compared on the test portion of Dataset 1. Results are reported for individual alleles, listed on the x-axis. The regression power of the scores is quantified using Pearson's *r*, on the y-axis.

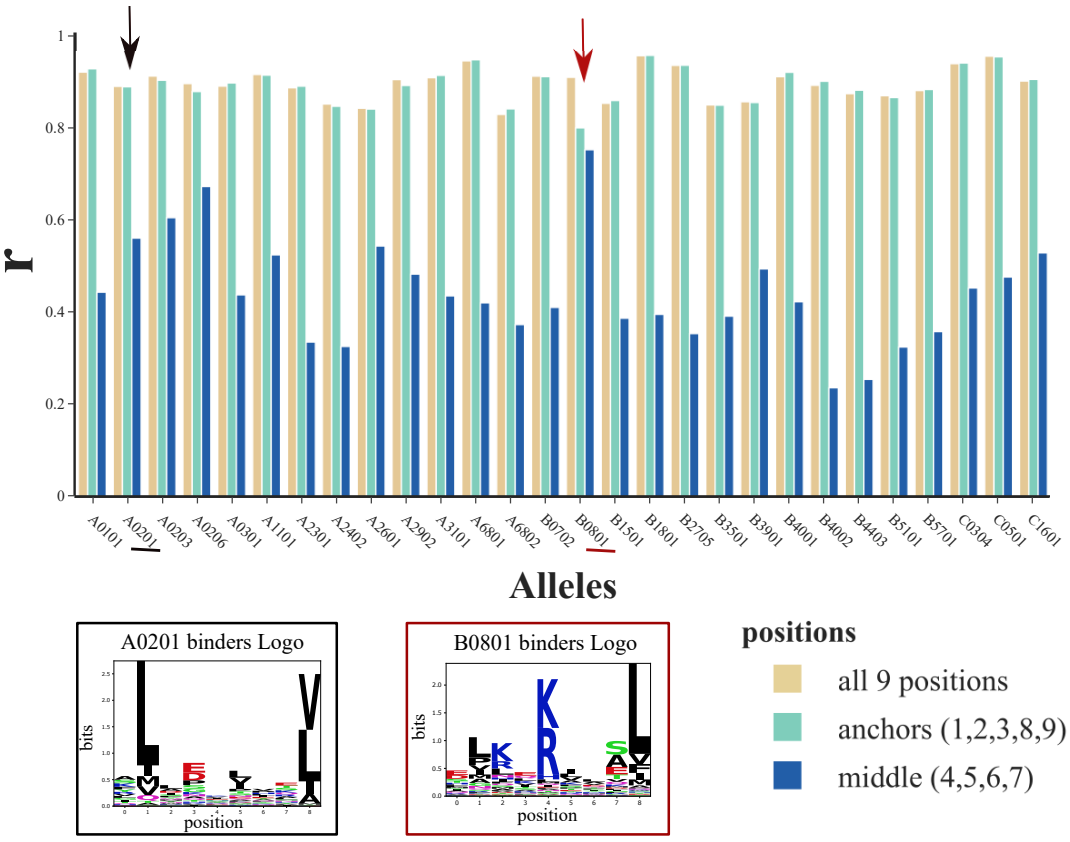

Figure 3. Results of the ablation study, in which 3pHLA-score was trained using different subsets of peptide positions: all nine positions, anchor positions (1, 2, 3, 8, 9) and middle positions (4, 5, 6, 7). Results are reported for individual alleles indicated on the x-axis. The regression power of scoring functions is quantified using Pearson's *r* and plotted on the y-axis. The logo representation of the HLA-A0201 and HLA-B0801 binders is presented to compare the importance of the middle position 5 for HLA-B0801.

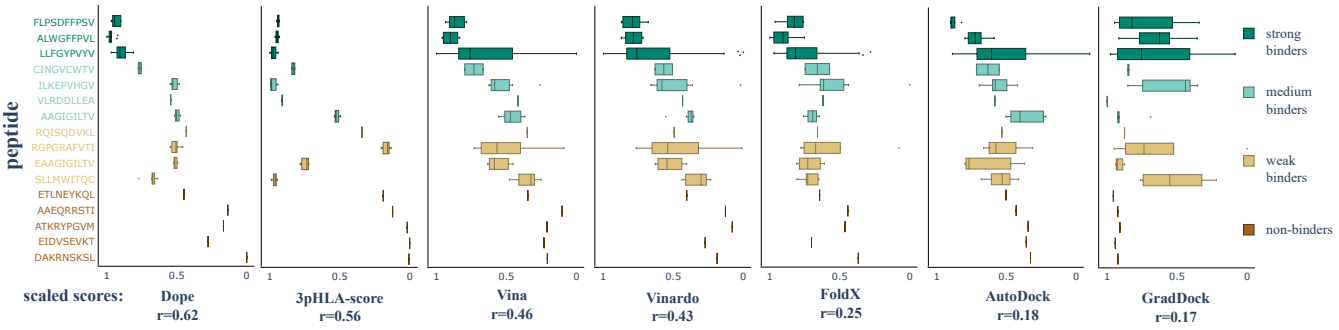

Figure 4. Performance of different scoring functions in evaluating the binding affinity of structures from the independent Dataset 3. Pearson's correlation coefficient is indicated next to the name of the scoring function. Peptides involved in the structures of Dataset 3 (see Table S4) are listed on the y-axis. The peptide names and corresponding box plots are colored and arranged along the y-axis according to their experimental binding affinity (ranging from dark green, strong binders, at the top, to dark orange, non-binders, at the bottom). Predicted scores scaled to the range 1-0 are plotted on the x-axis (1-highest predicted binder; 0-non-binder). The correlation is calculated for the predicted binding affinity of each of the 82 structures present in Dataset 3 with respect to their experimental binding affinities.

3pHLA-score improves structure-based peptide-HLA binding affinity prediction

Anja Conev, Didier Devaurs, Mauricio Menagatti Rigo, Dinler Amaral Antunes, Lydia Kavraki

Supplementary material

Peptide-HLA structure terminology

The HLA class I receptor is a heterodimer composed of a heavy chain (α) and a light chain ($beta_2$-microglobulin) (depicted in Figure S1a). The light chain is not highly variable and is not encoded by the HLA gene. The heavy chain is highly variable and contains three regions (α1, α2, α3). Regions a1 and a2 form the binding site. The binding site is the groove between a1 and a2 helices where peptides bind to HLAs and form the peptide-HLA complex (Figure S1b). Within the binding site of class I HLA receptors, six smaller pockets are defined. These pockets are labeled A to F, with pockets B and F accommodating the so-called anchor positions at the N-terminus and C-terminus of the peptide, respectively[55] (Figure S1c). Peptides that bind to HLA class I molecules are usually 9-11 amino acids long. pHLA binding is conditioned by the structural and energy fit of peptides to the binding site.

Evaluation metrics

To assess the regression power of scoring functions, the following metrics are used:

- Pearson's correlation coefficient r

$$r = \frac{cov(Y, Y')}{\sigma_Y \sigma_{Y'}} \tag{S.1}$$

  where Y are the observed values and Y' the models' predictions, cov is covariance and σ is the standard deviation. r quantifies the linear relationship between the observed and predicted values. The observed values in our case are the experimental binding affinities of the pHLA complex that are considered as labels in the dataset. The predicted values are values given by different scores that we evaluate (3pHLA-score, standard-pHLA-score, Vina, Vinardo). r ranges from -1 to 1 and the relationship is considered to be strong when the absolute value $|r|$ is above 0.7.

- Spearman's correlation coefficient ρ

$$\rho = \frac{cov(rk_Y, rk_{Y'})}{\sigma_{rk_Y} \sigma_{rk_{Y'}}} \tag{S.2}$$

  where rk_Y are the ranks of the observed values and $rk_{Y'}$ the ranks of predictions, cov is covariance and σ is the standard deviation. ρ quantifies the monotonic relationship between the observed and predicted values. The observed values in our case are the experimental binding affinities of the pHLA complex that are considered as labels in the dataset. The predicted values are values given by different scores that we evaluate (3pHLA-score, standard-pHLA-score, Vina, Vinardo). ρ ranges from -1 to 1 and the relationship is considered to be strong when the absolute value $|\rho|$ is above 0.7.

Note that different scoring functions output binding affinities in different units (i.e., GradDock predicts binding affinity in nM, AutoDock4 output is in kcal/mol). Not all units were consistent with our labels (i.e., nM). Additionally, the primary use of the scoring functions in virtual screening tasks is to correctly rank the scored structures. This is why we use correlation metrics rather than coefficient of determination to compare regression power of the scoring functions.

To assess the power of scoring functions to make a distinction between the binders and non-binders (classification power) the following metrics are used:

- Area Under the Receiver Operating Characteristic (AUROC) which is the area under the curve when the true positive rate is plotted against the false positive rate with varying thresholds. It gives an estimate of how well models can rank the examples. To calculate AUROC we use the binary labels for binding and non-binding pHLA complex and the predicted scores (3pHLA-score, standard-pHLA-score, Vina, Vinardo). AUROC ranges from 0 to 1, and values closer to 1 correspond to better predictive power.

Appeared in Scientific Reports, 2022
https://doi.org/10.1038/s41598-022-14526-x

- Area Under the Precision-Recall Curve (AUPRC) which is the area under the curve when the precision is plotted against recall with varying thresholds. It gives an estimate of whether models can correctly identify the positive examples without predicting too many false positives. Unlike AUROC, AUPRC is robust to imbalanced datasets. To calculate AUPRC we use the binary labels for binding and non-binding pHLA complex and the predicted scores (3pHLA-score, standard-pHLA-score, Vina, Vinardo). AUPRC ranges from 0 to 1, and values closer to 1 correspond to better predictive power.

Label transformation

$$x_{transformed} = \begin{cases} 1 - log_{50000}(x), & \text{when } x \leq 50000 \\ 0, & \text{otherwise} \end{cases} \tag{S.3}$$

where x is the binding affinity label expressed in nM units.

Per-peptide-position feature vector

To extract the per-peptide-position features we follow a protocol described in the subsection *3pHLA-score* of the section *Methods*. The final per-peptide-position feature vector has the following form:

$$\begin{bmatrix} fa\_atr_1 & fa\_rep_1 & fa\_intra\_rep_1 & ... & ref_1 \\ fa\_atr_2 & fa\_rep_2 & fa\_intra\_rep_2 & ... & ref_2 \\ \vdots & & & \vdots & \\ fa\_atr_N & fa\_rep_N & fa\_intra\_rep_N & ... & ref_N \end{bmatrix}_{N \times 19} \tag{S.4}$$

Each column corresponds to a ref2015 energy term (listed in Table S1). Each row corresponds to a different peptide residue position (as indicated by the numerical index). For example - the first row contains all the energy terms (fa_atr, fa_rep, fa_intra_rep, ..., ref) for the residue at position 1 of the peptide. N is the number of peptide residues.

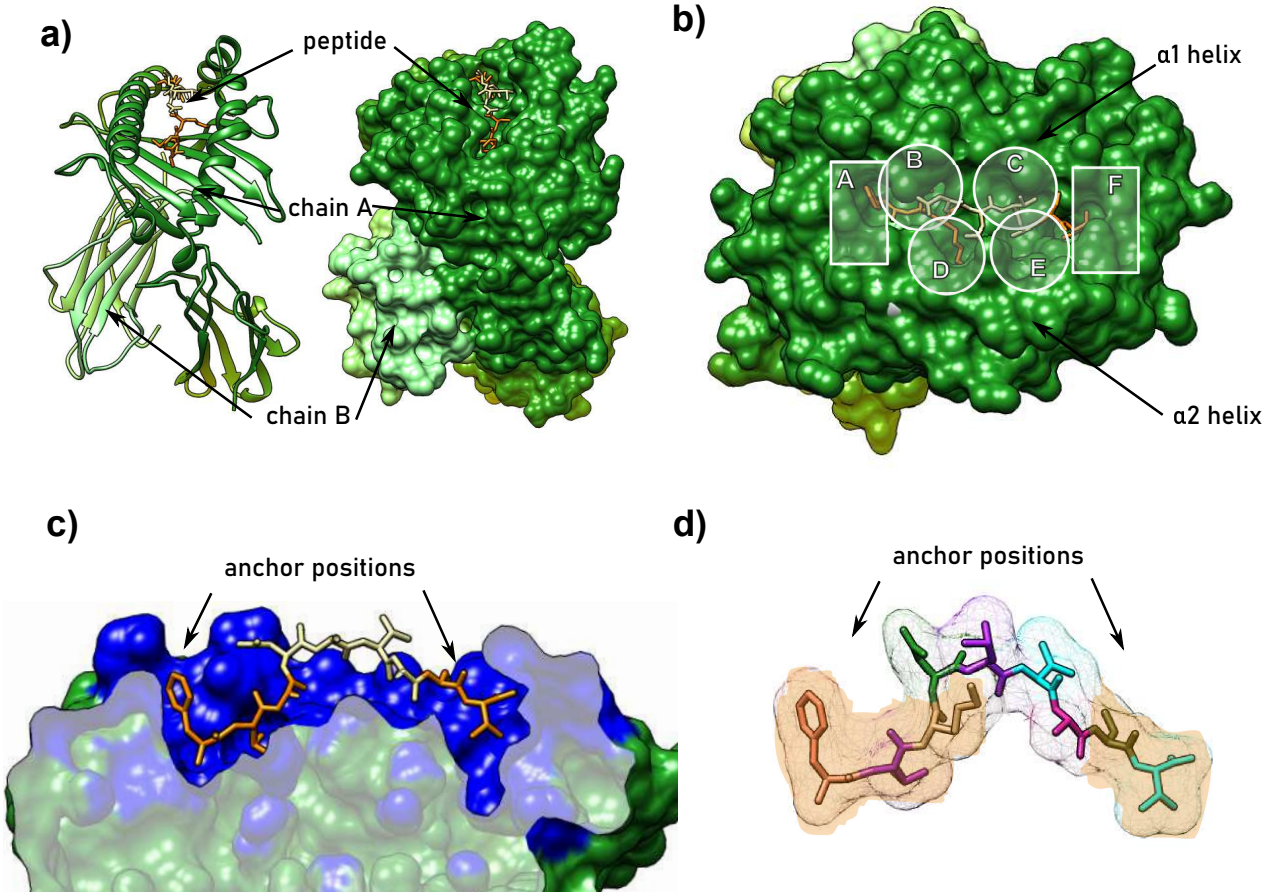

Figure S1. Basic terminology of the pHLA complex introduced using the HLA-A0201 receptor with the FLKDLVASV peptide bound. **(a)** Ribbon (left) and surface (right) representations of the complex (HLA receptor in green and peptide in orange); both chains A (dark green) and B (light green) are indicated. The anchor positions of the peptide are highlighted with a darker orange shade. **(b)** Zoomed-in view of the binding site with $\alpha 1$ and $\alpha 2$ helices indicated as well as 6 known pockets of the binding site (A-F). **(c)** A cross-section of the binding site (colored in blue) shows the depth of pockets A and F, as well as anchor positions within the peptide (orange). **(d)** Peptide representation without the HLA binding site each position within the peptide is colored with a different color. Surface around the peptide is represented with a colored mash. The anchor positions (positions 1, 2, 3, 8, 9) of the peptide are indicated with the orange shade.

Appeared in Scientific Reports, 2022
https://doi.org/10.1038/s41598-022-14526-x

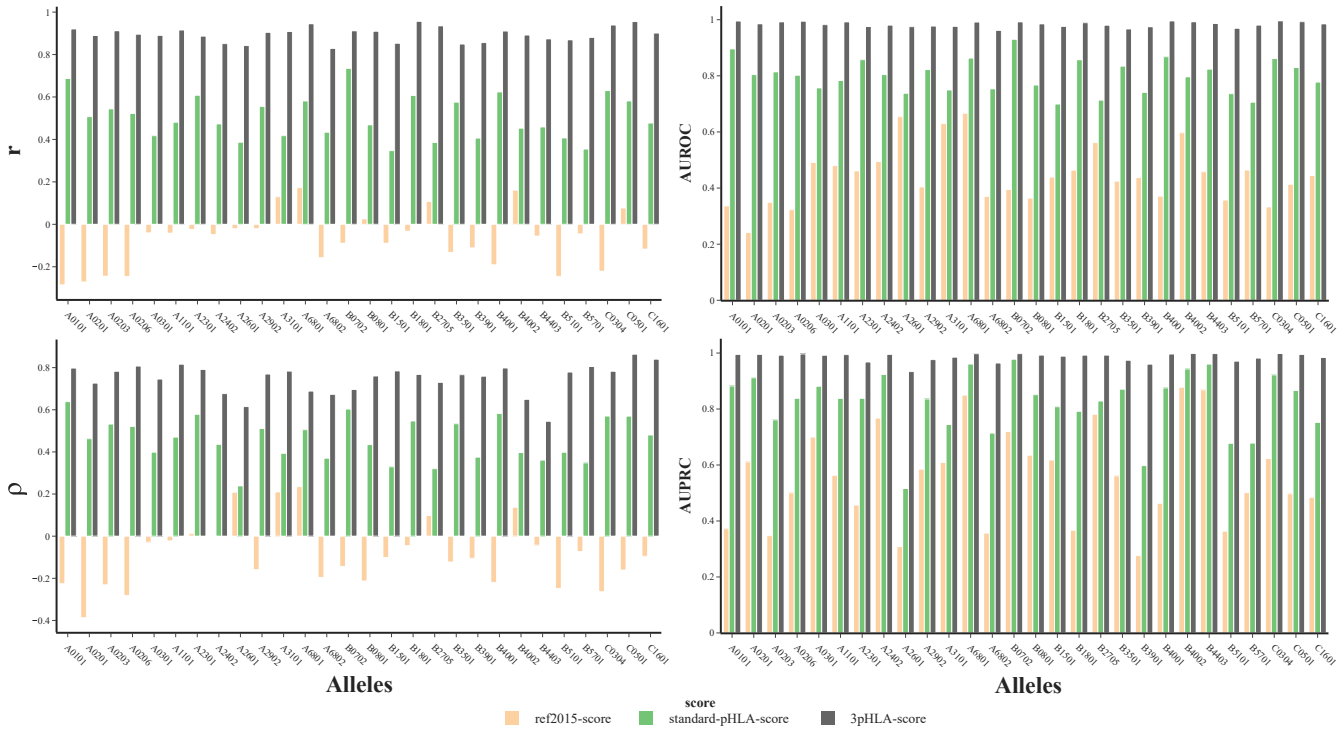

Figure S2. Three scoring functions (ref2015-score, standard-pHLA-score and 3pHLA-score) are evaluated and compared in a per-allele setting on the test portion of Dataset 1 - alleles are indicated on the x-axis. The regression power of the scoring functions is quantified with Pearson's r and Spearman's ρ metrics and plotted on the y-axis. The classification power of the scoring functions is quantified with the *AUROC* and *AUPRC* metrics and plotted on the y-axis. 3pHLA-score clearly outperforms ref2015-score and standard-pHLA-score.

Appeared in Scientific Reports, 2022
https://doi.org/10.1038/s41598-022-14526-x

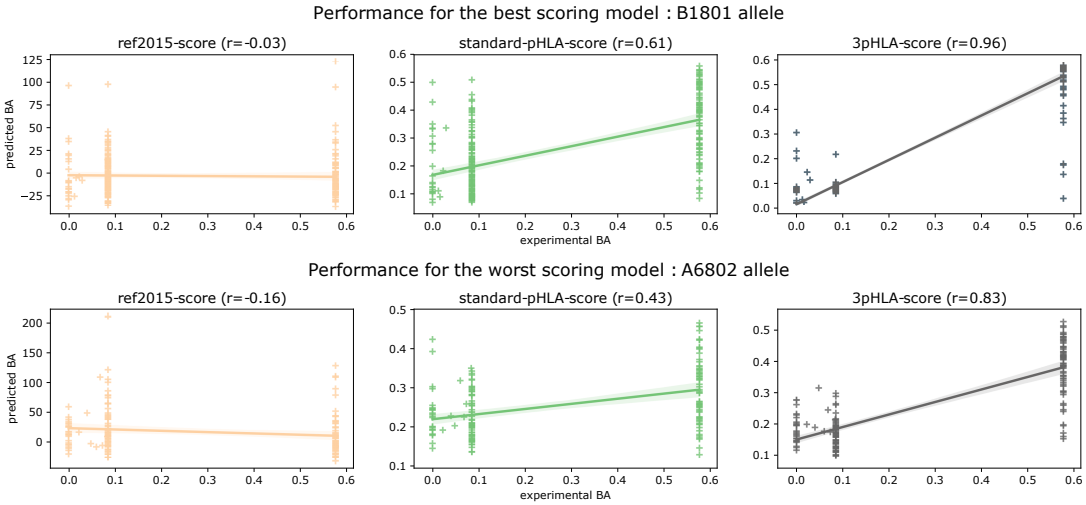

Figure S3. Scatter plot showing the correlation between experimental and predicted binding affinities for the three scoring functions (ref2015-score, standard-pHLA-score and 3pHLA-score). The predictions are made on the test portion of Dataset 1. The composition of the test set is such that the experimental binding affinities of binders are skewed around 0.6 while the values of non-binders range from 0.0 to 0.1. Here we present the results for two representative alleles - one for which the 3pHLA-score performs best (B1801 allele) and one where 3pHLA-score has the worst performance of all trained models (A6802 allele).

Appeared in Scientific Reports, 2022
https://doi.org/10.1038/s41598-022-14526-x

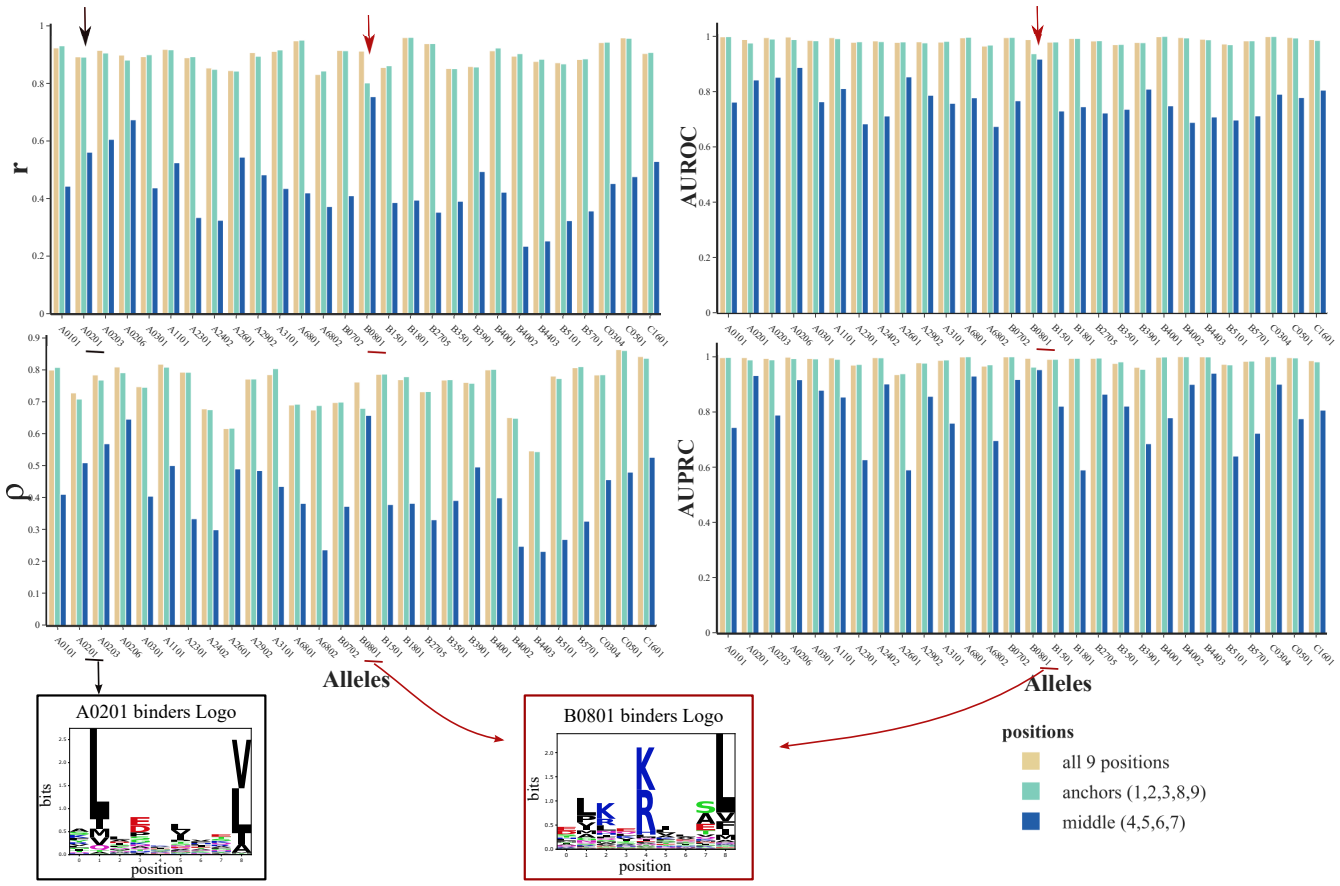

Figure S4. 3pHLA-scores trained with different peptide positions are compared (all 9 positions, anchor positions or middle positions) in a per-allele setting - alleles are indicated on the x-axis. The regression power of the scoring functions is quantified with Pearson's r and Spearman's ρ metrics and plotted on the y-axis. The classification power of the scores is quantified with *AUROC* and *AUPRC* and plotted on the y-axis. The predictive power of 3pHLA-score is similar when anchor positions are used instead of all 9 positions, while it drops when using only middle positions. The only exception is the HLA-B0801 allele: it is known to have a less common binding pattern, with a dominant anchor at position 5, which is depicted in the logo representation of its peptide binders compared to one of the more usual motifs, such as HLA-A0201.

Appeared in Scientific Reports, 2022
https://doi.org/10.1038/s41598-022-14526-x

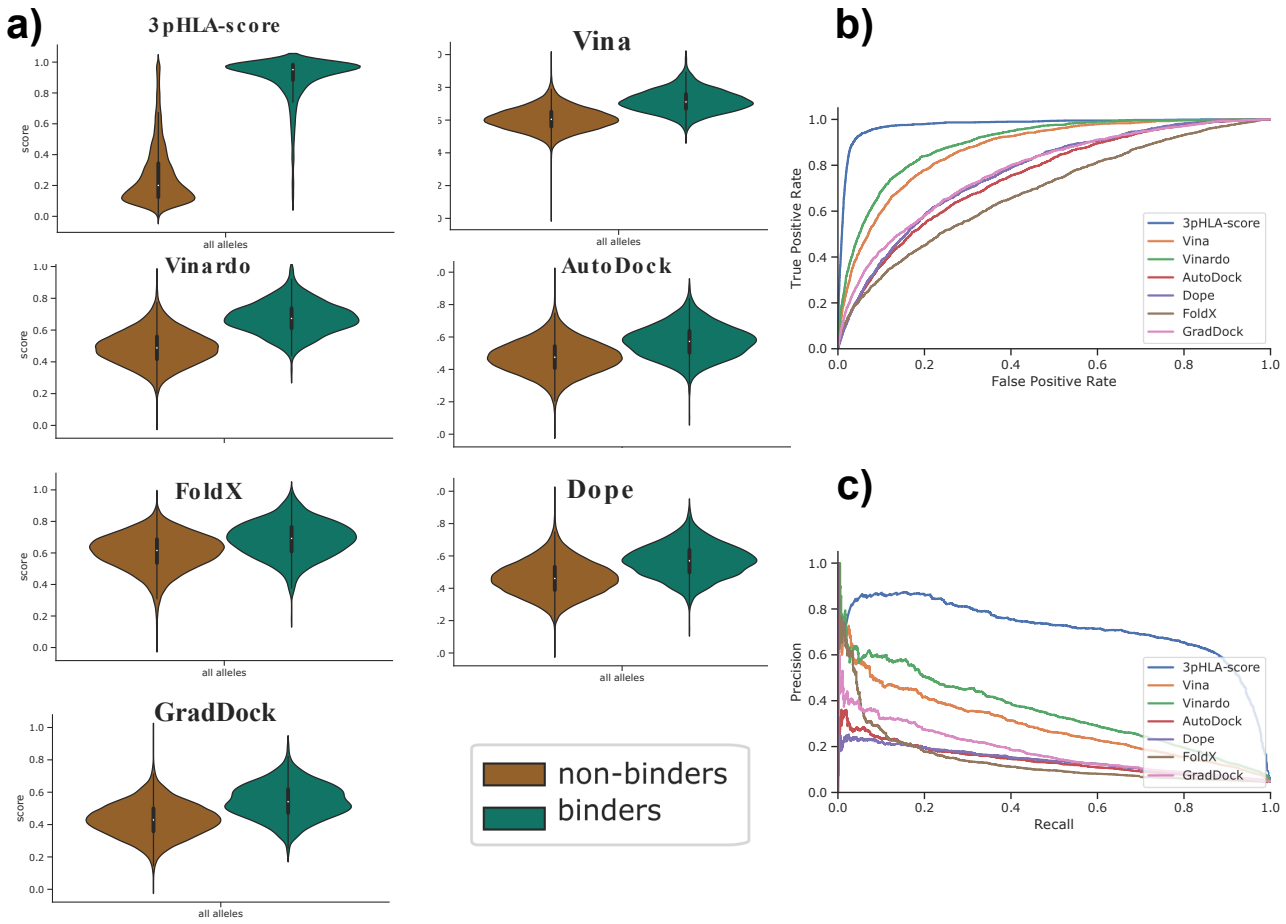

Figure S5. Performance of different scoring functions (listed in Table S5) in the virtual screening setting. Results are aggregated across alleles. a) Violin plots show the distribution of predicted binding affinities for binders (green) and non-binders(brown) and give an estimate of how well different scoring functions distinguish binders from non-binders in this setting. b) ROC-curves for different scoring functions in the virtual screening setting; c) PR-curves for different scoring functions in the virtual screening setting.

Appeared in Scientific Reports, 2022
https://doi.org/10.1038/s41598-022-14526-x

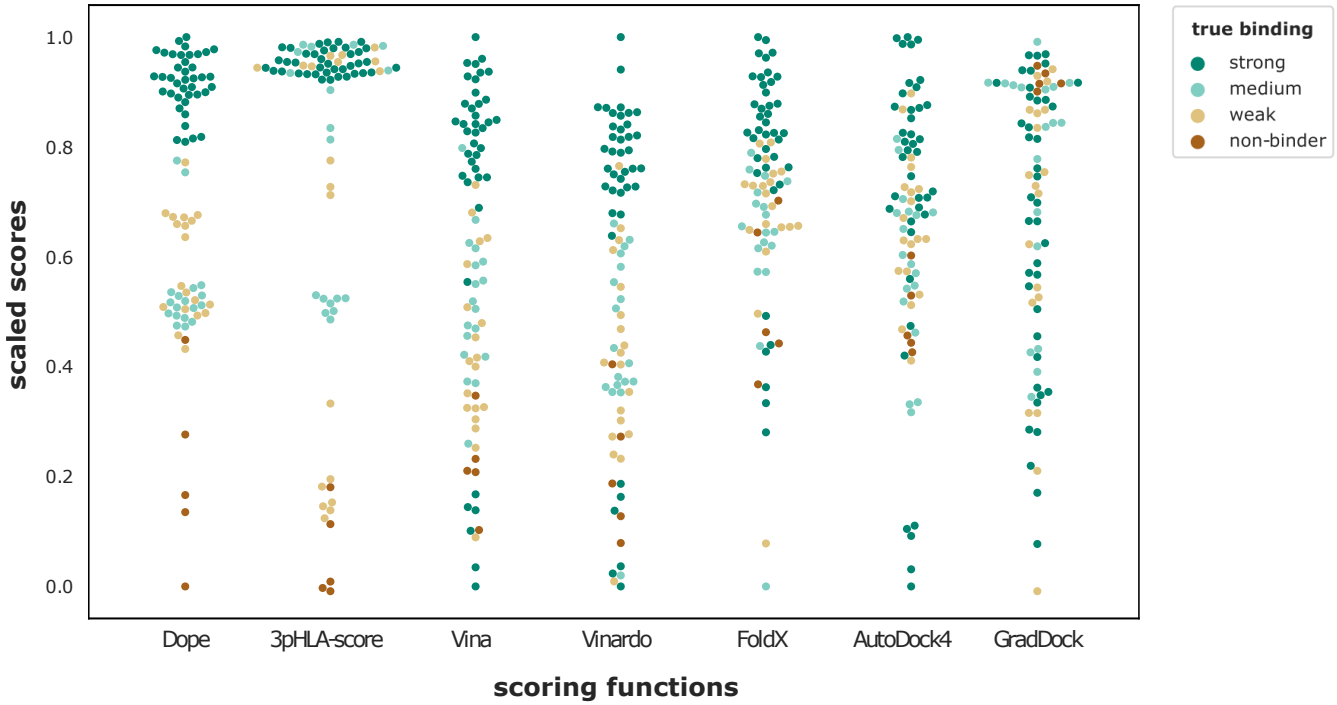

Figure S6. Scaled scores given by different scoring functions to structures from Dataset 3. The scores are scaled to fit 0-1 range and plotted on the y-axis. Each point represents a single structure and is colored based on the strength of experimental binding affinity.

Supplementary Table S1. Rosetta ref2015 energy terms[37]

Term	Description
fa_atr	Attractive energy between two atoms on different residues separated by distance, d
fa_rep	Repulsive energy between two atoms on different residues separated by distance, d
fa_intra_rep	Repulsive energy between two atoms on the same residue, separated by distance, d
fa_sol	Gaussian exclusion implicit solvation energy between protein atoms in different residues
lk_ball_wtd	Orientation-dependent solvation of polar atoms assuming ideal water geometry
fa_intra_sol	Gaussian exclusion implicit solvation energy between protein atoms in the same residue
fa_elec	Energy of interaction between two non-bonded charged atoms separated by distance, d
hbond_lr_bb	Energy of short range hydrogen bonds
hbond_sr_bb	Energy of long range hydrogen bonds
hbond_bb_sc	Energy of backbone-side chain hydrogen bonds
hbond_sc	Energy of side chain to side chain hydrogen bonds
dslf_fa13	Energy of disulfide bridges
rama_prepro	Probability of backbone ϕ,ψ angles given amino acid type
p_aa_pp	Probability of amino acid identity given backbone ϕ,ψ angles
fa_dun	Probability that a chosen rotamer is native-like given backbone ϕ,ψ angles
omega	Backbone-dependent penalty for cis ω dihedrals that deviate from 0° and trans ω dihedrals that deviate from 180°
pro_close	Penalty for an open proline ring and proline ω bonding energy
yhh_planarity	Sinusoidal penalty for non-planar tyrosine X^3 dihedral angle
ref	Reference energies for amino acid types

Supplementary Table S2. Pearson's correlation coefficient and corresponding two-sided p-values between the experimental binding affinities and predicted binding affinities. The results are evaluated on the test portion of Dataset 1, as part of the set of experiments where we compare different training protocols and are reported for each allele.

	ref2015-score		standard-pHLA-score		3pHLA-score	
Allele	r	p-value	r	p-value	r	p-value
A0101	-0.29	$4.36 * 10^{-7}$	0.69	$1.07 * 10^{-43}$	0.92	$4.18 * 10^{-125}$
A0201	-0.27	$9.15 * 10^{-15}$	0.51	$1.33 * 10^{-52}$	0.89	$3.85 * 10^{-269}$
A0203	-0.24	$3.86 * 10^{-5}$	0.54	$9.01 * 10^{-23}$	0.91	$1.37 * 10^{-108}$
A0206	-0.25	$3.48 * 10^{-3}$	0.52	$4.15 * 10^{-11}$	0.90	$3.90 * 10^{-50}$
A0301	-0.04	$3.38 * 10^{-1}$	0.42	$1.95 * 10^{-25}$	0.89	$5.25 * 10^{-196}$
A1101	-0.04	$4.87 * 10^{-1}$	0.48	$1.65 * 10^{-17}$	0.92	$2.68 * 10^{-111}$
A2301	-0.02	$7.47 * 10^{-1}$	0.61	$6.80 * 10^{-20}$	0.89	$1.36 * 10^{-62}$
A2402	-0.05	$4.25 * 10^{-1}$	0.47	$9.13 * 10^{-17}$	0.85	$1.59 * 10^{-78}$
A2601	-0.02	$7.47 * 10^{-1}$	0.39	$9.87 * 10^{-10}$	0.84	$5.92 * 10^{-64}$
A2902	-0.02	$7.76 * 10^{-1}$	0.56	$4.07 * 10^{-17}$	0.90	$4.98 * 10^{-73}$
A3101	0.13	$7.80 * 10^{-2}$	0.42	$3.59 * 10^{-9}$	0.91	$5.62 * 10^{-71}$
A6801	0.17	$7.65 * 10^{-3}$	0.58	$7.65 * 10^{-23}$	0.95	$2.92 * 10^{-116}$
A6802	-0.16	$4.38 * 10^{-2}$	0.43	$6.39 * 10^{-09}$	0.83	$9.97 * 10^{-43}$
B0702	-0.09	$4.90 * 10^{-2}$	0.73	$1.16 * 10^{-83}$	0.91	$1.73 * 10^{-189}$
B0801	0.03	$6.59 * 10^{-1}$	0.47	$3.26 * 10^{-17}$	0.91	$1.66 * 10^{-111}$
B1501	-0.09	$1.13 * 10^{-1}$	0.35	$2.19 * 10^{-10}$	0.85	$1.02 * 10^{-90}$
B1801	-0.03	$5.99 * 10^{-1}$	0.61	$8.40 * 10^{-28}$	0.96	$4.43 * 10^{-141}$
B2705	0.11	$4.82 * 10^{-2}$	0.39	$2.32 * 10^{-13}$	0.94	$5.68 * 10^{-153}$
B3501	-0.13	$3.28 * 10^{-2}$	0.58	$2.35 * 10^{-24}$	0.85	$1.19 * 10^{-73}$
B3901	-0.11	$1.65 * 10^{-1}$	0.41	$1.17 * 10^{-7}$	0.86	$1.17 * 10^{-46}$
B4001	-0.19	$7.82 * 10^{-3}$	0.62	$3.12 * 10^{-22}$	0.91	$2.19 * 10^{-75}$
B4002	0.16	$7.80 * 10^{-2}$	0.45	$2.04 * 10^{-7}$	0.89	$1.53 * 10^{-42}$
B4403	-0.06	$3.63 * 10^{-1}$	0.46	$2.00 * 10^{-15}$	0.87	$1.04 * 10^{-85}$
B5101	-0.25	$5.66 * 10^{-5}$	0.41	$5.67 * 10^{-12}$	0.87	$2.88 * 10^{-82}$
B5701	-0.05	$3.95 * 10^{-1}$	0.35	$4.61 * 10^{-12}$	0.88	$1.18 * 10^{-117}$
C0304	-0.22	$1.01 * 10^{-3}$	0.63	$1.45 * 10^{-25}$	0.94	$2.93 * 10^{-102}$
C0501	0.08	$2.52 * 10^{-1}$	0.58	$1.36 * 10^{-21}$	0.96	$4.14 * 10^{-119}$
C1601	-0.12	$9.10 * 10^{-2}$	0.48	$2.31 * 10^{-13}$	0.90	$1.39 * 10^{-77}$

Supplementary Table S3. Pearson's correlation coefficient and corresponding two-sided p-values between the experimental binding affinities and predicted binding affinities. The results are evaluated on the test portion of Dataset 1, as part of the set of experiments where we compare 3pHLA-score trained on different sets of residue positions.

	all 9 positions		anchor positions (1,2,3,8,9)		middle positions (4,5,6,7)	
Allele	r	p-value	r	p-value	r	p-value
A0101	0.92	$4.18 * 10^{-125}$	0.93	$4.01 * 10^{-131}$	0.44	$5.49 * 10^{-16}$
A0201	0.89	$3.85 * 10^{-269}$	0.89	$8.87 * 10^{-268}$	0.56	$4.37 * 10^{-66}$
A0203	0.91	$1.37 * 10^{-108}$	0.90	$5.66 * 10^{-103}$	0.60	$5.24 * 10^{-29}$
A0206	0.90	$3.90 * 10^{-50}$	0.88	$8.03 * 10^{-46}$	0.67	$1.26 * 10^{-19}$
A0301	0.89	$5.25 * 10^{-196}$	0.90	$1.91 * 10^{-203}$	0.44	$6.01 * 10^{-28}$
A1101	0.92	$2.68 * 10^{-111}$	0.91	$2.63 * 10^{-110}$	0.52	$5.39 * 10^{-21}$
A2301	0.89	$1.36 * 10^{-62}$	0.89	$9.35 * 10^{-64}$	0.33	$3.66 * 10^{-6}$
A2402	0.85	$1.59 * 10^{-78}$	0.85	$8.94 * 10^{-77}$	0.33	$3.47 * 10^{-8}$
A2601	0.84	$5.92 * 10^{-64}$	0.84	$1.92 * 10^{-63}$	0.54	$2.76 * 10^{-19}$
A2902	0.90	$4.98 * 10^{-73}$	0.89	$4.45 * 10^{-68}$	0.48	$1.09 * 10^{-12}$
A3101	0.91	$5.62 * 10^{-71}$	0.91	$3.96 * 10^{-73}$	0.43	$6.89 * 10^{-10}$
A6801	0.95	$2.92 * 10^{-116}$	0.95	$1.64 * 10^{-118}$	0.42	$1.57 * 10^{-11}$
A6802	0.83	$9.97 * 10^{-43}$	0.84	$4.64 * 10^{-45}$	0.37	$9.02 * 10^{-7}$
B0702	0.91	$1.73 * 10^{-189}$	0.91	$2.21 * 10^{-188}$	0.41	$4.08 * 10^{-21}$
B0801	0.91	$1.66 * 10^{-111}$	0.80	$1.47 * 10^{-65}$	0.75	$6.22 * 10^{-54}$
B1501	0.85	$1.02 * 10^{-90}$	0.86	$2.28 * 10^{-93}$	0.39	$1.09 * 10^{-12}$
B1801	0.96	$4.44 * 10^{-141}$	0.96	$5.67 * 10^{-142}$	0.39	$3.36 * 10^{-11}$
B2705	0.94	$5.68 * 10^{-153}$	0.94	$3.13 * 10^{-153}$	0.35	$2.73 * 10^{-11}$
B3501	0.85	$1.19 * 10^{-73}$	0.85	$1.92 * 10^{-73}$	0.39	$6.41 * 10^{-11}$
B3901	0.86	$1.17 * 10^{-46}$	0.85	$2.71 * 10^{-46}$	0.49	$4.46 * 10^{-11}$
B4001	0.91	$2.19 * 10^{-75}$	0.92	$7.44 * 10^{-80}$	0.42	$9.71 * 10^{-10}$
B4002	0.89	$1.53 * 10^{-42}$	0.90	$1.35 * 10^{-44}$	0.24	$9.67 * 10^{-3}$
B4403	0.87	$1.04 * 10^{-85}$	0.88	$5.16 * 10^{-89}$	0.25	$2.57 * 10^{-5}$
B5101	0.87	$2.88 * 10^{-82}$	0.87	$1.11 * 10^{-80}$	0.32	$7.39 * 10^{-8}$
B5701	0.88	$1.18 * 10^{-117}$	0.88	$5.12 * 10^{-119}$	0.36	$3.22 * 10^{-12}$
C0304	0.94	$2.93 * 10^{-102}$	0.94	$2.88 * 10^{-103}$	0.45	$2.23 * 10^{-12}$
C0501	0.96	$4.14 * 10^{-119}$	0.95	$1.29 * 10^{-117}$	0.48	$5.30 * 10^{-14}$
C1601	0.90	$1.39 * 10^{-77}$	0.91	$3.63 * 10^{-79}$	0.53	$1.71 * 10^{-16}$

Supplementary Table S4. Independent dataset of non-APE-Gen modeled structures - 16 peptides of the HLA-A0201 receptor[50]. Note for the five decoy peptides were modeled using Docktope tool[54], as they are not found in the PDB.

	Peptide	Method	Affinity (nM)	PDB codes
strong binders	ALWGFFPVL	purified MHC/competitive/radioactivity	2.7	1B0G, 1LP9, 2UWE, 2JCC, 2J8U
	FLPSDFFPSV	cellular MHC/competitive/radioactivity	0.57	1HHH, 3OX8, 3OXR, 3OXS
	LLFGYPVYV	purified MHC/competitive/radioactivity	3.8	1HHK, 1IM3, 2AV7, 2AV1, 1DUZ, 1AO7, 1BD2, 3IXA, 4E5X, 4FTV, 5IRO
medium binders	CINGVCWTV	purified MHC/competitive/radioactivity	55	3MRG
	ILKEPVHGV	purified MHC/competitive/radioactivity	192.3	1HHJ, 1P7Q, 2X4U, 1AKJ
	VLRDDLLEA	purified MHC/competitive/fluorescence	365	3FT4
	AAGIGILTV	purified MHC/competitive/radioactivity	395	2GUO, 2GUO, 3QEQ, 3QDJ, 3QFD
weak binders	RGPGRAFVTI	purified MHC/competitive/radioactivity	4600	3ECB, 3DMM, 1QO3, 1BII, 1DDH, 5IVX
	SLLMWITQC	purified MHC/competitive/radioactivity	21070	1S9W, 2PYE, 2P5E, 2P5W, 2F54, 2F54, 2F53, 2BNR
	RQISQDVKL	purified MHC/competitive/radioactivity	1925	4NO5, 4NO5
	EAAGIGILTV	purified MHC/competitive/fluorescence	14560	2GT9, 4QOK
non-binders	AAEQRRSTI	cellular MHC/competitive/fluorescence	>70000	Docktope model
	DAKRNSKSL	cellular MHC/competitive/fluorescence	>70000	Docktope model
	EIDVSEVKT	cellular MHC/competitive/fluorescence	>70000	Docktope model
	ATKRYPGVM	cellular MHC/competitive/fluorescence	>70000	Docktope model
	ETLNEYKQL	cellular MHC/competitive/fluorescence	>70000	Docktope model

Supplementary Table S5. Overview of the investigated scoring functions.

scoring function	**methodology**
AutoDock4[38]	empirical/forcefield
Vina[39]	empirical
Vinardo[40]	empirical
Foldx[42]	empirical/forcefield
GradDock[31]	pMHC trained ref2015
DOPE[41]	knowledge-based
3pHLA-score	pHLA trained ref2015

Supplementary Table S6. Pearson's correlation coefficient and corresponding two-sided p-values for different scoring functions evaluated on Dataset 3.

	r	p-value
DOPE[41]	0.62	$7.78 * 10^{-10}$
3pHLA-score	0.56	$5.43 * 10^{-8}$
Vina[39]	0.46	$1.49 * 10^{-5}$
Vinardo[40]	0.43	$6.12 * 10^{-5}$
FoldX[42]	0.25	0.02
AutoDock4[38]	0.18	0.11
GradDock[31]	0.17	0.13

Supplementary Table S7. Per-allele AUROC values achieved by different scoring functions in the virtual screening experiment. Best performing value in a row is bolded.

Allele	3pHLA-score	Vina	Vinardo	AutoDock4	DOPE	FoldX	GradDock
A0101	**0.997859**	0.953229	0.979900	0.845758	0.843517	0.844652	0.851253
A0201	**0.977805**	0.860607	0.914830	0.771790	0.893805	0.763165	0.877715
A0301	**0.983478**	0.872490	0.907125	0.767695	0.670895	0.719420	0.806323
A1101	**0.986580**	0.799687	0.819128	0.595910	0.569345	0.613045	0.747380
A2402	**0.991947**	0.948178	0.977718	0.850310	0.943699	0.738318	0.888586
A2902	**0.986024**	0.939271	0.958358	0.878182	0.921928	0.715233	0.856219
B0702	**0.984132**	0.910110	0.887766	0.701091	0.730810	0.757989	0.824495
B0801	**0.979995**	0.799905	0.895650	0.720445	0.817349	0.676633	0.787534
B1501	**0.968293**	0.841742	0.888293	0.673023	0.739590	0.686241	0.727943
B2705	**0.983617**	0.721595	0.822185	0.692447	0.706670	0.622695	0.721730
B3501	**0.954505**	0.931365	0.908800	0.796937	0.807605	0.710998	0.774555
B4001	**0.993772**	0.872666	0.902276	0.721731	0.654362	0.622001	0.690133
B4002	**0.976913**	0.887100	0.912087	0.762341	0.743367	0.538744	0.711106
B4403	**0.993922**	0.878464	0.933182	0.782636	0.644102	0.534307	0.645015
B5101	**0.964695**	0.942995	0.947765	0.817850	0.841995	0.941030	0.871555
B5701	**0.981696**	0.935688	0.949525	0.836808	0.831701	0.702084	0.734085

Supplementary Table S8. Per-allele AUPRC values achieved by different scoring functions in the virtual screening experiment. Best performing value in a row is bolded.

Allele	3pHLA-score	Vina	Vinardo	AutoDock4	DOPE	FoldX	GradDock
A0101	**0.956977**	0.523399	0.690218	0.192611	0.181234	0.337905	0.290420
A0201	**0.715660**	0.205126	0.298295	0.142011	0.285571	0.183424	0.357065
A0301	**0.740916**	0.277324	0.412307	0.202703	0.091861	0.146112	0.290163
A1101	**0.774023**	0.157361	0.189710	0.072240	0.055838	0.073304	0.140114
A2402	**0.919628**	0.422715	0.670886	0.206803	0.397063	0.132031	0.368762
A2902	**0.861374**	0.523034	0.598588	0.347429	0.348952	0.180041	0.306128
B0702	**0.695089**	0.376450	0.346057	0.091574	0.104016	0.197584	0.176701
B0801	**0.668274**	0.140807	0.354462	0.100157	0.158445	0.110397	0.164774
B1501	**0.800140**	0.257857	0.372958	0.099685	0.093826	0.139662	0.139493
B2705	**0.804437**	0.105399	0.186082	0.100384	0.087655	0.064670	0.168544
B3501	**0.699721**	0.443929	0.335562	0.154223	0.148952	0.146906	0.171589
B4001	**0.896198**	0.214475	0.256000	0.100702	0.068339	0.081446	0.091538
B4002	**0.761602**	0.241524	0.324725	0.135011	0.091169	0.066542	0.091808
B4403	**0.854348**	0.320935	0.469818	0.167838	0.069099	0.066236	0.077369
B5101	**0.795268**	0.531897	0.538648	0.181736	0.195025	0.636270	0.318873
B5701	**0.835307**	0.480516	0.550943	0.237850	0.181619	0.130319	0.115888

Alternative ML regression techniques

In the subsection *Machine learning models* of the *Methods* section we describe how we used Random Forest Regression models to train standard-pHLA-score and 3pHLA-score. It is possible to create variants of standard-pHLA-score and 3pHLA-score using the same protocol but replacing the Random Forest Regression with any other machine learning regression technique. Here we showcase the performance of alternative regression techniques: Linear Regression (LR), Support Vector Machine (SVM) Regression, Partial Least Squares (PLS) Regression and Random Forest (RF) Regression. We used the same dataset (training portion of Dataset 1) to train the models. We extracted the standard features and used them as input to train standard-pHLA-score. We extracted the per-peptide-position features and used them as input to train 3pHLA-score. We performed hyperparameter tuning for each regression model using 5-fold cross-validation and evaluated the performance of the models on the test portion of Dataset 1.

We report the Pearson's correlation between the experimental binding affinities and affinities predicted with standard-pHLA-score (Table S9, Figure S7) and 3pHLA-score (Table S10, Figure S7) trained using different regression techniques. For all regression techniques across all alleles we observe the same pattern as reported for the RF models in the main text: using per-peptide-position features as input to the models increases the performance of the models.

RF has overall best performance across alleles out of all regression methods we used (both for standard featurization and per-peptide-position featurization).

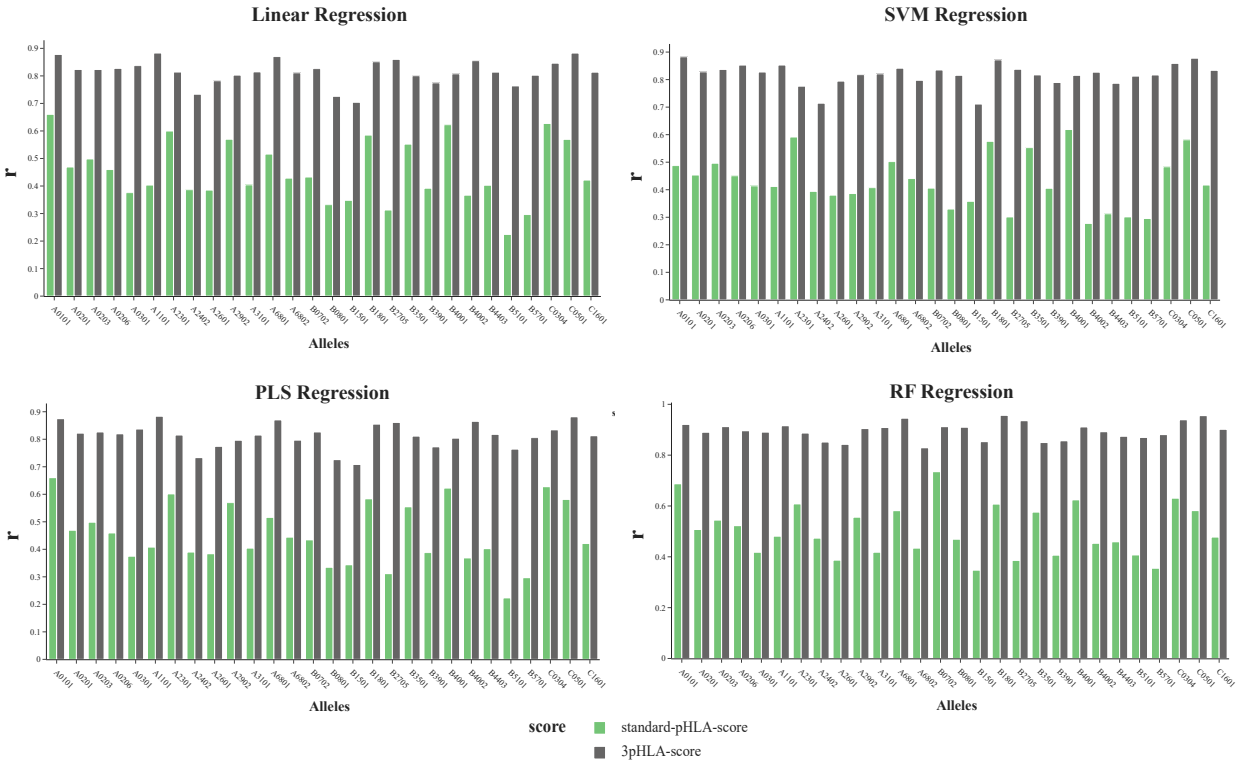

Figure S7. The standard-pHLA-score, and 3pHLA-score are trained using different ML regression techniques: Linear Regression, Support Vector Machine (SVM) Regression, Partial Least Squares (PLS) Regression and Random Forest (RF) Regression. Their performance is evaluated and compared on the test portion of Dataset 1. Results are reported for individual alleles, listed on the x-axis. The regression power of the scores is quantified using Pearson's *r*, on the y-axis.

Supplementary Table S9. Pearson's correlation coefficient and corresponding two-sided p-values between the experimental binding affinities and predicted standard-pHLA-score. The predictors were trained using the standard featurization as input to different machine learning models (linear regression - LR; support vector machine - SVM; partial least squares - PLS; random forest - RF). Models were trained on the training portion of Dataset 1. Results were obtained on the test portion of Dataset 1 and are reported for each allele.

	LR		SVM		PLS		RF	
Allele	r_p	p-value	r_p	p-value	r_p	p-value	r_p	p-value
A0101	0.66	$2.57 * 10^{-39}$	0.49	$1.63 * 10^{-19}$	0.66	$2.45 * 10^{-39}$	0.69	$1.07 * 10^{-43}$
A0201	0.47	$3.34 * 10^{-44}$	0.45	$4.61 * 10^{-41}$	0.47	$3.23 * 10^{-44}$	0.51	$1.33 * 10^{-52}$
A0203	0.50	$8.49 * 10^{-19}$	0.50	$1.27 * 10^{-18}$	0.50	$8.51 * 10^{-19}$	0.54	$9.01 * 10^{-23}$
A0206	0.46	$1.23 * 10^{-08}$	0.45	$2.55 * 10^{-08}$	0.46	$1.27 * 10^{-08}$	0.52	$4.15 * 10^{-11}$
A0301	0.38	$1.50 * 10^{-20}$	0.41	$5.47 * 10^{-25}$	0.37	$2.01 * 10^{-20}$	0.42	$1.95 * 10^{-25}$
A1101	0.40	$2.63 * 10^{-12}$	0.41	$8.25 * 10^{-13}$	0.41	$1.42 * 10^{-12}$	0.48	$1.65 * 10^{-17}$
A2301	0.60	$2.92 * 10^{-19}$	0.59	$1.18 * 10^{-18}$	0.60	$2.20 * 10^{-19}$	0.61	$6.80 * 10^{-20}$
A2402	0.39	$3.05 * 10^{-11}$	0.39	$1.2 * 10^{-11}$	0.39	$2.01 * 10^{-11}$	0.47	$9.13 * 10^{-17}$
A2601	0.39	$1.16 * 10^{-9}$	0.38	$1.97 * 10^{-9}$	0.38	$1.34 * 10^{-9}$	0.39	$9.87 * 10^{-10}$
A2902	0.57	$4.28 * 10^{-18}$	0.39	$2.66 * 10^{-8}$	0.57	$4.30 * 10^{-18}$	0.56	$4.07 * 10^{-17}$
A3101	0.40	$1.26 * 10^{-8}$	0.41	$9.29 * 10^{-9}$	0.40	$1.26 * 10^{-08}$	0.42	$3.59 * 10^{-09}$
A6801	0.52	$1.59 * 10^{-17}$	0.50	$1.5 * 10^{-16}$	0.52	$1.59 * 10^{-17}$	0.58	$7.65 * 10^{-23}$
A6802	0.43	$1.03 * 10^{-8}$	0.44	$3.53 * 10^{-9}$	0.44	$2.56 * 10^{-9}$	0.43	$6.39 * 10^{-9}$
B0702	0.43	$1.35 * 10^{-23}$	0.41	$1.09 * 10^{-20}$	0.43	$8.74 * 10^{-24}$	0.73	$1.16 * 10^{-83}$
B0801	0.33	$7.01 * 10^{-9}$	0.33	$9.17 * 10^{-9}$	0.33	$5.49 * 10^{-9}$	0.47	$3.26 * 10^{-17}$
B1501	0.35	$2.02 * 10^{-10}$	0.36	$5.57 * 10^{-11}$	0.34	$3.38 * 10^{-10}$	0.35	$2.19 * 10^{-10}$
B1801	0.58	$1.91 * 10^{-25}$	0.58	$1.59 * 10^{-24}$	0.58	$2.62 * 10^{-25}$	0.61	$8.41 * 10^{-28}$
B2705	0.31	$4.57 * 10^{-9}$	0.30	$1.77 * 10^{-08}$	0.31	$5.38 * 10^{-9}$	0.39	$2.32 * 10^{-13}$
B3501	0.55	$3.79 * 10^{-22}$	0.55	$3.18 * 10^{-22}$	0.55	$2.28 * 10^{-22}$	0.58	$2.35 * 10^{-24}$
B3901	0.39	$3.55 * 10^{-7}$	0.41	$1.28 * 10^{-7}$	0.39	$4.66 * 10^{-7}$	0.41	$1.17 * 10^{-7}$
B4001	0.62	$4.02 * 10^{-22}$	0.62	$8.91 * 10^{-22}$	0.62	$4.31 * 10^{-22}$	0.62	$3.12 * 10^{-22}$
B4002	0.37	$3.82 * 10^{-5}$	0.28	$2.14 * 10^{-3}$	0.37	$3.42 * 10^{-5}$	0.45	$2.04 * 10^{-7}$
B4403	0.40	$6.55 * 10^{-12}$	0.31	$1.63 * 10^{-7}$	0.40	$6.85 * 10^{-12}$	0.46	$2.00 * 10^{-15}$
B5101	0.22	$2.42 * 10^{-4}$	0.30	$6.02 * 10^{-7}$	0.22	$2.53 * 10^{-4}$	0.41	$5.67 * 10^{-12}$
B5701	0.30	$1.08 * 10^{-8}$	0.30	$1.21 * 10^{-8}$	0.30	$1.03 * 10^{-8}$	0.35	$4.61 * 10^{-12}$
C0304	0.63	$3.09 * 10^{-25}$	0.48	$3.62 * 10^{-14}$	0.63	$2.89 * 10^{-25}$	0.63	$1.45 * 10^{-25}$
C0501	0.57	$1.46 * 10^{-20}$	0.58	$1.44 * 10^{-21}$	0.58	$1.50 * 10^{-21}$	0.58	$1.36 * 10^{-21}$
C1601	0.42	$1.85 * 10^{-10}$	0.42	$3.03 * 10^{-10}$	0.42	$1.97 * 10^{-10}$	0.48	$2.31 * 10^{-13}$

Supplementary Table S10. Pearson's correlation coefficient and corresponding two-sided p-values between the experimental binding affinities and predicted 3pHLA-score. The predictors were trained using the per-peptide-position featurization as input to different machine learning models (linear regression - LR; support vector machine - SVM; partial least squares - PLS; random forest - RF). Models were trained on the training portion of Dataset 1. Results were obtained on the test portion of Dataset 1 and are reported for each allele.

	LR		SVM		PLS		RF	
Allele	r_p	p-value	r_p	p-value	r_p	p-value	r_p	p-value
A0101	0.88	$2.91 * 10^{-98}$	0.87	$8.99 * 10^{-96}$	0.87	$1.03 * 10^{-96}$	0.92	$4.18 * 10^{-125}$
A0201	0.82	$8.53 * 10^{-195}$	0.82	$9.38 * 10^{-195}$	0.82	$9.39 * 10^{-194}$	0.89	$3.85 * 10^{-269}$
A0203	0.82	$1.44 * 10^{-69}$	0.83	$3.57 * 10^{-70}$	0.83	$1.37 * 10^{-70}$	0.91	$1.37 * 10^{-108}$
A0206	0.83	$4.05 * 10^{-36}$	0.83	$7.10 * 10^{-36}$	0.82	$5.61 * 10^{-35}$	0.90	$3.90 * 10^{-50}$
A0301	0.84	$6.49 * 10^{-151}$	0.82	$1.44 * 10^{-139}$	0.84	$1.09 * 10^{-150}$	0.89	$5.25 * 10^{-196}$
A1101	0.88	$1.14 * 10^{-92}$	0.84	$1.17 * 10^{-75}$	0.88	$6.67 * 10^{-93}$	0.92	$2.68 * 10^{-111}$
A2301	0.81	$1.44 * 10^{-44}$	0.76	$4.05 * 10^{-36}$	0.82	$8.10 * 10^{-45}$	0.89	$1.36 * 10^{-62}$
A2402	0.73	$1.21 * 10^{-47}$	0.70	$4.35 * 10^{-41}$	0.73	$1.23 * 10^{-47}$	0.85	$1.59 * 10^{-78}$
A2601	0.78	$1.99 * 10^{-49}$	0.77	$1.61 * 10^{-47}$	0.77	$9.13 * 10^{-48}$	0.84	$5.92 * 10^{-64}$
A2902	0.80	$5.50 * 10^{-45}$	0.80	$6.60 * 10^{-44}$	0.80	$8.86 * 10^{-44}$	0.90	$4.98 * 10^{-73}$
A3101	0.81	$6.32 * 10^{-45}$	0.80	$1.30 * 10^{-42}$	0.82	$4.88 * 10^{-45}$	0.91	$5.62 * 10^{-71}$
A6801	0.87	$3.11 * 10^{-74}$	0.82	$6.25 * 10^{-60}$	0.87	$2.95 * 10^{-74}$	0.95	$2.92 * 10^{-116}$
A6802	0.81	$1.32 * 10^{-39}$	0.77	$5.80 * 10^{-34}$	0.80	$2.95 * 10^{-37}$	0.83	$9.97 * 10^{-43}$
B0702	0.83	$6.63 * 10^{-123}$	0.82	$7.27 * 10^{-122}$	0.83	$7.30 * 10^{-123}$	0.91	$1.73 * 10^{-189}$
B0801	0.73	$1.59 * 10^{-48}$	0.80	$1.64 * 10^{-66}$	0.73	$1.42 * 10^{-48}$	0.91	$1.66 * 10^{-111}$
B1501	0.70	$1.38 * 10^{-48}$	0.69	$1.40 * 10^{-46}$	0.71	$2.03 * 10^{-49}$	0.85	$1.02 * 10^{-90}$
B1801	0.85	$7.44 * 10^{-75}$	0.86	$1.82 * 10^{-77}$	0.85	$4.71 * 10^{-76}$	0.96	$4.44 * 10^{-141}$
B2705	0.86	$1.43 * 10^{-99}$	0.82	$1.28 * 10^{-83}$	0.86	$4.02 * 10^{-100}$	0.94	$5.68 * 10^{-153}$
B3501	0.80	$2.38 * 10^{-59}$	0.80	$1.17 * 10^{-59}$	0.81	$4.36 * 10^{-62}$	0.85	$1.19 * 10^{-73}$
B3901	0.77	$6.88 * 10^{-33}$	0.75	$2.42 * 10^{-30}$	0.77	$1.43 * 10^{-32}$	0.86	$1.17 * 10^{-46}$
B4001	0.81	$1.10 * 10^{-45}$	0.80	$1.06 * 10^{-43}$	0.80	$5.72 * 10^{-45}$	0.91	$2.19 * 10^{-75}$
B4002	0.86	$1.69 * 10^{-35}$	0.78	$1.72 * 10^{-25}$	0.87	$3.82 * 10^{-37}$	0.89	$1.53 * 10^{-42}$
B4403	0.81	$8.04 * 10^{-65}$	0.76	$7.25 * 10^{-53}$	0.82	$5.64 * 10^{-66}$	0.87	$1.04 * 10^{-85}$
B5101	0.76	$1.08 * 10^{-51}$	0.80	$1.44 * 10^{-60}$	0.76	$8.67 * 10^{-52}$	0.87	$2.88 * 10^{-82}$
B5701	0.80	$7.98 * 10^{-82}$	0.81	$5.92 * 10^{-83}$	0.81	$4.19 * 10^{-83}$	0.88	$1.18 * 10^{-117}$
C0304	0.85	$6.17 * 10^{-61}$	0.84	$3.40 * 10^{-60}$	0.83	$9.60 * 10^{-58}$	0.94	$2.93 * 10^{-102}$
C0501	0.88	$1.73 * 10^{-74}$	0.87	$1.03 * 10^{-68}$	0.88	$4.97 * 10^{-74}$	0.96	$4.14 * 10^{-119}$
C1601	0.81	$7.69 * 10^{-51}$	0.82	$1.20 * 10^{-52}$	0.81	$1.06 * 10^{-50}$	0.90	$1.39 * 10^{-77}$

3pHLA-score comparison with sequence-based approaches in the epitope discovery setting

We developed the 3pHLA-score with the purpose of structure-based virtual screening. In that context, we have compared its performance mainly to structure-based scoring functions. However, as we mention in the introduction, current methods for scoring peptide HLAs are mostly sequence-based. Sequence-based scores do not use structure as input and thus can not be used for structure-based virtual screens. Nevertheless, it is interesting to see how 3pHLA-score compares to the most widely used sequence-based scores. Here we evaluate the performance of 3pHLA-score, MHCFlurry2.0[10] and NetMHCpan4.1[58] in an epitope discovery setting (Dataset 2).

Average AUROC and AUPRC values across all alleles are reported in Table S11 for the compared scoring functions. Figure S8 shows the AUROC and AUPRC curves along with violin plots that depict the distribution of predicted scores across the binder and non-binder peptides from Dataset 2. Finally, Tables S12, S13 show the AUROC and AUPRC obtained for each allele. As expected, sequence-based approaches have very good performance across all alleles with MHCFlurry2.0 having the highest average AUROC and AUPRC (0.993 and 0.865 respectively). 3pHLA-score lags behind the sequence-based approaches with AUROC and AUPRC of 0.977 and 0.712. 3pHLA-score still has comparable performance for most of the alleles in terms of AUROC values (Table S13).

It is important to note that the dataset on which this experiment is performed (Dataset 2) is left out of the training of 3pHLA-score. However, we do not know if MHCFlurry2.0 or NetMHCpan4.1 have had a part of this dataset in their training, which might give them a slight advantage. Structure-based scoring functions are inherently more difficult to train. To the best of our knowledge structure-based scoring functions have not yet come close to the performance of sequence-based methods. The comparable performance of 3pHLA-score shows promise that structure-based approached can reach the accuracies of sequence-based approaches and can bridge the gaps that we mention in the introduction.

Supplementary Table S11. AUROC and AUPRC values aggregated for the virtual screening experiment across HLA alleles. The highest values are bolded.

	AUROC	AUPRC
3pHLA-score	0.977	0.712
MHCFlurry2.0[10]	**0.993**	**0.865**
NetMHCpan4.1[58]	0.991	0.855

Supplementary Table S12. Per-allele AUPRC values achieved by two sequence-based scoring functions and 3pHLA-score in the virtual screening experiment. The best performance in each row is bolded.

Allele	3pHLA-score	MHCFlurry2.0	NetMHCpan4.1
A0101	0.956977	0.948338	**0.967984**
A0201	0.715660	0.811894	**0.823205**
A0301	0.740916	**0.859079**	0.832935
A1101	0.774023	**0.917182**	0.905915
A2402	0.919628	**0.961677**	0.922089
A2902	0.861373	**0.919257**	0.883255
B0702	0.695089	**0.940622**	0.931435
B0801	0.668274	**0.949789**	0.901494
B1501	0.800140	**0.883350**	0.807565
B2705	0.804437	**0.950496**	0.793621
B3501	0.699721	**0.890328**	0.862872
B4001	0.896198	**0.907023**	0.883645
B4002	0.761602	**0.834503**	0.786427
B4403	0.854348	**0.904832**	0.877956
B5101	0.795268	**0.935606**	0.875531
B5701	0.835306	**0.913895**	0.888805

Supplementary Table S13. Per-allele AUROC values achieved by two sequence-based scoring functions and 3pHLA-score in the virtual screening experiment. The best performance in each row is bolded.

Allele	3pHLA-score	MHCFlurry2.0	NetMHCpan4.1
A0101	0.997860	0.998440	**0.998790**
A0201	0.977805	0.985580	**0.986395**
A0301	0.983478	**0.994025**	0.988960
A1101	0.986580	**0.995910**	0.995730
A2402	0.991955	**0.998285**	0.997485
A2902	0.986033	**0.995858**	0.994907
B0702	0.984140	**0.997170**	0.997115
B0801	0.980005	**0.996925**	0.995970
B1501	0.968315	**0.994275**	0.990650
B2705	0.983617	**0.997285**	0.991645
B3501	0.954505	**0.987875**	0.983180
B4001	0.993775	**0.997815**	0.996590
B4002	0.976924	**0.988520**	0.984520
B4403	0.993925	**0.995845**	0.994845
B5101	0.964695	**0.991275**	0.986595
B5701	0.981700	**0.995110**	0.993540

Appeared in Scientific Reports, 2022
https://doi.org/10.1038/s41598-022-14526-x

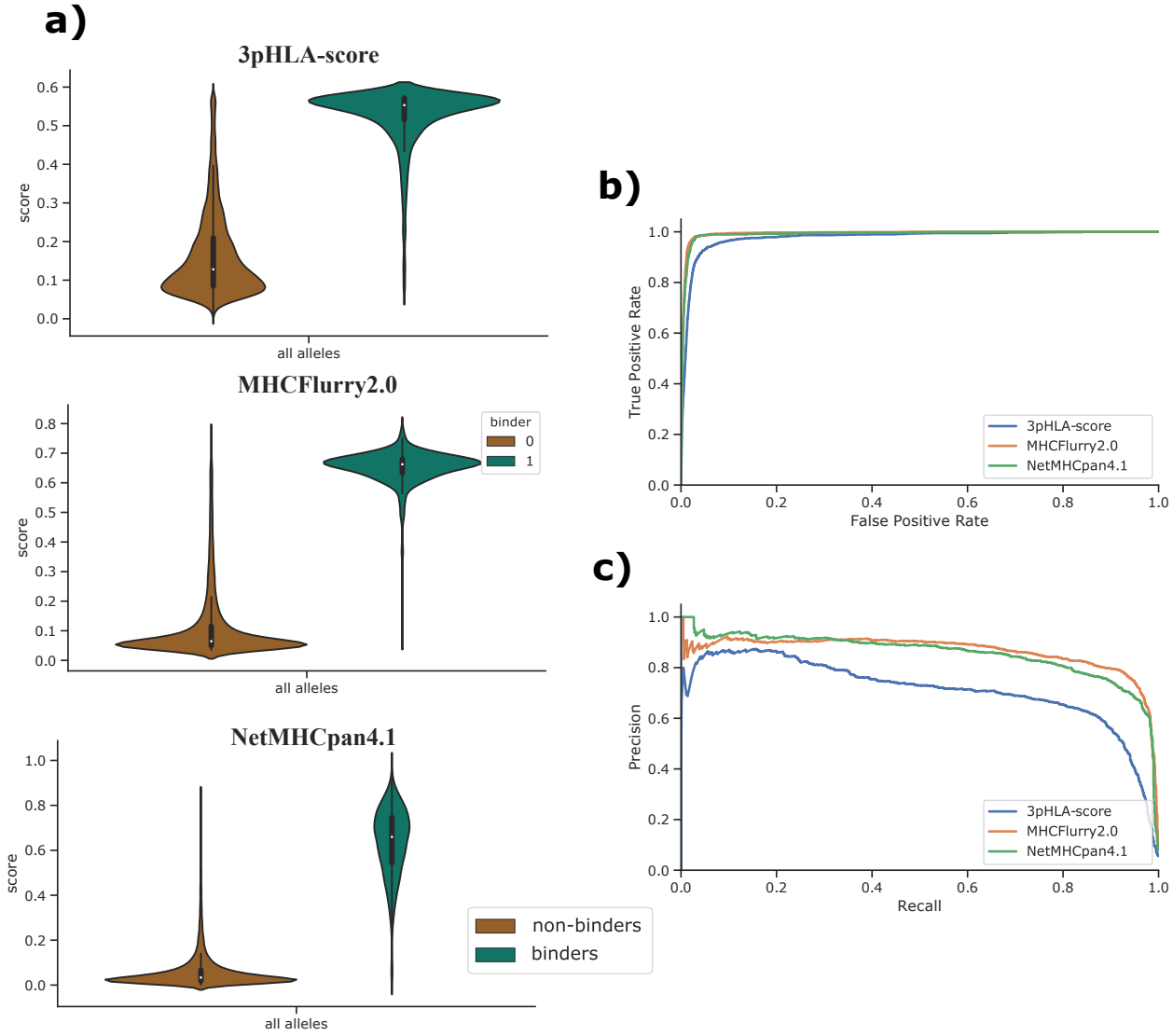

Figure S8. Comparing 3pHLA-score to sequence-based approached (MHCFlurry2.0, NetMHCpan4.1) in the virtual screening setting. Results are aggregated across alleles. a) Violin plots show the distribution of predicted binding affinities for binders (green) and non-binders (brown) and give an estimate of how well different scoring functions distinguish binders from non-binders in this setting. b) ROC-curves for different scoring functions in the virtual screening setting; c) PR-curves in the virtual screening setting.

Appeared in Scientific Reports, 2022
https://doi.org/10.1038/s41598-022-14526-x