TripHLApan: predicting HLA molecules binding peptides based on triple coding matrix and transfer learning

Meng Wang, Chuqi Lei, Jianxin Wang, Yaohang Li and Min Li
Corresponding author: Min Li, School of Computer Science and engineering, Central South University, Changsha 410083, China. Tel.: +86-731-88879335;
E-mail: limin@mail.csu.edu.cn

Abstract
Human leukocyte antigen (HLA) recognizes foreign threats and triggers immune responses by presenting peptides to T cells. Computationally modeling the binding patterns between peptide and HLA is very important for the development of tumor vaccines. However, it is still a big challenge to accurately predict HLA molecules binding peptides. In this paper, we develop a new model TripHLApan for predicting HLA molecules binding peptides by integrating triple coding matrix, BiGRU + Attention models, and transfer learning strategy. We have found the main interaction site regions between HLA molecules and peptides, as well as the correlation between HLA encoding and binding motifs. Based on the discovery, we make the preprocessing and coding closer to the natural biological process. Besides, due to the input being based on multiple types of features and the attention module focused on the BiGRU hidden layer, TripHLApan has learned more sequence level binding information. The application of transfer learning strategies ensures the accuracy of prediction results under special lengths (peptides in length 8) and model scalability with the data explosion. Compared with the current optimal models, TripHLApan exhibits strong predictive performance in various prediction environments with different positive and negative sample ratios. In addition, we validate the superiority and scalability of TripHLApan’s predictive performance using additional latest data sets, ablation experiments and binding reconstitution ability in the samples of a melanoma patient. The results show that TripHLApan is a powerful tool for predicting the binding of HLA-I and HLA-II molecular peptides for the synthesis of tumor vaccines. TripHLApan is publicly available at https://github.com/CSUBioGroup/TripHLApan.git.

Keywords: tumor vaccine; HLA; peptide; pan-specific prediction model

INTRODUCTION
Human leukocyte antigen (HLA) is the most polymorphic gene group in the human population, with more than 16 200 different types [1, 2]. These groups of genes help detect and identify foreign threats and trigger immune responses. Immunotherapy has emerged as a powerful cancer treatment in recent years, accurately predicting whether a peptide can be presented by a particular HLA molecule or not has become a critical problem in clinical immunotherapy [3–5]. HLA molecules are mainly divided into two categories: class-I and class-II. HLA class-I molecules mainly present intracellular endogenous peptides to the surface of CD8+ T cells, which are mostly derived from the degradation of intracellular proteins. HLA class-II molecules present exogenous peptides mainly through the phagocytosis of some phagocytes [6].

In the past 20 years, some computational tools have been developed to predict HLA molecules binding peptides [7, 8]. For the prediction task of HLA-I-peptide binding, early prediction tools, such as SYFPEITHI [9] and SMMPMBC [10], mainly use probabilistic methods to calculate position specificity to establish models. In recent years, more forecasting tools [11, 12] adopt machine learning, particularly deep learning methods, to build models. For example, netMHCSpanab [13] combines the immunogenicity characteristics of HLA-I-peptide complex to establish the artificial neural network model. ACME [14] uses the combination of...

Meng Wang received the B.S. and M.S. degree in safety engineering from Central South University, Changsha, China, in 2017 and 2020, respectively. Currently, He is working toward the Ph.D. degree in computer science and technology at Central South University, Changsha, China. His current research interests include bioinformatics and protein–peptide interactions.

Chuqi Lei received the B.S. degree in computer science from Central South University, Changsha, China, in 2020. Currently, she is working toward the Ph.D. degree in computer science and technology at Central South University, Changsha, China. Her current research interests include bioinformatics and protein–compound interactions.

Jianxin Wang received the B.S., M.S. and Ph.D. degrees in computer science from Central South University, Changsha, China, in 2001, 2004 and 2008, respectively. He is currently a Professor at the School of Computer Science and Engineering, Central South University. His main research interests include bioinformatics and deep learning.

Yaohang Li received the B.S. degree in Computer Science and Engineering from the South China University of Technology in 1997, and the M.S. and Ph.D. degrees in Computer Science from the Florida State University in 2000 and 2003, respectively. He is a Professor in the Department of Computer Science at Old Dominion University. His research interests are in Computational Biology, Monte Carlo methods, Big Data Analysis and Parallel/Distributed/Grid Computing.

Min Li received the B.S. degree in communication engineering and the M.S. and Ph.D. degrees in computer science from Central South University, Changsha, China, in 2001, 2004 and 2008, respectively. She is currently a Professor at the School of Computer Science and Engineering, Central South University. Her main research interests include bioinformatics and system biology.

Accepted: March 13, 2024
convolutional neural network and Attention modules to establish an interpretable model for affinity prediction. MHCSeqNet [15] applies Embedding [16] and skip-gram [17] models in Natural Language Processing to HLA-I-peptide prediction. TransPHLA is a generalized pan-specific model that is not restricted by HLA alleles or peptide length based on the transformer model [18]. Anthem builds an aggregating one-dependence estimators model trained by the integration of the outcomes of each scoring function from the set. And it facilitates model training using user input data with potential improvements in the performance [19]. Some tools model using only patient whole-exome sequencing and RNA sequencing data from their own laboratories to avoid the impact of data pollution on the model [20].

Similar to the development of HLA-I tools, the prediction tools of HLA-II molecules are gradually evolving from classical machine learning to deep learning. However, there are few tools for predicting HLA-II molecules binding to peptides, and their prediction accuracy is not high. At present, some successful tools, including NN-align [21], NetMHCIIpan [22], PUFFIN [23, 24], MARI2 [25], MHC.AttnNet [26], etc., only achieve good predictions on certain specific alleles [24, 27]. There are also some tools that can simultaneously predict the binding of two classes of HLA molecule with peptides, such as MHC.AttnNet [26] and MHCnuggets [28].

Early tools are allele-specific mainly. The allele-specific tools train a unique predictive model for each HLA molecular allele subtype. For example, NetMHC 4.0 [29] uses a feed forward neural network to classify HLA-I molecules. MHCnuggets [28] firstly clusters alleles, trains a model for the allele with the most binding peptides and then conducts transfer learning in clusters. Allele-specific tools often provide strong predictive performance, depending on the available amount of training data per allele and the quality of allele clustering [30, 31]. However, these tools have difficulty to generalize over all alleles, especially those that have not yet been discovered. Pan-specific approach [11, 12, 26, 32–37] mixes all HLA data together to train a pan-specific model that can predict all alleles with known sequences, such as MixMHCpred [38] and the latest version of NetMHCpan 4.1 [22]. As more and more HLA alleles are discovered, pan-specific tools play an increasingly important role in predicting HLA molecules binding peptides.

Although computational methods have been developed and made some success in the past few years, they have some challenges that limit the application. The main challenges are as follows: (1) Insufficient predicted allele types. Especially for HLA-II–peptides binding prediction task, only limited types of alleles can be predicted. (2) Extremely sensitive to the training data volume. For certain peptide lengths, the prediction performance of the existing prediction tools with a lot of available samples is acceptable. However, the predictor performance decreases sharply once the training samples for longer peptide lengths become insufficient. This phenomenon occurs on both allele-specific and peptide length–specific data sets [39]. (3) Lower predictive accuracy. Although most tools perform well on the data sets with large amounts of peptide lengths of 9 and 10, their predictive performance significantly decreases when peptide lengths are 8 and over 11 [39]. This may be due to the rare utilization of the biological properties and the connection of data known in the field. A part of characteristics may be masked during the data processing. In addition, most tools ignore the contextual information between amino acids.

To address the above issues, we propose TripHLApan, a new pan-specific model, for predicting HLA-peptide binding. Firstly, we extract the interaction site regions for HLA molecules binding peptides by analyzing 3D structure information in Uniprot database [40] data. Moreover, we compare the correlation between HLA molecular binding motifs and several HLA molecular coding strategy, and obtain the best allele molecular sequence extraction scheme. Then, we construct comprehensive feature profiles of peptides and HLA molecules, including the physicochemical and biochemical properties of amino acids, the probability of substitutions between amino acids and the endogenous hidden information related to binding. Finally, we design a new triple channel BiGRU + Attention [42] to model the potential information of peptide and HLA molecular sequence. The triple channel encoding input enables the model to fully learn different sequence information and synthesize their contributions to the combination. The BiGRU model captures the relationships between contexts. And the Attention model that focuses on the BiGRU hidden layer is used to reorganize the model weight of information from different training stages of BiGRU. The application of transfer learning ensures the independent training of the special data (peptides in length 8) and the scalability of the model in the future. Performance comparisons on several test sets show that TripHLApan outperforms the current best model.

**METHODS**

**Overview of TripHLApan**

TripHLApan is the first HLA–peptide binding prediction tool to encode and train in parallel using multiple amino acid characteristics (Figure 1a). In this model, the peptide sequence and HLA sequence are preprocessed and encoded. Three different encoding matrices are used to represent the sequences. Then, the encoded sequences are fed into the BiGRU + Attention model, and the outputs are concatenated together to form a matrix. After that, the output matrix is passed into three fully connected layers and one sigmoid layer, where the final binding probability is computed. We carry out a 5-fold cross-validation to avoid the contingency of model training data partitioning.

One can find that the TripHLApan’s BiGRU + Attention model is well adapted to the binding problem between HLA molecules and peptides. Firstly, the combination of peptides and proteins is largely determined by the complementary nature of their 3D structures. In the case of 3D structure information missing, BiGRU model based on sequence context information can well capture such sequence-based global information. Secondly, after being exported from BiGRU, TripHLApan does not directly take the last layer of the output matrix as the feature of the whole sequence like most models. Instead, it first uses the Attention model to redistribute the weights according to the importance of the subsequences learned by BiGRU model at the positions of various amino acid residues. Then it takes the last hidden layer as features of the whole sequence and inputs them into the next fully connected layer. The advantage of BiGRU + Attention method is that it is able to identify the influence of the subsequences at both ends of the peptide sequence, determining whether the whole peptide can bind or not. This is consistent with the phenomenon that the two ends of a peptide often show peptide-binding motifs [39].

The previous studies [43–45] suggest that the binding process of peptides with length 8 to HLA-I may involve the change of HLA-I molecular structure. Therefore, in the process of HLA-I model training, we first use the data sets of peptides in length 9–14 for model training, and then transfer the model to peptides in length 8 (Figure 1b). This training strategy ensures that the binding model of peptide with lengths 9–14 is not influenced by the data of peptide with length 8, and also ensures that the
prediction model of peptide with length 8 retains the learned HLA-I binding characteristics of peptides with lengths 9–14.

In order to better simulate the binding relationship between HLA molecules and peptides, we have analyzed the location characteristics of HLA binding peptides on different locus by Uniprot database. As shown in Figure 1c, protein 1AKJ is the binding structure of HLA-A and peptide. The peptide is completely wrapped into two helical structures and one β structure of HLA-A. Locating the key positions of several structures in the sequence index, we find that the positions of HLA-A binding peptide are all in the first 200 amino acids of the sequence. Several other class-I HLA molecules also show similar rules. Therefore, the first 200 amino acid fragments of HLA-I are selected to represent the whole HLA-I molecule and input into the model. For HLA-II, different scenes are observed. In Figure 1c, 1JK8 and 3LQZ, respectively, represent the binding relationship between HLA-DQA/DQB and HLA-DP1A/DP1B with peptides. The two chains of HLA molecule play a role together in the binding relationship with peptide. We specifically extract two 100 mer subsequences of different locus according to their different peptide binding sites to represent the two chains of HLA-II molecules, and feed them into the model, respectively.

**Data sets**

The data sets of HLA-I-peptide used for training and testing in this paper are from the IEDB database (downloaded on 15 March 2021) [46, 47] and several publications [10, 24, 28, 48, 49]. The data sets of HLA-II-peptide are from the IEDB database (downloaded on 15 March 2021). Only the items with mass spectrometry eluted ligands (in short, MS EL) label are selected for model training. Compared with affinity label, MS EL label covers antigen processing and presentation steps, and its experimental conditions
are more standardized. The HLA allele sequences are typically collected from the Immuno Polymorphism Database [2].

The purification process of data sets is described in Supplementary. Finally, for HLA-I, the number of samples in the final training set is 2788 602 (positive sample number: 464 767); the number of samples in the test set is 21 246 (positive sample number: 3941) and the number of samples in the unseen set (referring to the data set of HLA-I molecules not occurred in the training set, used to measure the generalization ability of the model) is 66 096 (positive sample number: 11 016). For HLA-II, the number of samples are 963 186 (positive: 160 531), 192 846 (positive: 32 141) and 5262 (positive: 877), respectively.

Preprocessing and data encoding
In preprocessing HLA-I molecular sequence, most of the previous prediction tools use the extraction of fixed position residues as the form of pseudo-sequences. But pseudo-sequences cannot capture the contextual relationship between adjacent amino acids within the sequence, which may lose important biological information. In order to find out the effect of encoding methods on the results, we investigate the relationship between sequence similarity and motif similarity of binding peptides under several different encoding schemes. We select 45 alleles with positive samples of more than 100 peptides and calculate the similarity between their sequences and the similarity of peptide motifs. The results are illustrated in Figure 2. There are 11 allele families with similar binding peptide motifs. Pair similarity based on full sequence of alleles can capture the similarity of 10 motif groups except the first one, while the other two pseudo sequences [45, 49] can capture only seven motif groups. This indicates that these two pseudo sequences can only capture part of the key information of alleles binding to the peptides, but they have difficulty in extracting more comprehensive information. Allele representation based on the full sequence is a more stable and comprehensive way for sequence information extraction. The allelic similarity based on the full sequence is poorly matched with the binding motifs of the first group, which may be due to the low similarity of the parts of the sequence unrelated to the binding relationship in this group of alleles. This suggests that extracting the useful sequence information embedded in the full sequence is useful for modeling their binding process. More details in the model can be found in the Supplementary.

RESULTS
TripHLApan outperforms baseline methods on HLA-I binding prediction
We compare TripHLApan with several carefully selected tools for classification performance, including the highly cited PickPocket [50], netMHCstabpan [13] and MixMHCpred [38], which are considered the best predictive tool in a recent survey [51], and five recent pan-specific tools: MHCSeqNet [15], MHCflurry 2.0 [52, 53], NetMHCpan 4.1 [22], MATHLA [35], MHCnuggets [28]. PickPocket mainly calculates the combination score through similarity calculation. NetMHCstabman and MHCflurry 2.0, respectively, integrate or calculate additional binding information (netMHCstabman: immunogenicity characteristics; MHCflurry 2.0: MHC allele-dependent effects (BA prediction) and allele-independent effects (AP prediction)). MHCSeqNet and MATHLA use models that perform well in other prediction tasks (MHCSeqNet: Embedding and skip-gram models; MATHLA: LSTM and multi-head attention mechanism). MHCnuggets adopts a strategy of first performing allele clustering, then transferring learning. NetMHCpan4.1 establishes a deep learning model for peptide binding motifs. MixMHCpred proposes a deconvolution algorithm based on position weight matrix, incorporating multiple specific features such as peptide length distribution. The selected comparison tool set covers a variety of different types and is representative in this field.

In the independent test set, we further define the test and unseen sets here. Their meanings are as follows: The allele types in the test set occur in the training set, and the allele types in the unseen set do not occur in the training set. Figure 3a shows the number of overlapping allele types between the unseen sets and these tools’ training data set. Even this testing is slightly unfair to TripHLApan on the unseen set since the benchmarking tool’s training process contains more items from our unseen sets. We use several indicators unrelated to threshold setting to evaluate the results of the tools: the area under the receiver operating...
Figure 3. TripHLApan is compared with the prediction results of baseline tools. A. The number of overlapping HLA types of the unseen set used in this paper and the training sets of nine tools including TripHLApan. B/C/D. AUCs/AUPRs and top-PPVs (the fraction of positive peptides within the top N) on the data sets with peptides of different lengths on the rate of positive and negative samples at 1:5/1:1/1:10/1:50.

characteristic curve (AUC), area under curve of PR (AUPR) and the fraction of positive predictive value within the top N (top-PPV).

The results in Table 1 show that TripHLApan is a powerful predictive tool on both test and unseen sets. TripHLApan has not only strong predictive power but also strong generalization ability. To observe the predictive differences between TripHLApan and the baseline tools at a finer granularity, we further divide these two test sets according to the lengths of different peptides.
Table 1: AUCs and AUPRs of TripHLApan compared with baseline tools on test set and unseen set. The results of TripHLApan are the averages of the 5-fold model.

| Methods       | Test set  | Unseen set |
|---------------|-----------|------------|
|               | AUC       | AUPR       | AUC       | AUPR       |
| TripHLApan    | 0.978     | 0.933      | 0.958     | 0.881      |
| netMHCPan     | 0.905     | 0.839      | 0.949     | 0.883      |
| MHCSeqNet     | 0.815     | 0.354      | 0.853     | 0.47       |
| MixMHCpred    | 0.896     | 0.779      | 0.941     | 0.843      |
| MHCFlurry     | 0.933     | 0.871      | 0.938     | 0.861      |
| netMHCstabpan | 0.888     | 0.74       | 0.922     | 0.78       |
| PickPocket    | 0.877     | 0.716      | 0.923     | 0.77       |
| MATHLA        | 0.891     | 0.704      | 0.92      | 0.741      |
| MHChnuggets   | 0.854     | 0.544      | 0.759     | 0.534      |

Figure 3b shows that in the test set, TripHLApan outperforms all other tools in predicting peptides of different lengths on the rate of positive and negative samples at 1:5/1:1/1:10/1:50. Moreover, TripHLApan’s performance degrades much slower than the other methods for longer peptides. In particular, on the rate of positive and negative samples at 1:5, TripHLApan’s AUC value outperforms the second-ranked models by 1.2%, 6.1%, 18.8%, 10.3% and 24.9% for peptide lengths of 10, 11, 12, 13 and 14 in the independent test set, respectively. The unseen set, whose goal is to measure the model’s generalization ability, has alleles not seen in the TripHLApan’s training set. The results show that when the peptide length is over 12, TripHLApan exhibits significantly higher AUC values than the other tools. Furthermore, TripHLApan has the highest AUPR values (see Table 1 and Figure 3c), indicating that TripHLApan is able to handle different data set environments and can actually achieve good prediction results. Figure 3c and d shows the AUPRs and top-FPPs of selected tools with positive and negative samples at 1:5/1:1/1:10/1:50 (More detailed results can be found in Supplementary Figures S1 and S2). One can find that the prediction accuracy of TripHLApan is comparable with NetMHCPan, MixMHCpred and MHCFlurry on the unseen set, and it shows better prediction ability in other test sets. This suggests that TripHLApan can effectively learn the patterns between sequences and binding patterns and thus pick out the most likely combinations of potential compounds, even when the sample ratios are much more stringent, such as 1:10 and 1:50. Then, TripHLApan’s strong learning ability on the new alleles is verified. We randomly select four-fifths portion of the unseen set to perform transfer learning using TripHLApan, and the remaining one-fifth portion is used for testing. The results are shown in Figure 3b ‘unseen set (left)’ where the prediction from TripHLApan is improved after transfer learning. This indicates that TripHLApan can adapt easily to a variety of prediction environments. Even on the allele conjugate that has never been seen before, TripHLApan can take advantage of known data for transfer learning to enhance its prediction capability. To eliminate the unseen set’s unfairness to TripHLApan and the benchmark tools, we extract the alleles that do not appear in all tool’s training sets (named ‘all unseen set’), a total of 19 160 items on 13 HLA alleles. TripHLApan obtains the highest AUC/AUPR and top-FPP scores (see Figure 4a–d). Therefore, we believe that, with the increase of HLA molecular-peptide data, TripHLApan can also achieve good prediction results after simple transfer learning on new allele data set.

Testing on the latest data set

We have tested TripHLApan on the recent data set from the IEDB database to further verify its validity. Firstly, we extract the recent data set from the IEDB (downloaded on 7 October 2021), and de-replicate it with the train set, test set and unseen set used in this paper to ensure that the new test data are completely new to the model. Then, the new data set is purified in the same way described in this paper. Finally, a total of 68 277 data samples are obtained, including 14 705 positive samples. Figure 4e–g shows that TripHLApan still achieves the best prediction performance in the recent test set compared with the three models that perform better in the test set. In addition, we divide the recent data set into different subsets based on different peptide lengths to evaluate the performance differences of the tool on different peptide lengths. After excluding data sets with sample numbers less than 50, the prediction results of different subsets are shown in Figure 4h. TripHLApan is also a leading method in predicting peptide groups of different lengths. In conclusion, we compare the prediction performance of different tools based on the latest epitope data collected by IEDB, and the results show that TripHLApan has a more stable prediction performance under different prediction environments.

Proper model architecture and data preprocessing contribute to better model performance

We attribute TripHLApan’s strong predictive power to its appropriate model combination and correct strategy selection. Firstly, based on the previous biomolecules structural analysis and statistical analysis of data (Figure 1c, Figure 2), three parallel sequence encoding methods are used to extract the information contained in the sequence from the perspectives of biochemical properties of amino acids, substitution probability and inherent hidden information related to binding. It is important to note that the biochemical properties of amino acids have been neglected by most methods, but these properties turn out to be crucial in predicting the combination of peptides and HLA molecules. Second, BiLSTM and BiGRU models, which can learn the relationship between sequence context information, are selected not only because they are as effective as described, but also because they do show strong predictive power in previous studies [15]. According to our preliminary comparative experiments, BiGRU is selected as our basic model unit. Thirdly, we add Attention modules to TripHLApan to pay attention to the weight of each subsequence of peptides and HLA molecules. We choose this kind of Attention approach because of the observed phenomenon of ‘significant anchor residues at both ends of the peptide’ [38]. We use the Attention module to focus on the contributions between the hidden layers of GRU output. The advantage is that we can ’recall' the outputs learned at different 'moments', thereby achieving the goal of learning more comprehensive knowledge. We conduct a fine tune on the number of hidden layers that the attention module focuses on, and the results show that the 256 attention layers we have chosen are the most effective while ensuring the running speed, as shown in Figure S3.

Selecting such modules allows TripHLApan to better capture the complex correlations of the anchor sites and their adjacent sites, and maximize the exposure of anchor residues at both ends of the sequence. Finally, the transfer learning strategy reduces the performance loss caused by excessive mixing of data and retains the learned binding modes. In addition, the ‘dropout’ and ‘early termination’ strategies also play an important role in preventing overfitting and improving model stability.

Table 2 shows the results of some ablation models compared with the original model. One can find that TripHLApan is the model with the strongest comprehensive prediction ability.
However, TripHLApan (AAIndex), which is encoded using only the biochemical properties of amino acids, also achieves very similar prediction results, indicating the biochemical properties of amino acids are the most crucial features. The training strategy of transfer learning also helps improve the performance of TripHLApan on the data sets with peptides of different lengths. Figure 5 compares TripHLApan in detail with the model without transfer learning strategy [TripHLApan (mixed)], the model with ab initio training for different peptide lengths [TripHLApan (from scratch)] and the well-performing TripHLApan [AAIndex]. The training strategy of transfer learning limits the prediction ability of peptide length 8 to some extent, but it is extended for longer peptide length (11/12/13/14). By comparing the results in Table 1 and Figure 5 one can find that TripHLApan’s prediction performance at different lengths is occasionally worse than that of some reference tools, but its overall AUC is much better than the other tools. This suggests that TripHLApan’s ability to separate positive and negative samples remains stable across data sets with peptides of different lengths. This is an important ability for predictive tools, especially when most current tools use defined thresholds for dividing positive and negative samples. In general, TripHLApan collects more comprehensive sequence information and amino acid characteristics, and adopts biological prior knowledge. As a result, it greatly improves the accuracy of binding prediction of HLA-1 molecules with peptides.

**Figure 4.** TripHLApan is compared with the prediction results of baseline tools on the data set in which the alleles do not appear in all tool’s training sets and on the last data set. A. Allele types in the unseen sets. B–D. ROC/AUPR curves and top-PPVs on all unseen sets. E–G. ROC/AUPR curves and top-PPVs on test set of TripHLApan compared with the prediction results of baseline tools. H. AUCs on peptides of different lengths on test set of TripHLApan compared with the prediction results of baseline tools.

**Table 2:** AUCs of model ablation experiment. All the results are the averages of the 5-fold model

| Models   | TripHLApan | 1# | 2# | 3# | 4# | 5# | 6# | 7# |
|----------|------------|----|----|----|----|----|----|----|
| Test set | 0.981      | 0.98 | 0.985 | 0.978 | 0.978 | 0.978 | 0.979 | 0.979 |
| Unseen set | 0.942 | 0.938 | 0.933 | 0.937 | 0.935 | 0.944 | 0.922 | 0.916 |

Note: These abbreviations represent different versions of TripHLApan: 1#: without Self-Attention; 2#: AAIndex + Blosum62; 3#: AAIndex + Embedding; 4#: Blosum62 + Embedding; 5#: AAIndex; 6#: Embedding; 7#: Blosum62.

**TripHLApan can predict single-patient cancer immunopeptidome**

To test the clinical application of TripHLApan, we collect individual patient melanoma-related immunopeptidome data sets [54–56]. In this data set, the true binding relationship between HLA-I molecule and the peptide is unknown. We use TripHLApan to reconstruct the HLA-I molecule and peptide correspondence in various cell lines. TripHLApan’s ability to re-annotate peptide presentation molecules is tested by calculating the Pearson correlation coefficients (PCCs) between the frequencies of amino acids in the predicted positive sample at each location and the frequencies of amino acids in the real binding. The comparison tools NetMHCpan and MixMHCpred are excluded because they have already included these four cases in their training set. We also exclude PickPocket, MATHLA and MHCnuggets here, due to their poor performance on the top-PPV according to Figure 3d. Figure 6 shows the average PCCs of tools on the data sets with peptide lengths at 9 and 10. TripHLApan has demonstrated a powerful ability to reconstruct peptide-HLA-I molecular relationships.

TripHLApan’s deconvolution ability is more stable in various cell lines compared with other methods. Especially in Mel5 and Mel8, when the peptide length increases from 9 to 10, the relationship reconstruction ability of all other tools decreased significantly, while TripHLApan remains stable. This not only demonstrates TripHLApan’s ability to reconstruct relationships with peptide samples of length 9 (the most common peptide length),
Figure 5. Ablation experiment. ROC curves and AUCs on peptides of different lengths on the test set of TripHLApan and its three different model strategies, as well as the three baseline models that performed better.

Figure 6. Comparison of the average PCCs of tools on the four monoallelic samples. but also shows its stability over longer peptide samples. For more detailed results, see Supplementary Tables S1 and S2.

TripHLApan outperforms baseline methods on HLA-II binding prediction

For the task of predicting the binding of HLA-II molecules to peptides, the current tools have the following limitations: (1) the molecular types that can be predicted are limited, and (2) the prediction accuracy is generally low. Based on the above analysis of the only 3D structure data of HLA-II molecules and the outstanding performance of TripHLApan in HLA-I, we find that TripHLApan also has strong predictive ability in predicting the binding of HLA-II molecules and peptides. Similar to predicting Type I, data in the independent test set are divided into the test sets and unseen sets according to whether their allele type appears in the training set or not. Finally, 77 test sets and 63 unseen sets are divided according to alleles. TripHLApan has AUCs greater than 0.9 on 61 and 55 subsets of the test set and unseen set, respectively. In particular, TripHLApan’s results on the 30 alleles in the unseen sets are above 0.99, an important breakthrough in the development of predictive models for HLA-II molecular peptide binding. We compare the results with nine tools [21, 48, 57-59] integrated into the IEDB database. Excluding DP and DQ, which cannot be predicted by the baseline tools and other unsupported alleles, there are 35 subsets in the test set and eight subsets in the unseen set. TripHLApan is able to achieve optimum AUC values on their 23 and 8 alleles (Supplementary Tables S3 and S4). Figure 7 shows the AUC distribution comparison results on the allele sets that most of the tools can predict, where TripHLApan (AAIndex) denotes the result of a version of TripHLApan only using AAIndex features. Next, we compare the performance of MHcnuggets in class II that can also do the binding prediction task of HLA-I and II peptides. The results show that TripHLApan can achieve better prediction accuracy on 31 alleles in 36 alleles collections supported by MHcnuggets. More detailed results are also shown...
in Supplementary Tables S3 and S4. Overall, TripHLApan, as a pan-specific model that can predict more allelic types, has achieved a prediction accuracy not reachable before.

CONCLUSIONS

In this work, we propose TripHLApan, a new predictive model of HLA binding with peptides integrating multiple features. We first obtain the best allele molecular sequence extraction scheme by extracting the interaction site regions and comparing the correlation between HLA molecular binding motifs and several HLA molecular coding strategy. This is an important step that other tools overlook before building their models. Then, we integrate the biochemical characteristics of amino acids rarely used in other tools, and form a parallel sequence coding and training network with the probability of amino acid replacement and Embedding features. Finally, we use Attention module to learn the context information contained in each local sequence after BiGRU, avoiding the loss of sequence information caused by excessive attention to position-specific factors in the modeling process. TripHLApan is confirmed to be superior to the current state-of-the-art prediction tools by a 5-fold cross-validation.

On the task of HLA-I-peptide binding prediction, we also test the generalization ability of the model with allele binding sample sets that do not appear in the TripHLApan training set (note that this is unfair to TripHLApan, as these samples may have been used to train for baseline tools). TripHLApan’s prediction results are still at the forefront of comparison tools. After a simple transfer learning in the unseen set, TripHLApan’s results remain stable, demonstrating TripHLApan’s powerful ability to learn and adapt to its environment. TripHLApan achieves the highest AUC score of 0.979 on the 13 HLA alleles that do not appear in all tools’ training sets. Also, TripHLApan’s powerful predictive performance has been validated on the latest data set. Nevertheless, we conduct a series of model ablation experiments, which confirm the effectiveness of the framework and strategy adopted in TripHLApan. Finally, we validate the stability of TripHLApan’s predictive power in clinical use on a single patient melanoma-related immunopeptidome data set. In the HLA-II-peptide binding prediction task, we have achieved more significant improvement in prediction ability, which is mainly reflected in the following two aspects: (1) TripHLApan is able to predict more allele types. TripHLApan is a pan-specific prediction model that can be used to predict HLA molecules and peptides as long as their sequences are known. (2) TripHLApan has demonstrated effectiveness in HLA-II-peptide binding prediction task.

Based on the ablation experiment and the architecture analysis of the comparative model, we attribute the excellent performance of TripHLApan mainly to the following aspects: (1) the data analysis and coding strategy in the early stage are more suitable for the binding patterns of HLA molecules and peptides. The multi-encoding parallel model input enables the model to capture as much binding information as possible. (2) The architecture of BiGRU+Attention not only enables the model to learn the binding relationships between many crucial subsequences, but also refocuses the importance of each subsequence. However, there are some limitations to TripHLApan. The enhancement is not quite observable on predicting the samples with the most common length of peptides at 9 on HLA-I-peptide binding prediction. This is mainly due to the fact that the predictive power on peptide length at 9 of existing tools has already reached a highly accurate point. In this case, the predictive power of the model will be more affected by the quality of the data. On the other hand, the characteristic of the TripHLApan’s pan-specificity limits its performance for 9 lengths to some extent. In addition, the sequence-based dichotomous prediction models cannot determine the specific binding site and posture, and we will take 3D features into account in future studies, especially in this era when protein structure prediction has achieved fine-grained accuracy. With the further increase of three-dimensional integrated data, combining the structural information of the data and integrating this information into the model will become a focus and direction of future research.

**Key Points**
- We develop a new model TripHLApan for predicting HLA molecules binding peptides by integrating triple coding matrix, BiGRU + Attention models, and transfer learning strategy.
- We obtain the best allele molecular sequence extraction scheme by extracting the interaction site regions and comparing the correlation between HLA molecular binding motifs and several HLA molecular coding strategy.
- TripHLApan exhibits strong predictive performance in various prediction environments with different positive and negative sample ratios. In addition, we validate the superiority and scalability of TripHLApan’s predictive performance using additional latest data sets, ablation experiments and binding reconstitution ability in the samples of a melanoma patient. The results show that TripHLApan is a powerful tool for predicting the binding of HLA-I and HLA-II molecular peptides for the synthesis of tumor vaccines.

**SUPPLEMENTARY DATA**

Supplementary data are available online at https://academic.oup.com/bib.

**FUNDING**

This work is supported by the National Natural Science Foundation of China under Grant No. 61832019, the science and technology innovation program of Hunan Province (2021RC0048) and the Hunan Provincial Science and Technology Program (2019CB1007). This work is supported in part by the High Performance Computing Center of Central South University.

**DATA AVAILABILITY**

The dataset used for TripHLApan training and testing is from the IEDB database (https://www.iedb.org/). Single Patient Cancer Immunopeptidome data is from publications [54–56].

**REFERENCES**

1. Lefranc M-P, Giudicelli V, Douroux P, et al. IMGT®, the international ImMunoGeneTics information system®, 25 years on. Nucl Acids 2015;43(D1):D413–22.
2. Robinson J, Halliwell JA, Hayhurst JD, et al. The IPD and IMGT/HLA database: allele variant databases. Nucleic Acids Res 2015;43(D1):D423–31.
3. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. CA Cancer J Clin 2019;69(1):7–34.
et B, et Q, Chu Wang et G, Mikolov T, L, et Collobert J, Cui Liu Z, et S, Srihari Chandragandhi

11. Finck A, Gill SI, June CH. Cancer immunotherapy comes of age and looks for maturity. Nat Commun 2020;11(1):1–4.

12. Rubin MM, Schreiber RD. The odds of immunotherapy success. Science 2015;350(6257):158–9.

13. Neefjes J, Jongsmma ML, Paul P, Bakke O. Towards a systems understanding of MHC class I and MHC class II antigen presentation. Nat Rev Immunol 2011;11(12):823–36.

14. Yuvaraj N, Srihari K, Chandragandhi S, et al. Analysis of protein-ligand interactions of SARS-Cov-2 against selective drug using deep neural networks. Big Data Min Anal 2021;4(2):76–83.

15. Qi R, Guo F, Zou Q. String kernels construction and fusion: a survey with bioinformatics application. Front Comp Sci 2022;16(6):166904.

16. Rammensee H-G, Bachmann J, Emmerich NPN, et al. SYFPEITHI database for MHC ligands and peptide motifs. Immunogenetics 1999;50(3–4):213–9.

17. Kim Y, Sidney J, Pinilla C, et al. Derivation of an amino acid similarity matrix for peptide: MHC binding and its application as a Bayesian prior. BMC Bioinformatics 2009;10(1):1–11.

18. Wu J, Wang W, Zhang J, et al. DeepHLApan: a deep learning approach for neoantigen prediction considering both HLA-peptide binding and immunogenicity. Front Immunol 2019;10:2559.

19. Liu Z, Cui Y, Xiong Z, et al. DeepSeqPan, a novel deep convolutional neural network model for pan-specific class I HLA-peptide binding affinity prediction. Sci Rep 2019;9(1):1–10.

20. Rasmussen M, Fenyö H, Harndal M, et al. Pan-specific prediction of peptide–MHC class I complex stability, a correlate of T cell immunogenicity. J Immunol 2016;197(4):1517–24.

21. Hu Y, Wang Z, Hu H, et al. ACME: pan-specific peptide–MHC class I binding prediction through attention-based deep neural networks. Bioinformatics 2019;35(23):4946–54.

22. Pholphisut P, Pornputtapon N, Sriswadi S, Chuaungswanich E. MHCSeqNet: a deep neural network model for universal MHC binding prediction. BMC Bioinformatics 2019;20(1):1–10.

23. Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch. J Mach Learn Res 2011;12(APRICLE):493–537.

24. Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space. Preprint at https://arxiv.org/abs/1301.3781v3. 2013.

25. Chu Y, Zhang Y, Wang Q, 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):300–11.

26. Mei S, Li F, Xiang D, et al. Anthem: a user customised tool for fast and accurate prediction of binding between peptides and HLA class I molecules. Brief Bioinform 2021;22(5):1–16.

27. Gartner JJ, Parkhurst MR, Gross A, et al. A machine learning model for ranking candidate HLA class I neoantigens based on known neoepitopes from multiple human tumor types. Nature cancer 2021;2(9):563–74.

28. Jensen KK, Andreatta M, Marcatili P, et al. Improved methods for predicting peptide binding affinity to MHC class II molecules. Immunology 2018;154(3):394–406.

29. Reyennisson B, Alvarez B, Paul S, et al. 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 2020;48(W1):W449–54.

30. Dai Z, Huismann BD, Zeng H, et al. Machine learning optimization of peptides for presentation by class II MHCs. Bioinformatics 2021;37(19):3160–7.
44. Maenaka K, Maenaka T, Tomiyama H, et al. Nonstandard peptide binding revealed by crystal structures of HLA-B* 5101 complexed with HIV immunodominant epitopes. *J Immunol* 2000;165(6):3260–7.
45. Sarkizova S, Klaeger S, le PM, 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.
46. Vita R, Overton JA, Greenbaum JA, et al. The immune epitope database (IEDB) 3.0. *Nucleic Acids Res* 2015;43(D1):D405–12.
47. Klaera S, Sarkizova S, PM, et al. A combined analysis of antigen presentation and T-cell recognition reveals restricted immune responses in melanoma. *Cancer Discov* 2018;8(11):1366–75.
48. Wang P, Sidney J, Kim Y, et al. Peptide binding predictions for HLA DR, DP and DQ molecules. *BMC Bioinformatics* 2010;11(1):1–12.
49. Nielsen M, Lundegaard C, Blicher T, et al. NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. *PLoS One* 2007;2(8):e796.
50. 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):1293–9.
51. Mei S, Li F, Leier A, et al. A comprehensive review and performance evaluation of bioinformatics tools for HLA class I peptide-binding prediction. *Brief Bioinform* 2020;21(4):1119–35.
52. O’Donnell TJ, Rubinsteyn A, Bonsack M, et al. MHCflurry: open-source class I MHC binding affinity prediction. *Cell Syst* 2018;7(1):129–132.e4.
53. O’Donnell TJ, Rubinsteyn A, Laserson U. MHCflurry 2.0: improved pan-allele prediction of MHC class I-presented peptides by incorporating antigen processing. *Cell Syst* 2020;11(1):42–48.e7.
54. Kalaora S, Barnea E, Merhavi-Shoham E, et al. Use of HLA peptidomics and whole exome sequencing to identify human immunogenic neo-antigens. *Oncotarget* 2016;7(5):5110–7.
55. Bassani-Sternberg M, Bräunlein E, Klar R, et al. Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. *Nat Commun* 2016;7(1):1–16.
56. Sturniolo T, Bono E, Ding J, et al. Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. *Nat Biotechnol* 1999;17(6):555–61.
57. Sidney J, Assarsson E, Moore C, 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):2–14.