Iran. J. Biotechnol. October 2024;22(4): e3927 DOI: 10.30498/ijb.2024.459448.3927

Research Article

HLAPepBinder: An Ensemble Model for The Prediction Of HLA-Peptide Binding

Mahsa Saadat[1], Fatemeh Zare-Mirakabad,[1,*] Ali Masoudi-Nejad[2], Mohammad Farahanchi Baradaran[1], Nazanin Hosseinkhan[3]

[1]Computational Biology Research Center (CBRC), Department of Mathematics and Computer Science, Amirkabir University of Technology, Tehran, Iran
[2]Laboratory of Systems Biology and Bioinformatics (LBB), Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran
[3]Institute of Endocrinology and Metabolism, Iran University of Medical Sciences, Tehran, Iran

[*]*Corresponding author*: Fatemeh Zare-Mirakabad, Department of Mathematics and Computer Science, Amirkabir University of Technology, Tehran, Iran. Tel/ Fax: +98-2164545674, E-mail: f.zare@aut.ac.ir

Received: 2024/05/25 ; **Accepted:** 2024/12/03

Background: Human leukocyte antigens (HLAs) play a pivotal role in orchestrating the host's immune response, offering a promising avenue with reduced adverse effects compared to conventional treatments. Cancer immunotherapies use HLA class I molecules for T cells to recognize tumor antigens, emphasizing the importance of identifying peptides that bind effectively to HLAs. Computer modeling of HLA-peptide binding speeds up the search for immunogenic epitopes, which enhances the prospect of personalized medicine and targeted therapies. The Immune Epitope Database (IEDB) is a vital repository, housing curated immune epitope data and prediction tools for HLA-peptide binding. It can be challenging for immunologists to choose the best tool from the IEDB for predicting HLA-peptide binding. This has led to the creation of consensus-based methods that combine the results of several predictors. One of the major challenges in these methods is how to effectively integrate the results from multiple predictors.

Objectives: Previous consensus-based methods integrate at most three tools by relying on simple strategies, such as selecting prediction methods based on their proximity to HLA in training data. In this study, we introduce *HLAPepBinder*, a novel consensus approach using ensemble machine learning methods to predict HLA-peptide binding, addressing the challenges biologists face in model selection.

Materials and Methods: The key contribution is the development of an automatic pipeline named *HLAPepBinder* that integrates the predictions of multiple models using a random forest approach. Unlike previous approaches, *HLAPepBinder* seamlessly integrates results from all nine predictors, providing a comprehensive and accurate predictive framework. By combining the strengths of these models, *HLAPepBinder* eliminates the need for manual model selection, providing a streamlined and reliable solution for biologists.

Results: *HLAPepBinder* offers a practical and high-performing alternative for HLA-peptide binding predictions, outperforming both traditional methods and complex deep learning models. Compared to the recently introduced transformer-based model, TranspHLA, which requires substantial computational resources, *HLAPepBinder* demonstrates superior performance in both prediction accuracy and resource efficiency. Notably, it operates effectively in limited computational environments, making it accessible to researchers with minimal resources. The codes are available online at https://github.com/CBRC-lab/HLAPepBinder.

Conclusion: Our study introduces a novel ensemble-learning model designed to enhance the accuracy and efficiency of HLA-peptide binding predictions. Due to the lack of reliable negative data and the typical assumption of unknown interactions being negative, we focus on analyzing the unknown HLA-peptide bindings in the test set that our model predicts with 100% certainty as positive bindings. Using *HLAPepBinder*, we identify 26 HLA-peptide pairs with absolute prediction confidence. These predictions are validated through a multi-step pipeline involving literature review, BLAST sequence similarity analysis, and molecular docking studies. This comprehensive validation process highlights *HLAPepBinder*'s ability to make accurate and reliable predictions, contributing significantly to advancements in immunotherapy and vaccine development.

Keyword: HLA class I, HLA-peptide binding, Immunotherapy, Random Forest, T cell epitope

1. Background

The regulation of the host's immune response is significantly influenced by human leukocyte antigens (HLAs) (1). Positioned on the cell surface, these proteins play a crucial role in enabling the immune system to distinguish between the body's own cells and external entities. The HLA-mediated mechanism has become increasingly significant in the context of immunotherapy for cancer treatment, representing a promising approach with reduced adverse effects compared to conventional chemotherapy and radiotherapy (2–5). HLAs are categorized into two types: class I and class II.

HLA class I molecules play a key role in presenting tumor antigens on the surfaces of tumor cells to cytotoxic T cells (CD8+ T cells). This presentation allows CD8+ T cells to recognize and eliminate cancerous cells through immune-mediated mechanisms, which is a fundamental mechanism underlying cancer immunotherapies such as adoptive cell therapy and checkpoint blockade therapy. While HLA class I molecules are primarily responsible for presenting antigens to cytotoxic T cells, HLA class II molecules also play a role in presenting antigens to helper T cells (CD4+ T cells), contributing to the overall immune response against tumors (2,4,6). It is evident that HLA class I molecules play a crucial role in directly mediating the immune response against cancer cells by presenting tumor antigens to CD8+ T cells. In this study, we focus on HLA class I and abbreviate it as "HLA" for brevity and ease of reference.

In the field of immunotherapy, the identification of peptides capable of effectively binding to HLA molecules is important. The careful selection of peptides with a strong affinity for particular HLA types is crucial for optimizing the efficiency of immunotherapeutic methods. This guarantees a prompt and accurate activation of the immune system to detect and combat specific target cells, whether they are cancerous or infected cells.

The precision of the binding between HLA and peptides is of significant importance in promoting a directed immune response, underscoring the critical role of meticulously choosing peptides designed to an individual's HLA profile in immunotherapeutic strategies. The computational identification of peptides with the potential to bind to HLA, known as HLA-peptide binding prediction or T cell epitope prediction, accelerates the discovery of immunogenic epitopes. This not only enhances the prospects of personalized medicine strategies but also expedites the development of targeted immunotherapies for conditions such as cancer and infectious disorders (7,8).

The Immune Epitope Database (IEDB, https://www.iedb.org/) is a comprehensive repository of curated immune epitope data sourced from scientific literature and submissions. This valuable resource not only keeps information but also offers prediction tools for HLA-peptide binding (9). These tools can be classified into three primary categories based on the methodologies they utilize, including scoring function-based methods, machine learning-based methods and, consensus-based methods.

Within the first category (10–12), candidate peptide sequences are assessed using specific features such as sequence similarity and amino acid abundance. Variations among the tools lie in the statistical approaches employed to calculate binding scores. In the second category, machine learning-based methods (13–17), peptides are classified as binding or non-binding through model training based on extracted features. Construction of a machine learning-based model involves feature encoding based on peptide and HLA sequences, selecting an optimal machine learning algorithm, training the model, optimizing it, and evaluating performance. Notably, artificial neural networks are the most widely used algorithm in current HLA-peptide binding prediction tools. Immunologists consistently face the challenge of selecting the most appropriate tool within the IEDB for HLA-peptide binding prediction, given the diverse range of tools available (18). So, consensus-based methods (18,19) are defined as the third category. These methods merge some predictors of peptide binding to HLA in a weighted manner, resulting in a final prediction score based on the collective outcomes of individual predictors. The aim of these methods is to integrate results from multiple predictors, thereby improving overall prediction performance compared to individual predictions. Earlier consensus-based methods incorporate up to three tools, typically using straightforward strategies like selecting prediction methods based on their closeness to HLA in the training data. For instance, NetMHCcons (18) is a combination of NetMHC (13), NetMHCpan (14) and PickPocket (10), determining the appropriate prediction method by assessing the distance between the query HLA allotype and its nearest neighbor in the base method's train set. Additionally, existing consensus-based methods generally incorporate the results of only two or three predictor tools, despite the growing availability of additional tools that could be integrated. A key challenge, however is the seamless integration of these numerous predictive models. Thus, the development of a consensus framework to effectively combine the outputs of HLA-peptide predictor tools is both necessary and valuable.

2. Objective

In this paper, we present *HLAPepBinder*, a novel consensus approach that utilizes ensemble machine learning, to simplify model selection process for biologists. *HLAPepBinder* tackles the common challenge of predicting HLA-peptide binding by combining the strengths of multiple existing models while minimizing their individual limitations. This integration removes the complexity of choosing the right model for biologists, providing a more reliable and streamlined solution for making HLA-peptide binding predictions.

Instead of creating a complex new model from scratch—which is often time-consuming and requires significant resources—our study uses the predictive capabilities of well-established models. By combining their strengths, we offer a more efficient and practical alternative. Unlike previous approaches that manually combined just two or three models, *HLAPepBinder* automates the integration process and incorporates a wider range of predictors.

A major innovation of *HLAPepBinder* is its automatic pipeline, which uses a random forest (RF) model to combine predictions from multiple existing tools. By synthesizing outputs from different predictors, the RF model in *HLAPepBinder* provides a highly efficient and accurate method for generating final HLA-peptide binding predictions. Specifically, *HLAPepBinder* integrates nine baseline predictors from IEDB: ANN (17), Consensus (19), NetMHCpan BA (14), NetMHCpan EL (14), SMM (11), SMMPMBEC (12), PickPocket (10), NetMHCcons (18) and NetMHCstabpan (16). The strength of *HLAPepBinder* lies in its ability to combine diverse prediction models, each using distinct methodologies. For example, some models rely on HLA allele names (17), while others utilize full HLA sequences (14). Additionally, some models incorporate structural information in their predictions (10). By bringing together these varied approaches, *HLAPepBinder* delivers more accurate and broadly applicable predictions across a wide range of HLA subtypes and datasets. This ensemble not only outperforms each of the nine individual baseline tools from IEDB but also clearly demonstrates the practical advantages of integrating models to advance HLA-peptide binding prediction.

Recently, a more complex model called TranspHLA (20), based on transformer architecture, was introduced. While TranspHLA outperforms all nine baseline models, it demands substantial computational resources. In contrast, our results demonstrate that *HLAPepBinder* not only surpasses TranspHLA in prediction accuracy but also excels in processing speed and resource efficiency.

We develop and evaluate *HLAPepBinder* within the Google Colab environment, utilizing limited computational resources (12 GB of RAM and a standard CPU). Despite these constraints, the model demonstrates robust performance. In comparison, TranspHLA requires a high-performance system equipped 92 GB of RAM, 80 CPU cores, and an Nvidia RTX 3080 GPU to function optimally. Notably, *HLAPepBinder* operates with fewer parameters and reduces computational complexity, rendering it a more accessible option for researchers with restricted computational resources. The model's ability to maintain high efficiency under these conditions underscores its practical advantages over more computationally demanding models such as TranspHLA.

3. Material and Methods

This section outlines the HLA-peptide binding prediction as a computational problem. We then describe our proposed pipeline, *HLAPepBinder*, uses an ensemble approach to address the HLA-peptide binding problem.

3.1. Problem Definition

The HLA-peptide binding problem is defined as follows:

- **Inputs:**

1. Peptide sequence (X) : A short sequence denoted as $X=x_1...x_m$, where m represents the length of the sequence, and x_i corresponds to the amino acid at the i^{th}

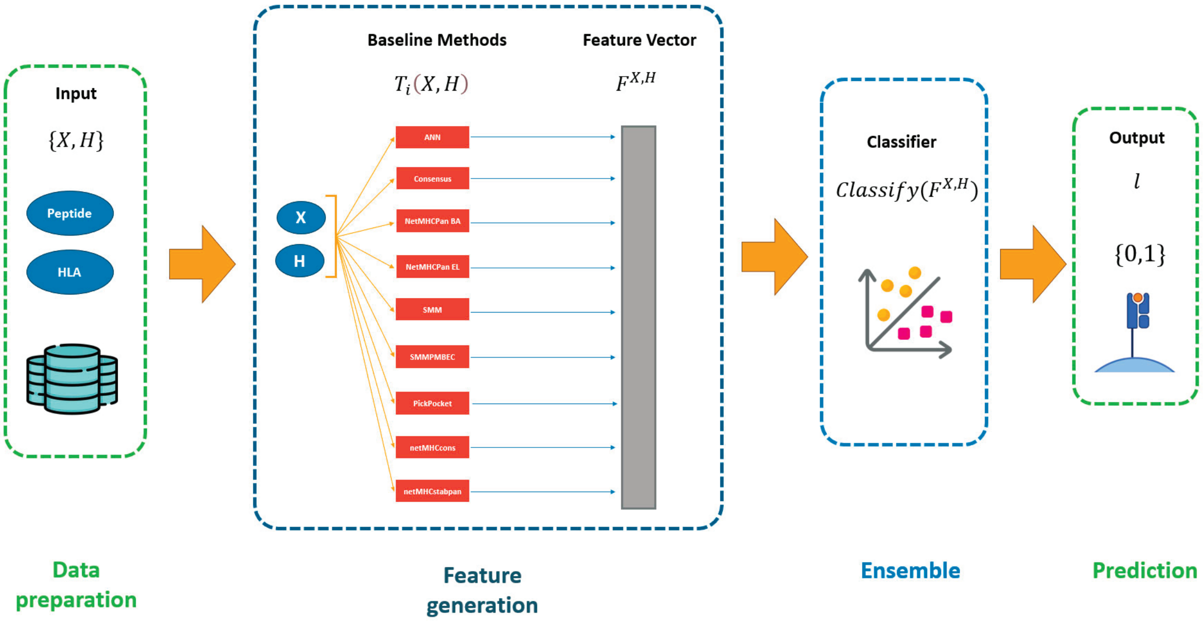


Figure 1. HLAPepBinder pipeline includes three main steps: Data preparation, Feature generation and Ensemble.

position within the sequence *X*.

2. HLA name (*H*): The HLA name follows a standardized nomenclature system that encodes detailed information about the specific HLA allele. For example, the designation ‘HLA-B*15:01’ identifies a variant of the HLA-B gene, where “15” represents its group and “01” denotes the first allele identified within that group. This naming convention is essential for identifying and characterizing specific HLA alleles, which play a crucial role in the immune system and have implications in various fields, including transplantation and disease susceptibility (21).

- **Output:** The prediction of binding between HLA and a peptide is denoted by a binary label named *l* where a value of 1 (or 0) indicates binding (or non-binding), respectively.

3.2. HLAPepBinder Pipeline

In this section, we introduce a pipeline named *HLAPepBinder*, incorporating state-of-the-art algorithms from IEDB to enhance the strengths of different models in IEDB and address their weaknesses. This pipeline consists of three main steps: data preparation, feature generation, and ensemble **(Fig. 1)**. More details are available in the subsequent subsections.

3.2.1. Data preparation

This study used the dataset introduced by Chu *et al.* (20), which was originally employed to benchmark their method, TranspHLA (20) against state-of-the-art models available in the IEDB. To ensure a rigorous performance comparison, they created a dataset named “External test set” that does not share any common data with the training data in TranspHLA (20) and the other models. Moreover, this balanced dataset provides comprehensive insights into the binding of HLA alleles with peptides of varying lengths, ranging from 8 to 12 residues. We define this dataset using the following two ordered lists, *D* and *B* as follows:

$$D = [(X, H)_i]_{i=1}^{n}, \qquad B = [b_i]_{i=1}^{n}, n = 95990$$

where *n* represents the number of data points. For each *i=1,2,…,95990* , we define b_i as:

$$b_i = \begin{cases} 1, & peptide\ X\ binds\ to\ HLA\ H\ in\ pair\ (X, H)_i, \\ 0, & otherwise. \end{cases}$$

We define $(X,H)_i$ as a positive (or negative) pair from the ordered list *D* where $b_i = 1$ (or 0). More details about the dataset are available in **Table 1**. This dataset includes five HLAs: HLA-A*01:01, HLA-A*02:01, HLA-A*24:02, HLA-B*08:01, and HLA-B*18:01.

Table 1. Dataset details of set *D*.

Number of HLAs	Number of peptides	Number of positive pairs	Number of negative pairs	Total number of pairs	Ref.
5	93154	48047	47943	95990	(20)

3.2.2. Feature Generation

Our primary objective is to integrate the outcomes from nine HLA-peptide binding predictor tools available on the IEDB website. The details of these nine tools are as follows:

- ANN (17) is composed of several neural networks, each with a distinct input representation model. The ANN 4.0 (15), also referred to as NetMHC 4.0, is an advanced model including a more intricate architecture, a larger dataset, and an expanded set of training features.
- Consensus (19) is the result of combining three predictors: NetMHC 4.0 (15), SMM (11) and CombLib (22).
- NetMHCpan BA (14) is a pan-specific version of the NetMHC (15) model trained on binding affinity datasets.
- NetMHCpan EL (14) is another version of NetMHCpan, specifically trained on eluted ligand datasets.
- SMM (11) utilizes the stabilized matrix method, operating by minimizing the distance between predicted scores and the measured affinities for the peptides.
- SMMPMBEC (12) is a tool that enhances the SMM method by incorporating a MHC-peptide binding energy covariance (PMBEC) matrix.
- PickPocket (10) predicts the binding between HLA and peptide using pocket similarities.
- NetMHCcons (18) specifically combines three state-of-the-art HLA-peptide binding prediction methods: NetMHC (17), NetMHCpan (14) and PickPocket (10).
- NetMHCstabpan (16) uses neural networks to predict binding affinity and complex stability while considering HLA allelic variability.

Assume that the nine tools are represented in the ordered list *T* as follows:
T = [ANN, Consensus, NetMHCpan BA, NetMHCpan EL, SMM, SMMPMBEC, PickPocket, NetMHCcons, NetMHCstabpan].

Each HLA-peptide pair $(X,H)_i \in D$ is input into the tool T_j to obtain the binding score $\mathrm{f}_j^{(X,H)_i} \in \mathbb{R}^+$,where ranges from 1 to 9. This score predicts the binding affinity between peptide *X* and HLA *H.* Following the computation of $\mathrm{f}_j^{(X,H)_i}$, we generate a feature vector named $F^{(X,H)_i}$ as follows:

$$F^{(X,H)_i} = \left[f_j^{(X,H)_i}\right]_{j=1}^{9}$$

Here, we construct the dataset *U* corresponding to the ordered sets *D* and *B*, to facilitate the training and testing of our pipeline, respectively as follows:

$$\forall i = 1,2,\dots,95990, \qquad (X,H)_i \in D \Longrightarrow F^{(X,H)_i} \in U.$$

3.2.3. Ensemble

In the final phase of the *HLAPepBinder* pipeline, we employ two distinct methods to integrate the binding prediction scores of the peptide *X* to HLA *H* using nine tools in the list *T* for the final output:

- Simple Voting (SV): we make a binary vector $$O^{(X,H)_i} = \left[o_j^{(X,H)_i}\right]_{j=1}^{9}$$ from $$F^{(X,H)_i} = \left[f_j^{(X,H)_i}\right]_{j=1}^{9}$$ as follows: $$o_j^{(X,H)_i} = \begin{cases} 1, & f_j^{(X,H)_i} \geq \theta_j, \\ 0, & otherwise, \end{cases}$$

where θ_j represents a predefined cutoff (20) used by the tool T_j to convert binding prediction scores to 1 or 0, indicating the binding or non-binding between the peptide and HLA, respectively. After making vector $O^{(X,H)_i}$, we predict the binding between *X* and *H* based on simple voting as follows:

$$SV(X,H)_i = \begin{cases} 1, & \sum_{j=1}^{9} o_j^{(X,H)_i} \geq 5, \\ 0, & otherwise. \end{cases}$$

- Random Forest (RF) (23): we divide dataset *U* into two parts. The first part is defined for training the RF model to learn binding or non-binding relationships between peptides and HLA, based on the binding scores of predictors as features; and the other one for testing. The RF approach combines outcomes from the nine predictors to enhance the accuracy. Recognized for their proficiency in ensemble learning, RF models employ multiple decision trees for intricate problem solving. By integrating insights from numerous decision trees, RF forms a robust and flexible model, effectively reducing the common issue of overfitting observed in individual decision trees. At the testing phase of the RF model, each pair *(X,H)* in the test set, is input into the model as F $^{(X,H)}$ for HLA-peptide binding prediction.

4. Results

In this section, we present a comparative analysis of our pipeline, *HLAPepBinder*, against state-of-the-art methods. Additionally, we extend the evaluation by testing *HLAPepBinder* on an independent dataset, demonstrating its efficacy on a larger and more diverse set of data. We also assess *HLAPepBinder*'s performance across different HLA subtypes to evaluate its generalizability. Lastly, we conduct case studies to examine the model's biological relevance and performance.

4.1. Evaluation Metrics

To assess our pipeline, we apply the accuracy, precision, recall, and F1-score metrics as follows (24):

- Accuracy represents the ratio of accurately classified samples to the total number of samples. It is computed by dividing the sum of true positives and true negatives by the overall sample count.
- Precision is the proportion of true positives out of the total number of predicted positives. It measures the ability of the model to correctly identify positive samples. Precision is calculated as the number of true positives divided by the sum of true positives and false positives.
- Recall, also known as sensitivity, is the proportion of true positives out of the total number of actual positives. It measures the ability of the model to correctly identify all positive samples. Recall is calculated as the number of true positives divided by the sum of true positives and false negatives.
- F1-score, the harmonic mean of precision and recall, offers a unified metric balancing both aspects. Its calculation is as follows:

.

$$F1-score=\frac{2}{(\frac{1}{Precision}+\frac{1}{Recall})}$$

4.2. The Assessment of HLAPepBinder Pipeline

As described in the "Material and Methods" section, we incorporate predictions from nine tools sourced from IEDB using our pipeline for HLA-peptide binding prediction. The primary aim of this section is to demonstrate that integration based on the RF model outperforms the SV model. To achieve this, we establish two versions of the *HLAPepBinder* pipeline: *HLAPepBinder_SV* and *HLAPepBinder_RF*.

We perform 5-fold cross-validation on dataset *U* (refer to the second phase of the pipeline) using *HLAPepBinder_RF*. In parallel, HLA-peptide binding predictions for each test fold are generated using *HLAPepBinder_SV*. **Table 2** presents the performance comparison of the two models, highlighting that the RF-based integration of the nine predictors significantly outperforms the SV model. Hereafter, we will refer to the RF-based model simply as *HLAPepBinder*.

Table 2. Performance comparison of RF and SV models in the Ensemble phase of HLAPepBinder.

Method	Accuracy	Precision	Recall	F1-Score
HLAPepBinder_SV	0.726	0.984	0.460	0.627
HLAPepBinder_RF	0.916	0.927	0.902	0.915

The importance of the nine tools was analyzed using RF **(Fig. S1).** This analysis highlights that the NetMHCpan EL model (14) plays a particularly influential role in predictive performance, aligning with the IEDB recommendation and emphasizing its importance in HLA-peptide binding prediction. Furthermore, the analysis reveals that earlier consensus models, such as NetMHCcons (18) and Consensus (19), contribute less substantially to the predictive process when compared to the IEDB-recommended method, NetMHCpan model (14).

4.3. Comparing the HLAPepBinder Pipeline Against State-of-the-Art Methods

In this section, we apply 5-fold cross-validation on dataset *U* (refer to the second phase of the pipeline) to validate the efficacy of *HLAPepBinder* against nine baseline methods sourced from IEDB. **Figure 2** presents the evaluation metrics comparing the prediction performance across these methods.

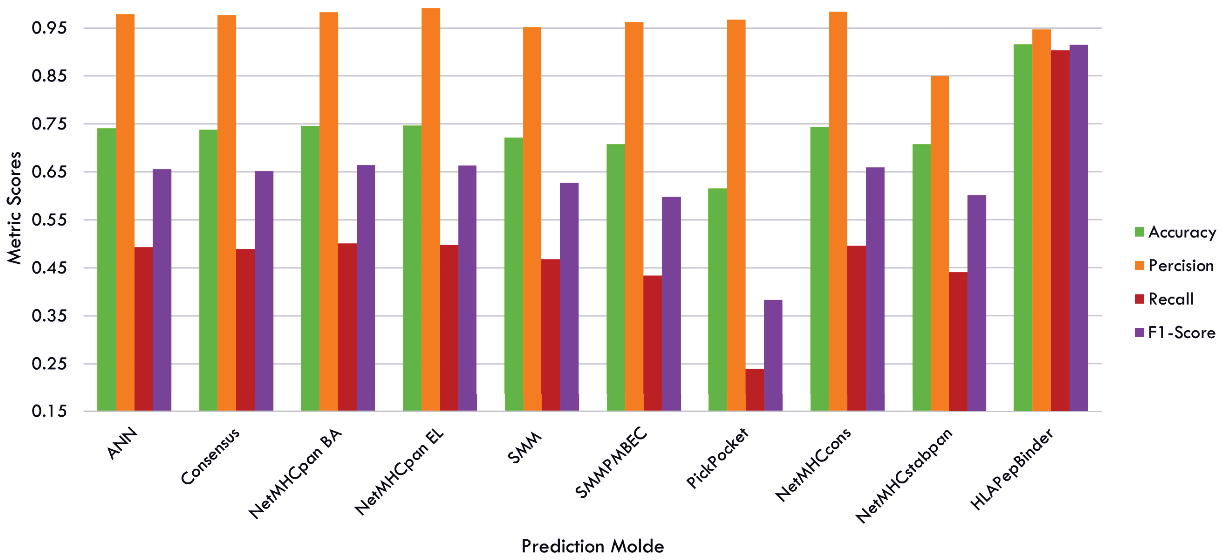


Figure 2. Comparing the performance of HLAPepBinder against nine baseline models from IEDB.

As illustrated in **Figure 2,** the *HLAPepBinder* consistently outperforms all baseline models across four criteria. It achieves an accuracy of 0.916, which is significantly higher than NetMHCpan EL (0.746) (14), the recommended model in IEDB. Furthermore, our pipeline, serving as a consensus model, exhibits higher performance compared to previous consensus models such as NetMHCcons (18) and Consensus (19). Unlike previous models that achieved high precision with a low F1-score, our model has successfully enhanced both precision and F1-score, reaching impressive values of ≈ 0.92 for both metrics.

Previous models have exhibited high precision but reported very low recall values. Given the definitions of these metrics, this suggests that while the models are effective in accurately identifying the bound pairs they predict (resulting in high precision), they miss a substantial number of actual bound pairs within the test dataset, leading to reduced recall. In such scenarios, the F1-score becomes particularly relevant, as it provides a balanced measure by representing the harmonic mean of precision and recall. The *HLAPepBinder* model has demonstrated significantly improved recall and, consequently, a higher F1-score. This improvement highlights that each of the baseline models captures a different subset of bound pairs, and combining these models in an ensemble has significantly enhanced overall performance.

4.4. Comparing the HLAPepBinder Pipeline Against State-of-the-Art Models on An Independent Dataset

Chu *et al.* (20) utilized a dataset referred to as the "Independent test dataset," distinct from set *D*, to compare their model named TranspHLA (20) to the baseline predictors in IEDB. In this section, our primary objective is to employ this dataset as an external test set for comparing *HLAPepBinder* against the baseline predictors in IEDB (14,18,19) and TranspHLA (20).

Table 3. Dataset details of set *E*.

Number of HLAs	Number of peptides	Number of positive pairs	Number of negative pairs	Total number of pairs	Ref.
73	115953	62100	61808	123908	(20)

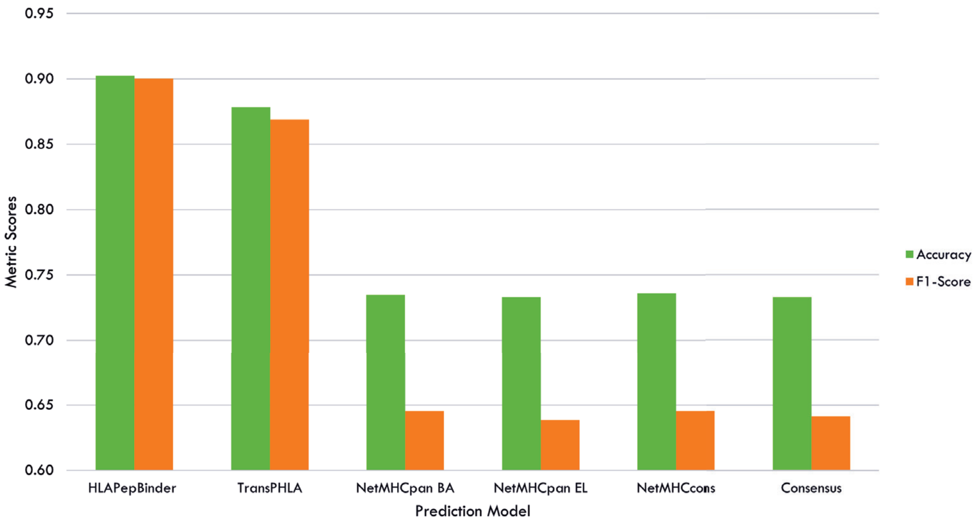


Figure 3. Performance comparison of HLAPepBinder with TranspHLA, consensus-based models, and the IEDB-recommended models in terms of accuracy and F1-score.

Notably, none of these models, including our own, have been trained on any part of this test data. We call this set *E*, with further details available in **Table 3.** *HLAPepBinder* is trained using dataset *U*, which is derived from the results of the nine baseline models applied to the set *D*. We then use it for HLA-peptide binding prediction on the independent test set *E*. **Figure 3** presents the outcomes of our pipeline in comparison to the results from other tools, as reported in the paper (20).

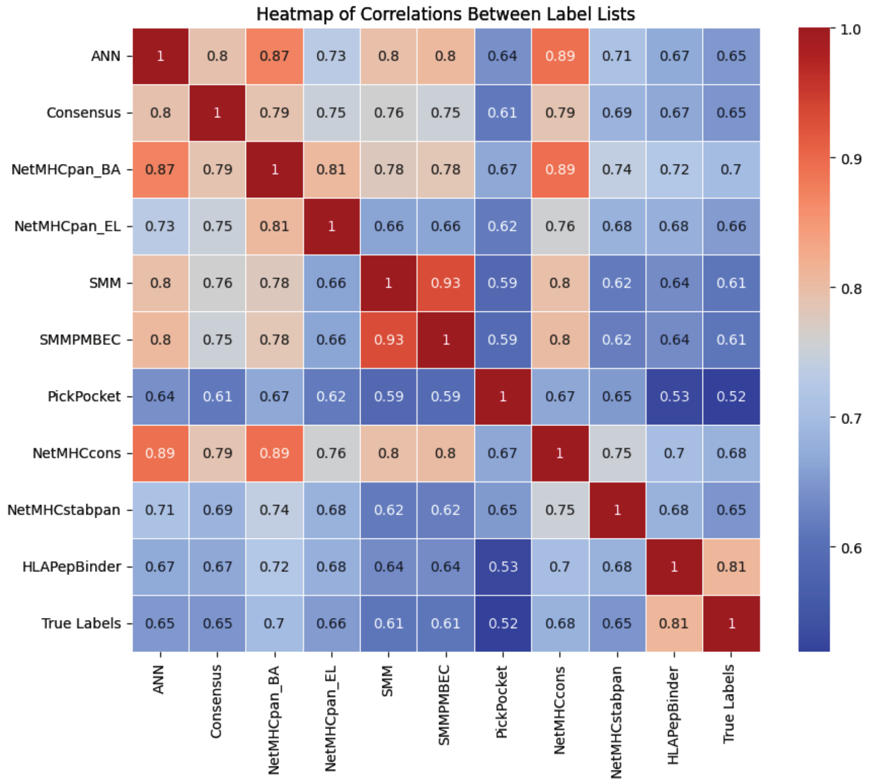
As illustrated in **Figure 3**, *HLAPepBinder* presents superior performance in comparison to the recently introduced TranspHLA model (20), as well as the consensus models found in IEDB (NetMHCcons (18) and Consensus (19)) and the recommended IEDB model, NetMHCpan EL (18). Furthermore, in addition to its high accuracy, it's noteworthy that *HLAPepBinder* significantly outperforms TranspHLA in terms of processing speed and resource consumption, in addition to its high accuracy.

To further evaluate the contribution of each base model within *HLAPepBinder*, we present a heatmap in **Figure 4**, depicting the prediction correlations between each of the nine baseline models and the true labels in dataset *E*. The heatmap demonstrates that *HLAPepBinder* exhibits significant correlations with each individual model, with none of the correlations being zero. Additionally, no correlation approaches one, indicating that no single model dominates the overall prediction. This finding confirms that *HLAPepBinder* effectively synthesizes diverse predictions from all nine models, rather than relying heavily on any single predictor. Moreover, *HLAPepBinder* shows the highest correlation with the true labels, highlighting its superior predictive performance and accuracy.

4.5. HLAPepBinder Performance on HLA Subtypes

In this section, we conduct a series of experiments to evaluate the generalizability of *HLAPepBinder* across different HLA subtypes using a leave-one-out validation approach. This approach enables us to assess how well the model could predict HLA-peptide binding for subtypes it has not encountered during training. Meanwhile, we evaluate the model's ability to predict the binding for subtypes that are absent from the training set but present in the test set. This test is key to determine whether *HLAPepBinder* could generalize its predictions to novel HLA subtypes. The results **(Table S1)** show consistent performance across different HLA subtypes, highlighting the model's robust generalizability across all cases.

This evaluation is also conducted using the external test dataset *(E)*. We examine *HLAPepBinder*'s performance across the three HLA groups—A, B, and C—excluding the specific A and B subtypes present in the main training set. The results, presented in



	ANN	Consensus	NetMHCpan_BA	NetMHCpan_EL	SMM	SMMPMBEC	PickPocket	NetMHCcons	NetMHCstabpan	HLAPepBinder	True Labels
ANN	1	0.8	0.87	0.73	0.8	0.8	0.64	0.89	0.71	0.67	0.65
Consensus	0.8	1	0.79	0.75	0.76	0.75	0.61	0.79	0.69	0.67	0.65
NetMHCpan_BA	0.87	0.79	1	0.81	0.78	0.78	0.67	0.89	0.74	0.72	0.7
NetMHCpan_EL	0.73	0.75	0.81	1	0.66	0.66	0.62	0.76	0.68	0.68	0.66
SMM	0.8	0.76	0.78	0.66	1	0.93	0.59	0.8	0.62	0.64	0.61
SMMPMBEC	0.8	0.75	0.78	0.66	0.93	1	0.59	0.8	0.62	0.64	0.61
PickPocket	0.64	0.61	0.67	0.62	0.59	0.59	1	0.67	0.65	0.53	0.52
NetMHCcons	0.89	0.79	0.89	0.76	0.8	0.8	0.67	1	0.75	0.7	0.68
NetMHCstabpan	0.71	0.69	0.74	0.68	0.62	0.62	0.65	0.75	1	0.68	0.65
HLAPepBinder	0.67	0.67	0.72	0.68	0.64	0.64	0.53	0.7	0.68	1	0.81
True Labels	0.65	0.65	0.7	0.66	0.61	0.61	0.52	0.68	0.65	0.81	1

Figure 4. Heatmap of prediction correlations between HLAPepBinder and baseline models with true labels in the dataset *E*.

Table 4, demonstrate the model's robustness and ability to generalize to unseen HLA types, especially within HLA group C, which was entirely absent from the training set.

Table 4. Performance of HLAPepBinder on the set *E* across HLA groups A, B, and C.

HLA Group	Accuracy	Precision	Recall	F1-score
HLA-A	0.903	0.914	0.890	0.902
HLA-B	0.905	0.903	0.908	0.906
HLA-C	0.915	0.914	0.918	0.916

4.6. Case Study

One of the primary challenges we encounter is the lack of experimental validation for non-binding HLA-peptide interactions. In essence, the negative data within our dataset remains unknown, leaving us without concrete examples of non-binding interactions. As a result, when predicting unknown HLA-peptide interactions, we must determine the validity of these predictions. To enhance our RF model, we modify it to output probabilities instead of binary classifications. This is achieved by calculating the proportion of votes for each class, which involves dividing the number of votes a class receives by the total number of trees in the forest.

From the test set, we identify HLA-peptide pairs that our model predicts with 100% probability, totaling 26 pairs across five distinct HLA subtypes. We evaluate these pairs to demonstrate the reliability of our model **(Table 5)**. Additionally, we discover three HLA-peptide pairs in the literature that exhibit satisfactory experimental binding (25–33). Notably, two of these peptides have been previously reported to bind to HLA class I subtypes.

To further assess the remaining peptides, we utilize BLAST with two sequence similarity thresholds—90% and 80%—and identify 13 additional pairs. For the

Table 5. Summary of HLA-peptide interaction analyses.

HLA	Peptide	Validation	Details	Reference
HLA-A*02:01	ALAQYLITA	Literature	Potential targets for immunotherapy	(25,26)
HLA-A*24:02	LIPEEFFQF	Literature	Potential targets for immunotherapy	(27–31)
HLA-B*08:01	NPTERVAAL	Literature	Role in Allorecognition	(32,33)
HLA-B*18:01	DEYIERLVW	Known peptide	Reported related to HLA class I	(35–38)
HLA-B*18:01	SEAGTHQEW	Known peptide	Reported related to HLA class I	(39)
HLA-B*08:01	DLIKFIMSL	BLAST Similarity (90%)	Reported related to HLA class I	(40–43)
HLA-B*08:01	ELMAHLTEM	BLAST Similarity (90%)	Reported related to HLA class I	(40,44–46)
HLA-B*08:01	FPARLKKVL	BLAST Similarity (90%)	Reported related to HLA class I	(27,28,47–49)
HLA-B*18:01	VETGVELM	BLAST Similarity (90%)	Reported related to HLA class I	(37)
HLA-B*08:01	EGTLRRRSL	BLAST Similarity (80%)	Reported related to HLA class I	(50)
HLA-A*01:01	KTDSSFSFM	BLAST Similarity (80%)	Reported related to HLA class I	(29,51)
HLA-A*24:02	EYAAVAQEL	BLAST Similarity (80%)	Reported related to HLA-C	(52)
HLA-B*18:01	EEKQHLLFM	BLAST Similarity (90%)	Reported related to HLA-B	(31,53–55)
HLA-B*08:01	EPAGRPPAL	BLAST Similarity (90%)	Reported related to HLA-B	(39,56)
HLA-B*18:01	TEPPFSGIY	BLAST Similarity (90%)	Reported related to HLA-B	(57)
HLA-B*18:01	EDFRYGYSY	BLAST Similarity (80%)	Reported related to HLA-B	(58)
HLA-B*08:01	KKKLRTLQL	BLAST Similarity (80%)	Reported related to HLA-B	(58)
HLA-B*08:01	DLRHYLSL	BLAST Similarity (90%)	Reported related to HLA-B*08:01	(59)
HLA-B*08:01	HLRNHQQI	Molecular Docking	Docking Score: -243.324	(Fig. 5A)
HLA-B*08:01	TLEPRGYSL	Molecular Docking	Docking Score: -243.637	(Fig. 5B)
HLA-B*18:01	EEGHVAVF	Molecular Docking	Docking Score: -241.008	(Fig. 5C)
HLA-B*18:01	NEYEVYSL	Molecular Docking	Docking Score: -239.856	(Fig. 5D)
HLA-B*18:01	VEFLGPVAL	Molecular Docking	Docking Score: -237.400	(Fig. 5E)
HLA-B*08:01	EPGGRPFYL	Molecular Docking	Docking Score: -235.915	(Fig. 5F)
HLA-B*08:01	YLKKHGIDV	Molecular Docking	Docking Score: -215.966	(Fig. 5G)
HLA-B*08:01	TKRRSPSL	Molecular Docking	Docking Score: -210.840	(Fig. 5H)

final eight pairs, we employ molecular docking with the HPEPDOCK tool (34) to predict potential binding interactions. A comprehensive summary of these analyses can be found in **Table 5**.

Molecular docking is a computational technique used to predict the binding affinity and interactions between a peptide and an HLA molecule. HPEPDOCK (34) simulates this binding process and provides a binding score, which helps assess the likelihood of HLA-peptide binding. The server generates the top ten docking models, and we select the highest-ranked model for detailed analysis. **Figure 5** and **Figure S2** illustrate the molecular docking results for the HLA-peptide pairs and presents the cartoon representation of the HLA-peptide complex. All eight pairs achieved docking scores below -200, indicating a strong likelihood of binding. Both figures demonstrate that the peptides are positioned within the grooves of the HLA molecules, further supporting the predicted binding interactions.

5. Discussion

The *HLAPepBinder* pipeline demonstrates a significant advancement in HLA-peptide binding prediction by effectively integrating predictions from nine state-of-the-art tools through a RF model. The results of our analyses establish HLAPepBinder_RF as a superior consensus model, outperforming both its SV-based counterpart (HLAPepBinder_SV) and individual baseline methods across various evaluation metrics. The comparison between HLAPepBinder_RF and HLAPepBinder_SV highlights the robustness of

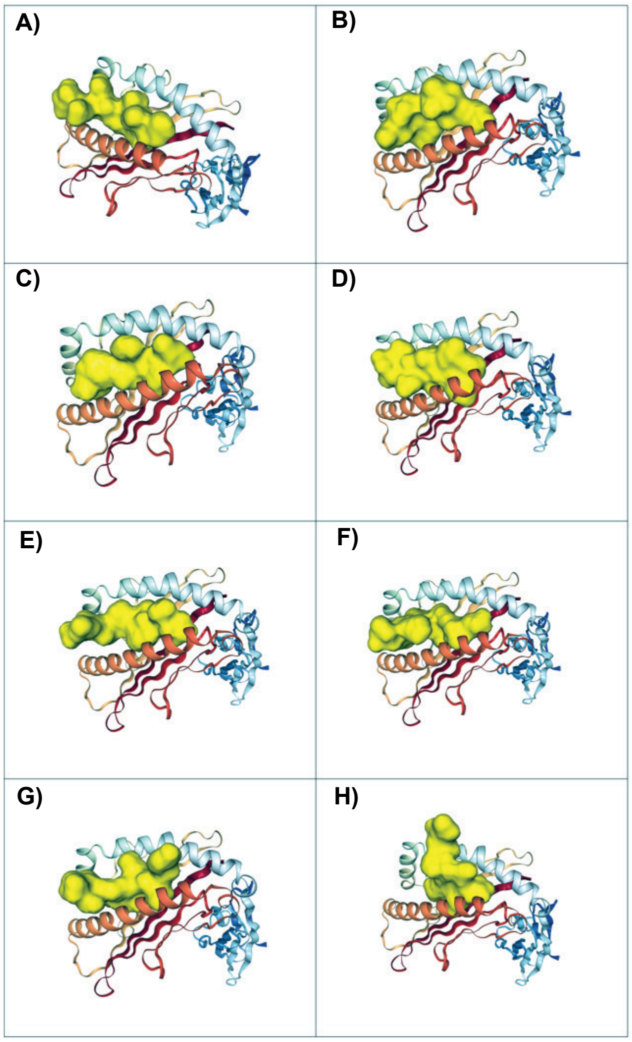


Figure 5. **Cartoon representation of molecular docking between HLA and peptide, highlighting the peptide's alignment within the HLA groove. A)** HLA-B*08:01 with HLRNHQQI, **B)** HLA-B*08:01 with TLEPRGYSL, **C)** HLA-B*18:01 with EEGHVAVF, **D)** HLA-B*18:01 with NEYEVYSL, **E)** HLA-B*18:01 with VEFLGPVAL, **F)** HLA-B*08:01 with EPGGRP-FYL, **G)** HLA-B*08:01 with YLKKHGIDV, and **H)** HLA-B*08:01 with TKRRSPSL.

the RF-based approach. As evidenced by the 5-fold cross-validation results on dataset , HLAPepBinder_RF achieves significantly higher accuracy (0.916), precision (0.927), recall (0.902), and F1-score (0.915) compared to HLAPepBinder_SV, which is limited by its low recall and F1-score. This improvement underscores the effectiveness of ensemble methods, particularly RF, in synthesizing diverse predictive patterns from multiple models. HLAPepBinder_RF achieves a better balance of precision and recall, addressing the limitations of earlier models that focus on precision but lack recall.

The feature importance analysis using the RF model provides valuable insights into the contributions of individual predictors. The NetMHCpan EL (14) model emerges as the most influential baseline tool. In contrast, earlier consensus models, such as NetMHCcons (18) and Consensus (19), play a relatively minor role in the *HLAPepBinder* pipeline. This finding supports the hypothesis that using the predictive strengths of specific advanced models, such as NetMHCpan EL, can significantly enhance overall performance.

HLAPepBinder consistently outperforms all baseline methods and consensus models from IEDB in terms of accuracy and F1-score, as shown in **Figure 2**. Moreover, it surpasses the performance of TranspHLA (20) on the independent test set *E*, achieving both higher accuracy and computational efficiency. This success highlights the robustness of the *HLAPepBinder* pipeline as a consensus-based approach, capable of harnessing the predictive diversity of individual models to deliver superior results.

The leave-one-out validation experiments reveal that *HLAPepBinder* maintains high predictive accuracy across various HLA subtypes. The model consistently achieves strong performance metrics even for subtypes absent in the training set, as shown in **Tables 4** and **Table S1**. Notably, the model exhibits exceptional generalizability to HLA group C, which was entirely excluded from the training phase, achieving an F1-score of 0.916 on test set *E*. This finding underscores the model's potential for application to novel HLA subtypes and its ability to generalize effectively to unseen data.

The correlation heatmap in **Figure 4** illustrates the contribution of each baseline model within *HLAPepBinder*. The absence of dominant correlations highlights the model's ability to synthesize diverse predictions rather than relying disproportionately on any single predictor. The high correlation between *HLAPepBinder* predictions and true labels further validates its superior predictive performance.

A major challenge in HLA-peptide binding prediction is the scarcity of experimentally verified negative data, as many interactions labeled as non-binding might actually be true binders, but remain experimentally unverified. Using *HLAPepBinder*, we identified 26 HLA-peptide pairs in the test set that are predicted with 100% certainty to bind. These predictions are validated through literature review, BLAST sequence similarity analysis, and molecular docking **(Table 5)**. Also, these literature review validated predictions include several promising candidates for immunotherapy and vaccine development (25–33). Notably, eight unverified pairs demonstrated strong binding potential with docking scores below -200, further reinforcing the model's reliability. This integration of molecular docking results with computational predictions bridges the gap between in silico methods and experimental validation, offering a robust evaluation framework.

6. Conclusions

Our research introduced a novel ensemble-learning model for HLA-peptide binding prediction, addressing a critical need in immunoinformatics. IEDB's website models are considered reliable due to their large and curated dataset, rigorous curation processes, peer review and regular updates. Model selection has been a significant challenge, as manually evaluating prediction models can be time-consuming and prone to errors. Our approach, known as an ensemble model, simplifies this process by providing a systematic way to use the strengths of existing predictors. This not only enhances predictive accuracy but also minimizes time investment and reduces error rates. It's worth noting that recent deep learning models are resource-expensive and time-consuming.

The core achievement of our research lied in the significantly improved predictive accuracy of *HLAPepBinder*. Our ensemble model consistently outperformed individual baseline models and alternative methods across various criteria. This heightened accuracy proves instrumental in understanding immune responses, optimizing vaccine design, and advancing therapeutic development, offering more reliable predictions of HLA-peptide binding. Beyond accuracy, our approach introduces efficiency and cost-effectiveness. By utilizing IEDB predictors, we not only build on collective expertise but also minimize the resources required for model development.

This enhancement of ensemble models, extending HLA-peptide binding prediction, offers valuable insights for addressing various bioinformatics challenges and paving the way for future research. In other words, there is potential in model ensembling related to HLA class II, improving the interpretability of these models, and optimizing their implementation across different computational environments. These areas represent promising opportunities for further investigation and advancement in the field.

References

1. Yewdell JW, Bennink JR. Immunodominance in major histocompatibility complex class I-restricted T lymphocyte responses. *Annu Rev Immunol*. 1999;**17**(1):51-88. doi: 10.1146/annurev.immunol.17.1.51
2. Najafimehr H, Hajizadeh N, Nazemalhosseini-Mojarad E, Pourhoseingholi MA, Abdollahpour-Alitappeh M, Ashtari S, *et al*. The role of Human leukocyte antigen class I on patient survival in Gastrointestinal cancers: a systematic review and meta- analysis. *Sci Rep*. 2020;**10**(1):728. doi: 10.1038/s41598-020-57582-x
3. Aptsiauri N, Garrido F. The Challenges of HLA Class I Loss in Cancer Immunotherapy: Facts and Hopes. *Clinic Can Res*. 2022;**28**(23):5021-5029. doi: 10.1158/1078-0432.CCR-21-3501
4. Chowell D, Morris LGT, Grigg CM, Weber JK, Samstein RM, Makarov V, *et al*. Patient HLA class I genotype influences cancer response to checkpoint blockade immunotherapy. *Science* (1979). 2018;**359**(6375):582-587. doi: 10.1126/science.aao4572
5. Imahashi N, Nishida T, Ito Y, Kawada J, Nakazawa Y, Toji S, *et al*. Identification of a novel HLA-A*24:02-restricted adenovirus serotype 11-specific CD8+ T-cell epitope for adoptive immunotherapy. *Mol Immunol*. 2013;**56**(4):399-405. doi: 10.1016/j.molimm.2013.05.232
6. Garrido F. HLA Class-I Expression and Cancer Immunotherapy. In: MHC Class-I Loss and Cancer Immune Escape. 2019.**p.** 79-90. doi: 10.1007/978-3-030-17864-2_3
7. Liu J, Fu M, Wang M, Wan D, Wei Y, Wei X. Cancer vaccines as promising immuno-therapeutics: platforms and current progress. *J Hematol Oncol*. 2022;**15**(1):28. doi: 10.1186/s13045-022-01247-x
8. Feola S, Chiaro J, Martins B, Russo S, Fusciello M, Ylösmäki E, *et al*. A novel immunopeptidomic-based pipeline for the generation of personalized oncolytic cancer vaccines. *Elife*. 2022;11. doi: 10.7554/eLife.71156
9. Dhanda SK, Mahajan S, Paul S, Yan Z, Kim H, Jespersen MC, *et al*. IEDB-AR: immune epitope database—analysis resource in 2019. Nucleic Acids Res. 2019 Jul 2;**47**(W1):W502–W506. doi: 10.1093/nar/gkz452
10. 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*. 2009;**25**(10):1293-1299. doi: 10.1093/bioinformatics/btp137
11. Peters B, Sette A. Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method. *BMC Bioinformatics*. 2005;**6**(1):132. doi: 10.1186/1471-2105-6-132
12. Kim Y, Sidney J, Pinilla C, Sette A, Peters B. Derivation of

an amino acid similarity matrix for peptide:MHC binding and its application as a Bayesian prior. *BMC Bioinformatics*. 2009;**10**(1):394. doi: 10.1186/1471-2105-10-394

13. Lundegaard C, Lamberth K, Harndahl M, Buus S, Lund O, Nielsen M. NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8–11. *Nucleic Acids Res*. 2008;**36**(suppl_2):W509-W512. doi: 10.1093/nar/gkn202
14. Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. NetMHCpan-4.0: Improved Peptide–MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *J Immunol*. 2017;**199**(9):3360-3368. doi: 10.4049/jimmunol.1700893
15. Andreatta M, Nielsen M. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics*. 2016;**32**(4):511-517. doi: 10.1093/bioinformatics/btv639
16. Rasmussen M, Fenoy E, Harndahl M, Kristensen AB, Nielsen IK, Nielsen M, *et al*. Pan-Specific Prediction of Peptide–MHC Class I Complex Stability, a Correlate of T Cell Immunogenicity. *J Immunol*. 2016;**197**(4):1517-1524. doi: 10.4049/jimmunol.1600582
17. Nielsen M, Lundegaard C, Worning P, Lauemøller SL, Lamberth K, Buus S, *et al*. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci*. 2003;**12**(5):1007-1017. doi: 10.1110/ps.0239403
18. Karosiene E, Lundegaard C, Lund O, Nielsen M. NetMHCcons: a consensus method for the major histocompatibility complex class I predictions. *Immunogen*. 2012;**64**(3):177-186. doi: 10.1007/s00251-011-0579-8
19. Moutaftsi M, Peters B, Pasquetto V, Tscharke DC, Sidney J, Bui H-H, *et al*. A consensus epitope prediction approach identifies the breadth of murine TCD8+-cell responses to vaccinia virus. *Nat Biotechnol*. 2006;**24**(7):817-819. doi: 10.1038/nbt1215
20. Chu Y, Zhang Y, Wang Q, Zhang L, Wang X, Wang Y, *et al*. A transformer-based model to predict peptide–HLA class I binding and optimize mutated peptides for vaccine design. *Nat Mach Intell*. 2022;**4**(3):300-311.
21. Luo Y, Kanai M, Choi W, Li X, Sakaue S, Yamamoto K, *et al*. A high-resolution HLA reference panel capturing global population diversity enables multi-ancestry fine-mapping in HIV host response. *Nat Genet*. 2021;**53**(10):1504-1516. doi: 10.1038/s41588-021-00935-7
22. Sidney J, Assarsson E, Moore C, Ngo S, Pinilla C, Sette A, *et al*. Quantitative peptide binding motifs for 19 human and mouse MHC class I molecules derived using positional scanning combinatorial peptide libraries. *Immunome Res*. 2008;**4**(1):2. doi: 10.1186/1745-7580-4-2
23. Tin Kam Ho. The random subspace method for constructing decision forests. *IEEE Trans Pattern Anal Mach Intell*. 1998;**20**(8):832-844. doi: 10.1109/34.709601
24. Metz CE. Basic principles of ROC analysis. *Semin Nucl Med*. 1978;**8**(4):283-298. doi: 10.1016/S0001-2998(78)80014-2
25. Faridi P, Woods K, Ostrouska S, Deceneux C, Aranha R, Duscharla D, *et al*. Spliced Peptides and Cytokine-Driven Changes in the Immunopeptidome of Melanoma. *Cancer Immunol Res*. 2022;**8**(10):1322-1334. doi: 10.1158/2326-6066.CIR-19-0894
26. Vadakekolathu J, Boocock DJ, Pandey K, Guinn B-A, Legrand A, Miles AK, *et al*. Multi-Omic Analysis of Two Common P53 Mutations: Proteins Regulated by Mutated P53 as Potential Targets for Immunotherapy. *Cancers* (Basel). 2022;**14**(16). doi: 10.3390/cancers14163975
27. Gfeller D, Guillaume P, Michaux J, Pak H-S, Daniel RT, Racle J, *et al*. The Length Distribution and Multiple Specificity of Naturally Presented HLA-I Ligands. *J Immunol*. 2018;**201**(12):3705-3716. doi: 10.4049/jimmunol.1800914
28. Solleder M, Guillaume P, Racle J, Michaux J, Pak H-S, Müller M, *et al*. Mass Spectrometry Based Immunopeptidomics Leads to Robust Predictions of Phosphorylated HLA Class I Ligands. *Mol Cell Proteomics*. 2020;**19**(2):390-404. doi: 10.1074/mcp.TIR119.001641
29. Marcu A, Bichmann L, Kuchenbecker L, Kowalewski DJ, Freudenmann LK, Backert L, *et al*. HLA Ligand Atlas: a benign reference of HLA-presented peptides to improve T-cell-based cancer immunotherapy. *J Immunother Cancer*. 2021(4). doi: 10.1136/jitc-2020-002071
30. Mei S, Ayala R, Ramarathinam SH, Illing PT, Faridi P, Song J, *et al*. Immunopeptidomic Analysis Reveals That Deamidated HLA-bound Peptides Arise Predominantly from Deglycosylated Precursors. *Mol Cell Proteomics*. 2020;**19**(7):1236-1247. doi: 10.1074/mcp.RA119.001846
31. Sarkizova S, Klaeger S, Le PM, Li LW, Oliveira G, Keshishian H, *et al*. A large peptidome dataset improves HLA class I epitope prediction across most of the human population. *Nat Biotechnol*. 2020;**38**(2):199-209. doi: 10.1038/s41587-019-0322-9
32. Laumont CM, Daouda T, Laverdure J-P, Bonneil É, Caron-Lizotte O, Hardy M-P, *et al*. Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames. *Nat Commun*. 2016;7:10238. doi: 10.1038/ncomms10238
33. Granados DP, Sriranganadane D, Daouda T, Zieger A, Laumont CM, Caron-Lizotte O, *et al*. Impact of genomic polymorphisms on the repertoire of human MHC class I-associated peptides. *Nat Commun*. 2014;5:3600. doi: 10.1038/ncomms4600
34. Zhou P, Jin B, Li H, Huang S-Y. HPEPDOCK: a web server for blind peptide-protein docking based on a hierarchical algorithm. *Nucleic Acids Res*. 2018;**46**(W1):W443-W450. doi: 10.1093/nar/gky357
35. Neidert MC, Kowalewski DJ, Silginer M, Kapolou K, Backert L, Freudenmann LK, *et al*. The natural HLA ligandome of glioblastoma stem-like cells: antigen discovery for T cell-based immunotherapy. *Acta Neuropathol*. 2018;**135**(6):923-938. doi: 10.1007/s00401-018-1836-9
36. Lanoix J, Durette C, Courcelles M, Cossette É, Comtois-Marotte S, Hardy M-P, *et al*. Comparison of the MHC I Immunopeptidome Repertoire of B-Cell Lymphoblasts Using Two Isolation Methods. *Proteomics*. 2018;**18**(12):e1700251. doi: 10.1002/pmic.201700251
37. Bassani-Sternberg M, Chong C, Guillaume P, Solleder M, Pak H, Gannon PO, *et al*. Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allostery regulating HLA specificity. *PLoS Comput Biol*. 2017;**13**(8):e1005725. doi: 10.1371/journal.pcbi.1005725
38. Venema WJ, Hiddingh S, de Boer JH, Claas FHJ, Mulder A, den Hollander AI, *et al*. ERAP2 Increases the Abundance of a Peptide Submotif Highly Selective for the Birdshot Uveitis-Associated HLA-A29. *Front Immunol*. 2021;**12**:634441. doi: 10.3389/fimmu.2021.634441
39. Faridi P, Woods K, Ostrouska S, Deceneux C, Aranha R, Duscharla D, *et al*. Spliced Peptides and Cytokine-Driven

Changes in the Immunopeptidome of Melanoma. *Cancer Immunol Res*. 2020;**8**(10):1322-1334. doi: 10.1158/2326-6066.CIR-19-0894

40. Nicholas B, Bailey A, Staples KJ, Wilkinson T, Elliott T, Skipp P. Immunopeptidomic analysis of influenza A virus infected human tissues identifies internal proteins as a rich source of HLA ligands. *PLoS Pathog*. 2022;**18**(1):e1009894. doi: 10.1371/journal.ppat.1009894
41. Chong C, Marino F, Pak H, Racle J, Daniel RT, Müller M, *et al*. High-throughput and Sensitive Immunopeptidomics Platform Reveals Profound Interferonγ-Mediated Remodeling of the Human Leukocyte Antigen (HLA) Ligandome. *Mol Cell Proteomics*. 2018;**17**(3):533-548. doi: 10.1074/mcp.TIR117.000383
42. Pandey K, Mifsud NA, Lim Kam Sian TCC, Ayala R, Ternette N, Ramarathinam SH, *et al*. In-depth mining of the immunopeptidome of an acute myeloid leukemia cell line using complementary ligand enrichment and data acquisition strategies. *Mol Immunol*. 2020;**123**:7-17. doi: 10.1016/j.molimm.2020.04.008
43. Mayer RL, Verbeke R, Asselman C, Aernout I, Gul A, Eggermont D, *et al*. Immunopeptidomics-based design of mRNA vaccine formulations against Listeria monocytogenes. *Nat Commun*. 2022;**13**(1):6075. doi: 10.1038/s41467-022-33721-y
44. Olsson N, Schultz LM, Zhang L, Khodadoust MS, Narayan R, Czerwinski DK, *et al*. T-Cell Immunopeptidomes Reveal Cell Subtype Surface Markers Derived From Intracellular Proteins. *Proteomics*. 2018;**18**(12):e1700410. doi: 10.1002/pmic.201700410
45. Chong C, Müller M, Pak H, Harnett D, Huber F, Grun D, *et al*. Integrated proteogenomic deep sequencing and analytics accurately identify non-canonical peptides in tumor immunopeptidomes. *Nat Commun*. 2020;**11**(1):1293. doi: 10.1038/s41467-020-14968-9
46. Ruiz Cuevas MV, Hardy M-P, Hollý J, Bonneil É, Durette C, Courcelles M, *et al*. Most non-canonical proteins uniquely populate the proteome or immunopeptidome. *Cell Rep*. 2021;**34**(10):108815. doi: 10.1016/j.celrep.2021.108815
47. Gloger A, Ritz D, Fugmann T, Neri D. Mass spectrometric analysis of the HLA class I peptidome of melanoma cell lines as a promising tool for the identification of putative tumor-associated HLA epitopes. *Cancer Immunol Immunother*. 2016;**65**(11):1377-1393. doi: 10.1007/s00262-016-1897-3
48. Chong C, Müller M, Pak H, Harnett D, Huber F, Grun D, *et al*. Integrated proteogenomic deep sequencing and analytics accurately identify non-canonical peptides in tumor immunopeptidomes. *Nat Commun*. 2020;**11**(1):1293. doi: 10.1038/s41467-020-14968-9
49. Bassani-Sternberg M, Pletscher-Frankild S, Jensen LJ, Mann M. Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. *Mol Cell Proteomics*. 2015;**14**(3):658-673. doi: 10.1074/mcp.M114.042812
50. Miller AM, Koşaloğlu-Yalçın Z, Westernberg L, Montero L, Bahmanof M, Frentzen A, *et al*. A functional identification platform reveals frequent, spontaneous neoantigen-specific T cell responses in patients with cancer. *Sci Transl Med*. 2024;**16**(736):eabj9905. doi: 10.1126/scitranslmed.abj9905
51. Weingarten-Gabbay S, Klaeger S, Sarkizova S, Pearlman LR, Chen D-Y, Gallagher KME, *et al*. Profiling SARS-CoV-2 HLA-I peptidome reveals T cell epitopes from out-of-frame ORFs. *Cell*. 2021;**184**(15):3962-3980.e17. doi: 10.1016/j.cell.2021.05.046
52. Di Marco M, Schuster H, Backert L, Ghosh M, Rammensee H-G, Stevanović S. Unveiling the Peptide Motifs of HLA-C and HLA-G from Naturally Presented Peptides and Generation of Binding Prediction Matrices. *J Immunol*. 2017;**199**(8):2639-2651. doi: 10.4049/jimmunol.1700938
53. Komov L, Melamed Kadosh D, Barnea E, Admon A. The Effect of Interferons on Presentation of Defective Ribosomal Products as HLA Peptides. *Mol Cell Proteomics*. 2021;**20**:100105. doi: 10.1016/j.mcpro.2021.100105
54. Goncalves G, Mullan KA, Duscharla D, Ayala R, Croft NP, Faridi P, *et al*. IFNγ Modulates the Immunopeptidome of Triple Negative Breast Cancer Cells by Enhancing and Diversifying Antigen Processing and Presentation. *Front Immunol*. 2021;**12**:645770. doi: 10.3389/fimmu.2021.645770
55. Weingarten-Gabbay S, Klaeger S, Sarkizova S, Pearlman LR, Chen D-Y, Gallagher KME, *et al*. Profiling SARS-CoV-2 HLA-I peptidome reveals T cell epitopes from out-of-frame ORFs. *Cell*. 2021;**184**(15):3962-3980.e17. doi: 10.1016/j.cell.2021.05.046
56. Guasp P, Lorente E, Martín-Esteban A, Barnea E, Romania P, Fruci D, *et a*l. Redundancy and Complementarity between ERAP1 and ERAP2 Revealed by their Effects on the Behcet's Disease-associated HLA-B*51 Peptidome. *Mol Cell Proteomics*. 2019;**18**(8):1491-510. doi: 10.1074/mcp.RA119.001515
57. Mifsud NA, Illing PT, Lai JW, Fettke H, Hensen L, Huang Z, *et al*. Carbamazepine Induces Focused T Cell Responses in Resolved Stevens-Johnson Syndrome and Toxic Epidermal Necrolysis Cases but Does Not Perturb the Immunopeptidome for T Cell Recognition. *Front Immunol*. 2021;**12**:653710. doi: 10.3389/fimmu.2021.653710
58. Yair-Sabag S, Tedeschi V, Vitulano C, Barnea E, Glaser F, Melamed Kadosh D, *et al*. The Peptide Repertoire of HLA-B27 may include Ligands with Lysine at P2 Anchor Position. *Proteomics*. 2018;**18**(9):e1700249. doi: 10.1002/pmic.201700249
59. Forlani G, Michaux J, Pak H, Huber F, Marie Joseph EL, Ramia E, *et al*. CIITA-Transduced Glioblastoma Cells Uncover a Rich Repertoire of Clinically Relevant Tumor-Associated HLA-II Antigens. *Mol Cell Proteomics*. 2021;**20**:100032. doi: 10.1074/mcp.RA120.002201