TransMut: a program to predict HLA-I peptide binding and optimize mutated peptides for vaccine design by the Transformer-derived self-attention model

Yanyi Chu (a96123155@sjtu.edu.cn)
Shanghai Jiao Tong University https://orcid.org/0000-0002-4969-3931
Yan Zhang
Qiankun Wang
Shanghai Jiao Tong University
Lingfeng Zhang
University of Ottawa
Xuhong Wang
Shanghai Jiao Tong University
Yanjing Wang
Shanghai Jiao Tong University
Jianmin Wang
Yonsei University
Xue Jiang
Dennis Salahub
University of Calgary
Yi Xiong
Shanghai Jiao Tong University https://orcid.org/0000-0003-2910-6725
Dong-qing Wei
Shanghai Jiao Tong University

Article

Keywords: Human leukocyte antigen, pHLA binding prediction, Transformer, self-attention, mutated peptides

Posted Date: September 30th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-785618/v1

Version of Record: A version of this preprint was published at Nature Machine Intelligence on March 23rd, 2022. See the published version at https://doi.org/10.1038/s42256-022-00459-7.

TransMut: a program to predict HLA-I peptide binding and optimize the mutated peptides for vaccine design by the Transformer-derived self-attention model

Yanyi Chu[1, 2†], Yan Zhang[3†], Qiankun Wang[1], Lingfeng Zhang[4], Xuhong Wang[5], Yanjing Wang[1], Jianmin Wang[6], Xue Jiang[1], Dennis Russell Salahub[2], Yi Xiong[1*], Dong-Qing Wei[1, 7*]

[†]These authors are co-first authors and contributed equally to this work.

[*]These authors are corresponding authors.

[1]State Key Laboratory of Microbial Metabolism, Shanghai-Islamabad-Belgrade Joint Innovation Center on Antibacterial Resistances, Joint International Research Laboratory of Metabolic & Developmental Sciences and School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200030, P.R. China.

[2]Department of Chemistry, CMS - Centre for Molecular Simulation and IQST - Institute for Quantum Science and Technology, University of Calgary, 2500 University Drive NW, Calgary, Alberta, Canada T2N 1N4.

[3]Department of Clinical Oncology, The University of Hong Kong-Shenzhen Hospital, Shenzhen, P.R. China;

[4]School of Electrical Engineering and Computer Science, University of Ottawa, 75 Laurier Ave., Ottawa, Ontario, Canada, K1N 6N5.

[5]The Ministry of Education Key Laboratory of System control and information

processing, Department of Automation, School of Electronic, Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200030, P.R. China.

[6]Integrative Biotechnology & Translational Medicine, Yonsei University, Incheon 21983, Republic of Korea.

[7]Peng Cheng Laboratory, Vanke Cloud City Phase I Building 8, Xili Street, Nanshan District, Shenzhen, Guangdong, 518055, P.R China.

Tel: +86 21-34204573; Email: xiongyi@sjtu.edu.cn, dqwei@sjtu.edu.cn

Abstract

Computational prediction of the interaction between human leukocyte antigen (HLA) and peptide (pHLA) can speed up epitope screening and vaccine design. Here, we develop the TransMut framework composed of TransPHLA for pHLA binding prediction and AOMP for mutated peptide optimization, which can be generalized to any binding and mutation task of biomolecules. Firstly, TransPHLA is developed by using a Transformer-derived self-attention model to predict pHLA binding, which is significantly superior to 11 previous methods on pHLA binding prediction, neoantigen and human papilloma virus vaccine identification. For vaccine design, the AOMP program is then developed to automatically optimize mutated peptides to search for mutant peptides with higher affinity to the target HLA and with high homology to the source peptide. Among 3660 non-binding pHLAs, 3630 were successfully mutated. Of these, 94% were verified by the IEDB recommended method, and 88% have homology higher than 80% to the source peptide.

Keywords: Human leukocyte antigen, pHLA binding prediction, Transformer, self-attention, mutated peptides.

Introduction

The binding of peptides with the major histocompatibility complex (MHC), also called human leukocyte antigen (HLA) in humans, is essential for antigen presentation, which is a necessary prerequisite for effective T cell recognition[1].

Only when the peptide is presented to the HLA molecules on the outer cell surface to form a peptide-HLA (pHLA) complex, and then recognized by the T cell, can it trigger a robust immune response[2]. HLA is generally divided into two categories: HLA class I (HLA-I) and HLA class II (HLA-II). HLA-I is encoded by three I loci, including HLA-A, HLA-B, and HLA-C, and it is expressed on the surface of all nucleated cells. In contrast, the encoded HLA-II can only be expressed in professional antigen-presenting cells[3]. Therefore, we focus on HLA-I molecules (hereinafter referred to as HLA) in this study. HLA mainly binds short peptides with a length of 8-12 amino acids, of which 9-mer peptides are the most common[4]. These short peptides are usually obtained by proteasome-mediated degradation of intracellular proteins. Then, presenting these pHLAs on the cell surface for $CD8^+$ T cells to recognize[5,6].

Since the HLA molecule is highly specific and has a polymorphism in the human population[7], coupled with the source antigen it is randomly processed by proteasome intracellularly, a small proportion of peptides can be presented to the HLA molecules[1]. Determining which peptides are selected for display in an individual's HLA type is crucial to epitope selection. The most reliable method is to verify the affinity of peptides and HLA through experiments, which are time- and labor-consuming. Given that the affinity of the peptide and HLA is closely related to whether it can be presented, many *in silico* methods have been developed to predict the affinity between the peptide and HLA (the related works are summarized in Section 1 of the Supplementary Information). The

existing methods are mainly based on machine learning models, especially neural networks, to predict the binding or affinity between peptides and some HLA alleles[8]. However, over the years of dataset accumulation and modeling improvement, although the accuracy is as high as 90% for peptides of length 9[9], other performance evaluation metrics and prediction capabilities for other peptide lengths are still not satisfactory[8]. The reason for the above situation is that, compared with other lengths, 9-mer peptides are easier to bind to HLA and have the most pHLA binding data. In contrast, the peptides with lengths 13 and 14 have few pHLA binding data. Moreover, many methods construct multiple allele-specific models[10], which make it impossible to be applied in HLA or peptide lengths that do not exist in the training data, and there is a tendency for poor performance for HLA or peptide lengths with few training data. Therefore, a great deal of effort is urgently needed to develop pan-specific methods[10], which are trained on multi-allele data and solve the problem of allele-specific methods, to accurately predict pHLA binding or not, especially for rare HLA and peptide lengths.

Moreover, it is attractive to synthesize short peptides to design highly targeted immune responses. Therefore, understanding the interaction of pHLA can promote the design of short peptide vaccines[11] and play an important role in the development of candidate vaccines for various diseases[12,13]. Several studies[14,15] demonstrate that neoantigens produced by non-synonymous mutations in cancer cells play a key role in the anti-tumor immune response.

Moreover, vaccines for neoantigens have been proven to be beneficial to clinical outcomes[16,17]. In addition, short peptide-based vaccines have many advantages over traditional vaccines[12,18], such as the lower risk of genetic recombination and integration, higher stability, and the easier to synthesize and store, etc. The principle of short peptide vaccines is that antigen peptides bind to specific HLA to form HLA-peptide-TCR complexes to elicit T cell immune responses. Theoretically, the antigen peptide should have high affinity with specific HLA. At present, the process of identifying neoantigens is as follows[8]. First, use a next-generation sequencing platform to characterize the non-synonymous mutations of the primary tumor, and then use docking or other computational methods to predict the binding probability of the mutant peptide and the patient's HLA[19]. Finally, the number of candidate mutant peptides has been reduced, thus speeding up the process of experimental validation[20,21]. However, the above-mentioned process is relatively complicated, requires high professional knowledge, and is difficult to be applied in large-scale development. Therefore, the development of an automatic program to optimize mutated peptides (AOMP) would represent a huge breakthrough in the neoantigen design field.

In this study, we designed a Transformer[22]-derived model for pHLA binding prediction (TransPHLA), as shown in Figure 1. It is a pan-specific method[10], which can achieve more stable and stronger performances and can be applied to rare and unseen HLA alleles. The core idea of the TransPHLA model is to

apply self-attention[22] to peptides, HLAs, and pHLA pairs to obtain the binding score. In addition, we also introduced some techniques to optimize the model: (i) Embedding block: Besides the encoding of amino acids in the sequence, we added positional embedding to describe the position information of the sequence; (ii) Encoder block: Multiple self-attentions are applied to focus on different components of sequences, and padding positions of the sequence will be masked to prevent misleading the model; (iii) Feature optimization block: The fully-connected layers with the gyro channel that rise first and then fall are used to process the features obtained by the previous self-attention block to obtain better feature representation; (iv) Projection block: Use multiple fully-connected layers to predict the final pHLA binding score. The above four sub-modules together constitute the whole model. We compared TransPHLA with 11 previous PHLA binding prediction methods, including the state-of-the-art method[23], IEDB recommended method[9], and 9 benchmark methods[9,24-31]. TransPHLA not only achieved better performance, but also solved the limitations of many methods in HLA allele and peptide length. Moreover, the speed of TransPHLA is very high, the prediction of 170000 samples only needs 2 minutes on a CPU, and our web server does not limit the number of user inputs. Further, we conduct two types of case studies to demonstrate the usability and validity of the proposed TransPHLA. The TransPHLA shows an extraordinary performance compared with 11 other methods of neoantigen identification[32,33], and it achieves a positive screening rate of 96%. For human

papilloma virus (HPV) vaccine identification[34], although the positive screening rate is not very high, it is also significantly superior to the other 11 methods.

In another crucial aspect, we developed an AOMP program based on the attention mechanism obtained by trained TransPHLA, which can be used in many aspects especially vaccine design. The details of AOMP are shown in Figure 2. When the user provides a pair of a source peptide and a target HLA, the AOMP program can search for mutant peptides with higher affinity for the target HLA and the mutation between 1-4 positions. This program not only guarantees the affinity between the mutant peptide and the target HLA, but also ensures the homology of the mutant peptide and the source peptide to trigger cross-immunization. We tested 366 combinations of different HLAs and peptide binder lengths, and randomly selected 10 non-binding peptides for each combination, for a total of 3660 negative pHLAs. Among them, 3630 successfully found the binding mutant peptide-HLA, and 94% of them were verified by the method recommended by IEDB[9], which confirmed the usability of our program. Further, 88% of the 3660 source peptides can find successful mutant peptides with a homology of more than 80% (mutate 1-2 sites), which is exciting for vaccine design.

The TransPHLA and AOMP program jointly form the TransMut framework, which is the first successful attempt to apply Transformer to the field of biomolecular mutations. This framework can be applied to any biomolecular mutation task, such as epitope optimization[35], drug design[36]. Especially in the

vaccine development, like TNFα-targeted vaccine, due to the biological activity of TNFα, it will cause inflammation in the body, and long-term medication will cause the risk of autoimmune diseases[37]. The core problem of TNFα vaccine development is how to reduce the biological activity of TNFα while maintaining sufficient immunogenicity[38]. Therefore, the AOMP program is just suitable to be applied to the project. First, use the Transformer-derived model to train the mutation direction data of the biomolecule, and then the attention score toward the mutation direction is obtained. Based on the attention score, the AOMP program will find a better mutant.

We provide a web server for TransPHLA and AOMP programs, the input and output results of which are shown in Figure 3. The data and code are freely available.

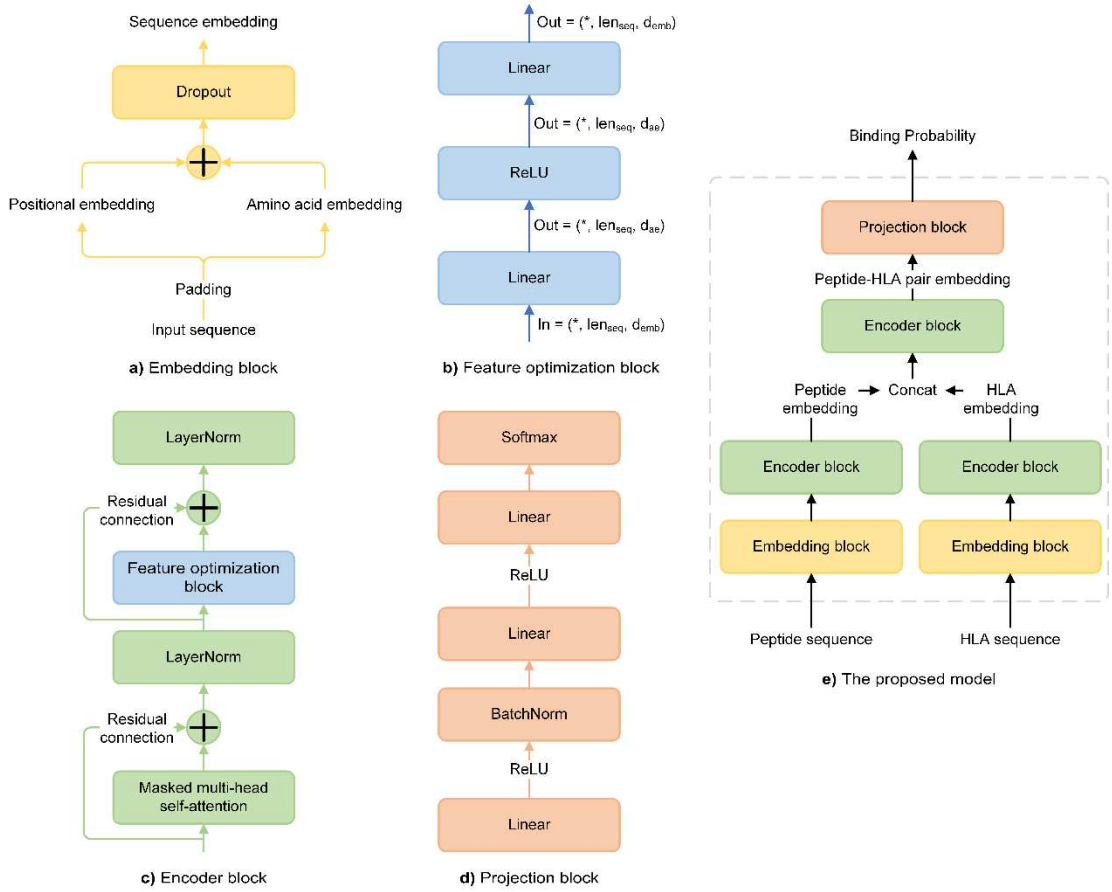


Figure 1. The proposed TransPHLA model.

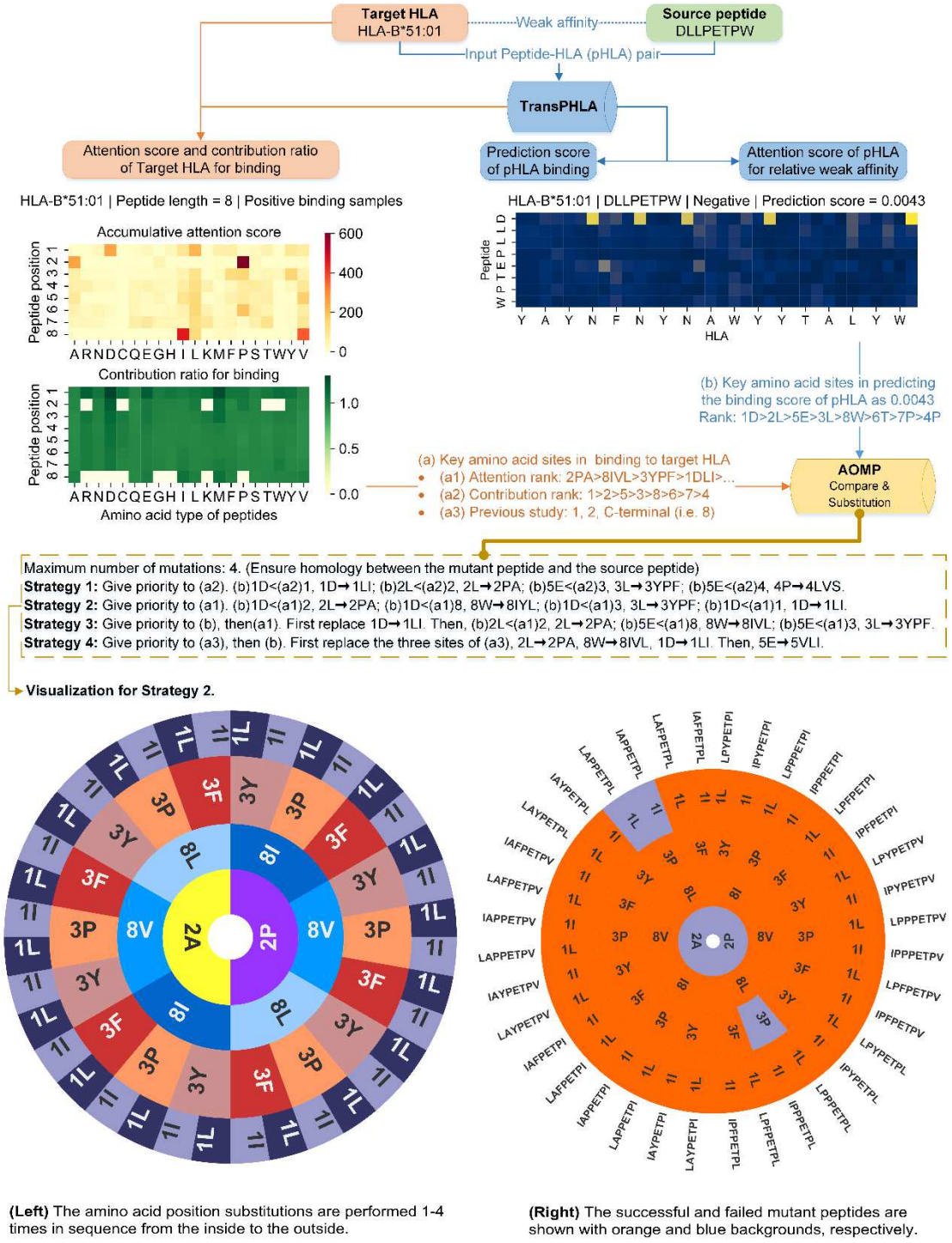
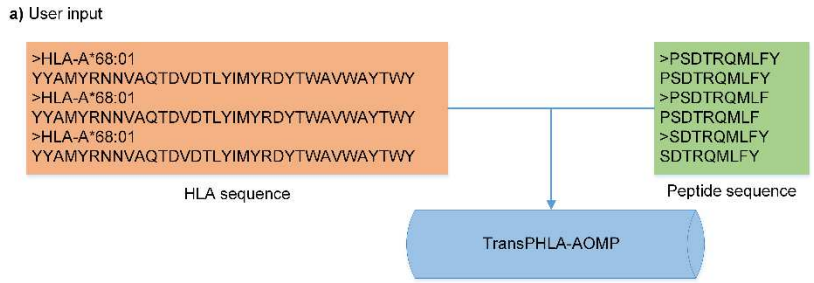


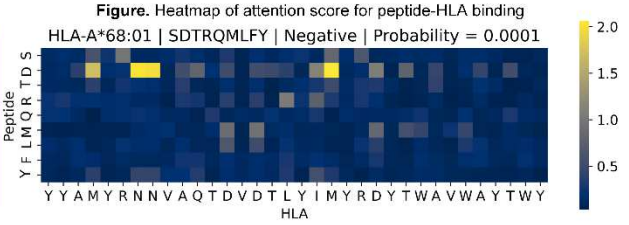
Figure 2. The workflow of automatically optimize mutated peptide (AOMP) program for example peptide DLLPETPW and HLA-B*51:01. The number and the letter, for example 8I, indicate that the amino acid at the 8-th position of the peptide obtained at the previous level is replaced with the amino acid I.



b) TransPHLA for peptide-HLA binding

Table. Prediction results

	HLA	peptide	Binding	Prediction score
1	HLA-A*68:01	PSDTRQMLFY	0	0.0017
2	HLA-A*68:01	PSDTRQMLF	0	0
3	HLA-A*68:01	SDTRQMLFY	0	0.0001



c) Automatically optimize mutated peptide (AOMP) program

HLA	Original peptide	Mutation peptide	Mutation Amino acid site	Mutation number	Sequence similarity	Binding	Prediction score
HLA-A*68:01	SDTRQMLFY	SATRQMLFY	2|D/A	1	0.888889	1	0.8336
HLA-A*68:01	SDTRQMLFY	DATRQMLFY	1|S/D,2|D/A	2	0.888889	1	0.9961
HLA-A*68:01	SDTRQMLFY	SVSRQMLFR	2|D/V,3|T/S,9|Y/R	3	0.666667	1	1
HLA-A*68:01	SDTRQMLFY	ETIRQMLFR	1|S/E,2|D/T,3|T/I,9|Y/R	4	0.666667	1	1

Figure 3. The user input and output results of the web server for TransPHLA and AOMP program.

Results

Comparison of TransPHLA with existing methods

To verify the effectiveness of the TransPHLA, we compare it with 9 benchmark methods from the Immune Epitope Database (IEDB), the recommended method from IEDB (i.e. NetMHCpan_EL[9]), and the state-of-the-art method published in 2021 (i.e. Anthem[26]). The benchmark methods are ANN[24], Consensus[28], NetMHCcons[29], NetMHCpan_BA[10], NetMHCstabpan[31], PickPocket[30], CombLib[27], SMM[25], and SMMPMBEC[26], which can be obtained from http://tools.iedb.org/main/tools-api/. Different methods provide different

scoring methods to determine whether pHLA can bind, such as predicted IC50, predicted score, and percentile rank. We prefer to use predicted IC50 and predicted score as the criteria for the regression task and binary classification task, and Consensus only provides percentile rank as the criterion. Supplementary Table 1 records the details of criteria strategies for different methods[10,27,39].

It is worth noting that not every method is compatible with every HLA and peptide. Except for NetMHCpan_BA, NetMHCpan_EL, and our method, different methods have different limitations. For example, Anthem is an allele-specific method that builds the model for each peptide length of each HLA, and uses complex sequence scoring functions to extract peptide features for each model; thus, it is difficult for users to predict samples when their HLA or peptide length are not provided. The SMM and SMMPMBEC only support peptide-related samples from 8 to 11 in length, and the CombLib only supports peptide-related samples with a length of 9. In a word, under the same data, not every method can predict all the samples provided by the user.

The comparison is performed on pHLA independent test, pHLA external test, neoantigen identification, and HPV vaccine identification (Figure 4).

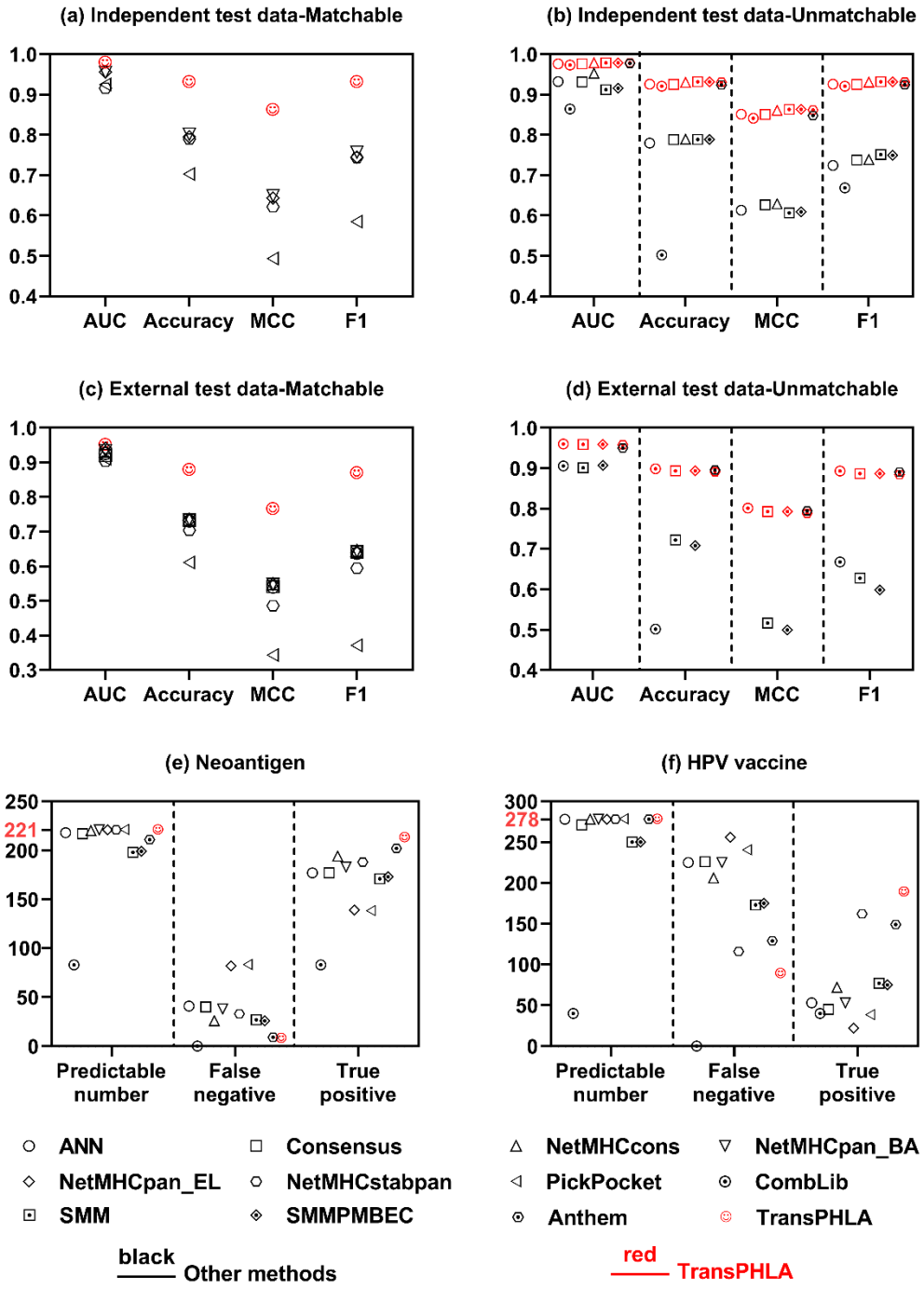


Figure 4. Comparison of the proposed TransPHLA method with 11 existing methods. For subfigures (a) and (c), matchable means that the data of methods in the subgraph is consistent, that is independent or external data. For subfigures (b) and (d), unmatchable means that the data of different methods in the subgraph is not the same, and they are different subsets of the provided

data. For each method, the TransPHLA performs prediction and pairwise comparison on the corresponding subset. For subfigures (e) and (f), the predictable number represents the number of predictable peptide-HLA-I binders.

pHLA test set

Figure 4 shows two perspectives: (i) methods can predict all the provided data (i.e. Matchable subfigures) and (ii) methods can predict only part of the provided data due to their respective limitations (i.e. Unmatchable subfigures). In Matchable subfigures, the data used for the performance comparison of the methods are all consistent, so the performance results can be compared uniformly and directly. In Unmatchable subfigures, HLAs and peptide lengths that can be predicted by different methods are different. Therefore, for each method in these subfigures, the data used for performance comparison is a subset of the provided data. To make the performance comparison fairer and more reasonable, the proposed method performs a pairwise comparison with each method on the corresponding subset data. It can be seen that the proposed method is superior to other methods regardless of whether it is independent data or external data. Although NetMHCpan_EL has achieved good performance on external data, its performance on independent data is greatly reduced. Independent data contains 112 types of HLA alleles, while external data contains only 5 HLA alleles. We mentioned before that the two

test data are complementary in the performance comparison of the methods, thus, only a method that works well on both two types of data can demonstrate its superiority. The Anthem method shows slightly inferior performance than TransPHLA, but it cannot be extended to some unknown HLAs or peptide lengths due to its limited published data.

Also, the performance of each method for each peptide length on the independent test set and external test set is discussed. We show the violin plot for the distribution of AUC, Accuracy, MCC, and F1 of the 12 methods on the independent test set and external test set (Supplementary Figures 1-8). The results also indicate the superiority of the proposed TransPHLA compared to the other 11 methods: (a) TransPHLA is not restricted by HLA allotype and peptide length; (b) For any peptide length, TransPHLA shows superior performance on all metrics with other methods; (c) TransPHLA has a tight distribution on four metrics, especially in peptide length 9, which reflects the potential of TransPHLA to increase the performance as the amount of training data increases. Furthermore, if pHLA data with the number of other peptide lengths or HLAs increases, our method is also likely to achieve better results; (d) MCC shows that TransPHLA is effective for any HLAs under any length. Besides, only Anthem can achieve this. But Anthem is limited by the published data and the performance is not higher than TransPHLA; (e) About 170000 pHLAs, TransPHLA uses 28 seconds to finish the prediction task on GeForce RTX 3080 GPU, and 2 minutes on CPU. Other methods are not so fast. For a

detailed analysis of the results, see Sections 2.1 and 2.2 of the Supplementary Information.

Overall, the proposed method shows its superior predictive ability compared with other methods.

Neoantigen identification

Epitope screening and identification mainly depend on the pHLA binding, especially in neoepitope-based immunotherapy recognized as the most promising cancer treatment, the primary determinant of neoantigen screening is the binding of peptide and autologous specific HLA molecule[40].

We collected neoantigen data from nonsmall-cell lung cancer(NSCLC), melanoma, ovarian cancer, and pancreatic cancer in recent work[32,33], including 221 experimentally verified pHLA binders, which were identified by pHLA multimer-based assays, Enzyme-Linked ImmunoSpot (ELISpot) assay flow cytometry, or other experiments. The comparison results of different methods on this data are shown in Figure 4(e) and Supplementary Table 2. Excitingly, our method was able to screen out 96.4% of neoantigens. Although CombLib has achieved 100% accuracy, it only supports peptides with a length of 9, which limits its application. The remaining 10 methods are not as accurate as our method, and they may be limited by predictable HLAs or peptide lengths.

HPV vaccine identification

In the world, about 80% of people have been exposed to HPV before the age of 50, which is the most common sexually transmitted disease[41]. Currently, HPV cannot be cured, but there are some preventive HPV vaccines. However, the therapeutic effect of these vaccines is limited, and the use rate is very low[42]. Therefore, it is very urgent to develop therapeutic vaccines to treat HPV infections and diseases.

An important study[34] has displayed some experimentally verified HPV vaccine data, including 278 experimentally verified pHLA binders from HPV16 proteins E6 and E7, composed of 8-11 mer peptides. The comparison results of different methods on this data are shown in Figure 4(f) and Supplementary Table 3. Although our method only shows a screening rate of 68%, it is still significantly superior to other methods and can assist experimental research.

TransPHLA uncovers the underlying patterns of pHLA binding

The attention mechanism included in TransPHLA provides biological interpretability for the model. Further, we explored the binding rules of pHLA according to our method. Evidence shows that the C-terminal, N-terminal, and anchor site[43] of the peptide are critical for binding to HLA, which are always located by the first, last, and second position of the peptide sequence. The attention score of the position is confirmed, as shown in Figure 5(a).

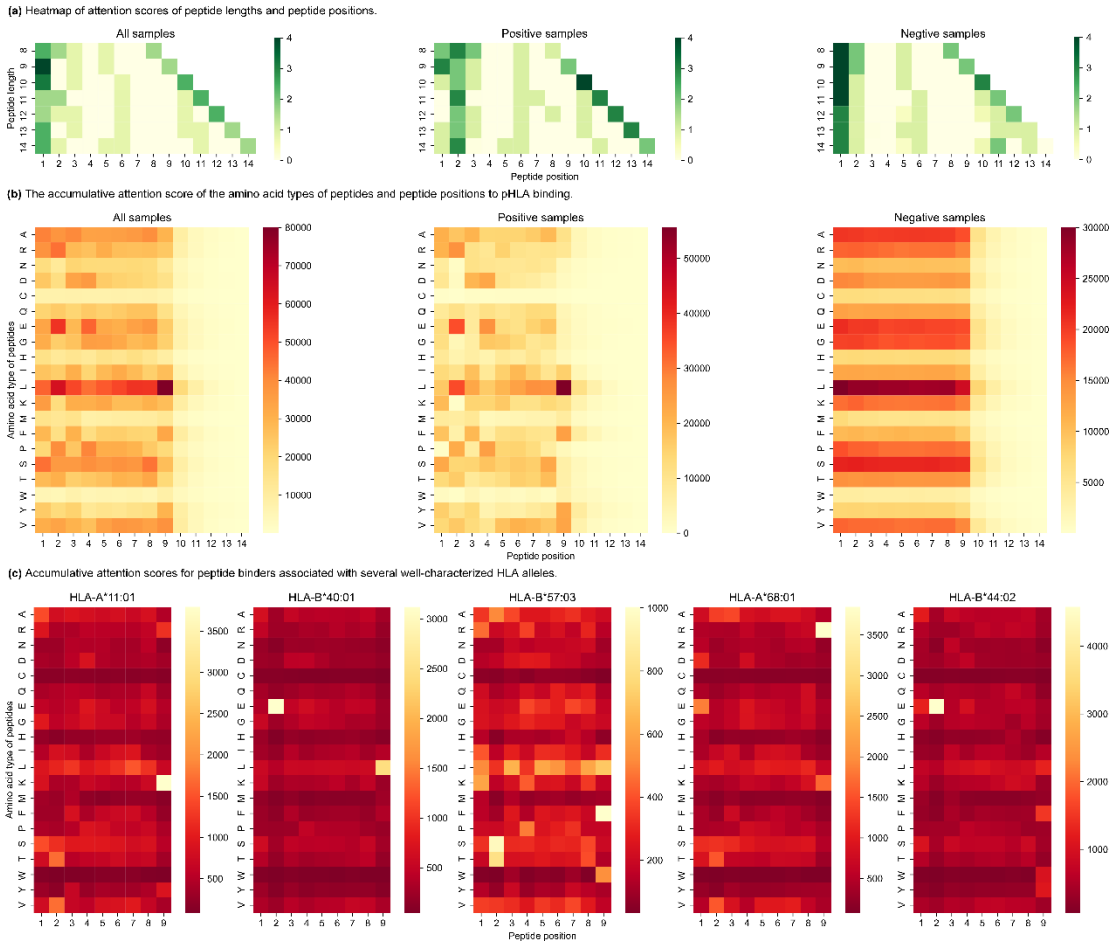


Figure 5. (a) Attention scores associated with all correctly predicted samples, correctly predicted positive samples and correctly predicted negative samples. (b) The contribution (i.e. accumulative attention score) of the amino acid types of peptides and peptide positions to peptide-HLA-I (pHLA) binding. (c) Accumulative attention scores for peptide binders associated with several well-characterized HLA-I alleles. Only 9-mer peptides are focused on here. The brighter residues are considered more important in pHLA binding.

Furthermore, we analyzed the contribution of the amino acid types on the positive and negative samples to the binding and non-binding at different peptide positions (Figure 5(b)), respectively. It can be found that the binding

and non-binding of pHLA are affected by different components of peptides. In addition, we analyzed the influence of 20 amino acids at different peptide positions for binding or non-binding for all 366 HLA-peptide length combinations, the attention scores and corresponding heatmaps can be downloaded from our webserver. These results will not only help us understand the mechanism of pHLA binding, but also can be used for vaccine design, as shown in Sections 2.3 and 4.5.

In addition, since the attention score represents the pattern of pHLA binding, it implies that the key amino acid sites on the peptide sequence which are important for binding or non-binding to the target HLA. To better persuade, we visualized the binding pattern of 5 HLA alleles according to ACME[44] (Figure 5(c)). As expected, TransPHLA found a similar pattern for amino acid types at different peptide positions to the previous studies[44,45]. For HLA-A*11:01, TransPHLA recognizes the anchor residue that the peptides with K (Lys) at position 9 (9-th K). For HLA-B*40:01, the important residues, which are 2-nd E (Glu) and 9-th L (Leu), were successfully identified by TransPHLA. For HLA-B*57:03, hydrophobic residues usually form the binding pocket, and we identified this preference through 9-th L, 9-th F (Phe), and 9-th W (Trp), which is consistent with 2BVP[46]. For HLA-A*68:01, 4HWZ[47] demonstrates the 9-th K and 9-th R (Arg) residues of the peptide significantly contribute to the binding. For HLA-B*44:02, the significance of 2-nd E has been proved by 1M6O[48]. All these results have been supported by previous studies, and demonstrate the

effectiveness of our methods.

AOMP program

AOMP program is proposed to search for mutant peptides with higher affinity given the specific important peptide with weak affinity for a specific HLA allele. For example, Figure 2 visualizes the process of AOMP and automatic mutation of strategy 2 for source peptide DLLPETPW and target HLA-B*51:01.

To demonstrate the effectiveness of our program, we tested all 366 HLA-peptide length combinations. For each HLA-peptide length combination, we randomly selected 10 negative pHLA that were correctly predicted by TransPHLA. Finally, we ran the automatic mutation program for 3660 negative pHLA, of which 3633 were successful. Among them, 1692, 1536, 350, and 55 were successfully mutated at least 1, 2, 3, and 4 amino acid sites, respectively. Further, 88% of the source peptides can find successful mutant peptides with a homology of more than 80%, which is exciting for vaccine design.

To further verify the authenticity and usability of the mutation results, we use the NetMHCpan_BA recommended by IEDB[9] to validate mutation results for 3660 pHLA. According to the predicted IC50 threshold of 500, 2905 (80%) are successful. And 3418 (94%) are successful according to percentile rank threshold of 2. Overall, 3419 (94%) are successful in either IC50 or percentile rank.

Discussion

pHLA binding and interaction are critical to epitope presentation and prerequisite for subsequent T-cell recognition which initiates an effective immune response. Epitope screening and identification mainly depend on the affinity of pHLA, especially in neoepitope-based immunotherapy recognized as the most promising cancer treatment, the primary determinant of neoantigen screening is the affinity of peptide and autologous specific HLA molecule. Therefore, accurate pHLA binding prediction is essential for the identification of immunotherapy targets, epitope screening, and vaccine design. On the other hand, the short peptide-based vaccine design is another important field for the treatment of diseases. However, the current vaccine design method is still in its infancy and cannot yet be automated.

Firstly, we proposed a TransPHLA method for pHLA binding prediction, which is a generalized pan-specific model that is not restricted by HLA allele and peptide length. We conducted two types of independent tests and two types of case studies (that are neoantigen and HPV vaccine identification). Compared with the state-of-the-art method (i.e. Anthem), IEDB recommended method (i.e. NetMHCpan_EL), and 9 benchmark methods, the proposed TransPHLA achieves significantly superior performance on all four experiments.

Then, based on TransPHLA, we develop an AOMP program to search for mutant peptides with higher affinity to the target HLA and high homology with the source peptide. Among 3660 negative pHLA for different HLAs and peptide

lengths, 3630 samples are successfully found for the binding mutant peptide-HLA, 94% were verified by the method recommended by IEDB, 88% with a homology of more than 80%, which is exciting for vaccine design.

This is the first attempt to propose a TransMut framework in the field of automatic mutation of biomolecules. Excitingly, it has achieved success in the field of pHLA binding prediction and peptide-based vaccine design. It is worth noting that this framework can be applied to any binding prediction and mutation task of biomolecules.

Methods

Dataset

In this study, the pHLA binding data (that is, positive data) are obtained from Anthem[23], which can be downloaded from https://github.com/17shutao/Anthem/tree/master/Dataset. The negative data is generated in a similar way as in previous studies[8,9,49]. For each binder length and each HLA allele, peptides of negative data are sequence segments that are randomly chosen from the source proteins of IEDB HLA-I immunopeptidomes. Although false-negative peptides may be generated, the possibility and proportion of such peptides are very small[1,50] and can be ignored. This strategy of constructing negative samples guarantees that the dataset is balanced (Supplementary Table 4).

To compare with previous methods conveniently, we followed the training and evaluation strategy of Anthem[23], which is the state-of-the-art pHLA binding

prediction method. There are three types of datasets for different purposes: the training set for model training and model selection, the independent test set and the external test set for model evaluation and methods comparison. The positive data sources of training and that of the independent test set are the same, including (i) 4 public HLA binders databases, which are IEDB[51], EPIMHC[52], MHCBN[53], and SYFPEITHI[54]; (ii) Allotype-specific HLA ligands identified by mass spectrometry in previously published studies[55-70], and (iii) Peptide binders from training datasets of other pHLA binding prediction tools[44,51,71-82]. The external test set is experimentally verified by Anthem[23].

Further, we have checked and deleted some error or duplicated samples, for example, “HLA-B*07:01”-related samples are ignored because its sequence contains errors. The statistics of the three types of datasets are listed in Supplementary Table 4. The number of pHLA binders for each peptide binder length of each HLA spans a large range, from 10^1 to 10^5, the details are shown in Supplementary Figure 9. On the other hand, the common peptide binder lengths are 8-14. For different peptide binder lengths, there are big gaps in the number of pHLA binders. In Supplementary Figure 10, the number of peptide binders with length 9 is very large, while 13 and 14 are very few. This leads to differences in the performance of the method on different peptide binder lengths.

Experiment settings

To follow the previous studies[8,23] for pHLA binding prediction, we conducted

the 5-fold cross-validation (CV) and independent test. Since the source of the independent test set and the training set are the same, the data distributions between the training set and independent test set are very similar (see Supplementary Figures 9 and 10). When the model is tested on the data with a similar distribution to the training data, it will be easier to obtain better test performance than a model that is not trained with a similar distribution of test data. That is, the proposed method and Anthem[23] may obtain greater advantages than other methods on the independent test set. Thus, we set up an external test to perform a fairer performance comparison of different methods.

5-fold CV

The 5-fold CV used in this study has two purposes, which are trained model evaluation and selection. It divides the training set into five parts equally, four of which are used for model training, and the remaining part is used for evaluation of the model with the same parameters. The training and evaluation process is repeated 5 times to ensure that each part of data participates four times for model training and once for model evaluation. Finally, the average result of 5 model evaluations is used as the final evaluation result. Usually, using CV can avoid the overfitting of the model to a certain extent.

Independent test

Independent test is also a popular strategy for performance evaluation of methods. Independent test data require that it cannot have any overlap with training data. It is used as unified data to evaluate the performance of different methods. Therefore, compared with the CV, using the same independent test data will more objectively evaluate the performance and generalization ability of different methods.

External test

For a fair comparison, we used experimental data as external test data to eliminate possible deviations due to the similar data distribution. According to Supplementary Figures 9 and 10, the data distribution of the external test is a little bit different from that of the training and independent test data. Like the independent test, it can also more objectively evaluate the performance and generalization ability of the method.

Performance evaluation metrics

For each predictive model or method, the following metrics are calculated:

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \quad (1)$$

$$\text{MCC} = \frac{(\text{TP}\times\text{TN})-(\text{FN}\times\text{FP})}{\sqrt{(\text{TP}+\text{FN})\times(\text{TN}+\text{FP})\times(\text{TP}+\text{FP})\times(\text{TN}+\text{FN})}} \quad (2)$$

$$\text{F}_1-\text{score} = \frac{2\cdot\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}}, \text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}, \text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \quad (3)$$

where TP is true positive, FP is false positive, FN is false negative, TN is true

negative, and MCC is Matthews Correlation Coefficient. In addition, we adopt AUC, i.e. the area under the receiver operating characteristic curve, as another performance evaluation metric. The above four metrics can comprehensively evaluate the performance of models or methods.

Except for MCC, which ranges from -1 to 1, other metrics range from 0 to 1. The higher value of the metric, the better the model or method. It is worthwhile to note that MCC cannot be calculated when two of TN, TP, FN, FP are 0, because the denominator is 0. And this phenomenon is not caused by both FN and FP being 0. Thus, if the MCC cannot be calculated for a specific peptide length of a specific HLA, it implies that the method is invalid for this HLA with this peptide length.

TransPHLA

The core idea of TransPHLA is the application of the self-attention mechanism[22], and it is composed of four blocks to achieve high prediction performance. The embedding block added positional embedding to the amino acid embedding, to generate the sequence embedding, and then a dropout technology is used to enhance the robustness. Through the embedding block, TransPHLA generates the input embedding for peptide and HLA respectively. Next, input the peptide and HLA embedding into the encoder block, respectively, which contains the masked multi-head self-attention mechanism and the feature optimization block. The feature optimization block is a combination of

fully connected layers in which the channel of the gyro first rises and then falls. This module makes the feature representation obtained by the attention mechanism better, mainly because more layers are added. Then, the output feature representations of peptide and HLA are concatenated as the embedding of the pHLA pair. After pHLA pair embedding passes through the encoder block, the projection block is used to predict the pHLA binding score. The details can be seen in Figure 1.

Sequence embedding composed of amino acid embedding and positional embedding

First, each peptide and HLA are padded to the maximum length of 15 and 34, respectively, to handle the variable input length. Then, the character embedding model is used to create a unique embedding for each amino acid, where the dimension of the embedding is defined as d_X. Taking the peptide SDKYGLGY as an example, which has a length of 8. From Supplementary Figure 11(a), embeddings of 6 different amino acids are different, and embeddings of padding rows are all the same.

On the other hand, the order of amino acids is critical to the structure of the peptide and HLA sequence, but the above embedding method does not consider it. Thus, we apply positional embedding to encode the position of the amino acid in the sequence. Given the position p in the sequence, the positional embedding is also encoded as a d_X-dimensional vector, the value of the i-th

element of this vector is $PE(p)_i$, the formulas are

$$PE(p)_{2i} = sin\left(p/10000^{2i/d_x}\right) \quad (4)$$

$$PE(p)_{2i+1} = cos\left(p/10000^{2i/d_x}\right) \quad (5)$$

where 2*i* represents the even dimensions, and 2*i*+1 represents the odd dimensions. This position embedding method can not only reflect the absolute position information of the amino acid, but also the relative position information. We visualize positional embedding in Supplementary Figure 11(b). It is worth noting that for any peptide or HLA, position embedding is the same. And we conducted the ablation experiment for positional embedding, and demonstrated its validility in the TransPHLA, the details can be seen in Section 5 of Supplementary Information.

Finally, the amino acid embedding and positional embedding are summed to obtain the sequence embedding (shown in Supplementary Figure 11(c)).

Masked multi-head self-attention mechanism

The attention mechanism is to focus on the important information and reduce the impact of unimportant information from a large amount of information. Its essence is mapping the Query *Q* to a set of Key-Value (*K*-*V*) pairs then obtain an output, where *K*-*V* pairs are the form of storing sequence elements in memory. It reflects the attention score (i.e. weight) according to the correlation or similarity of *Q* and *K*. The attention score represents the importance of information (i.e. *V*). The larger the attention score, the more

focused on the corresponding information.

Compared with RNNs, Transformer realizes parallelization and solves the long-term dependencies problem, hence, it can process the data faster than RNNs. Compared with CNNs, which extract local information commendably, Transformer extracts more global information, which is suitable for the information exploration of the whole sequence of peptide and HLA.

The self-attention mechanism belongs to a variant of the attention mechanism, which captures the internal correlation of a sequence, and reduces the dependence on external information. In this mechanism, *Q*, *K*, and *V* matrices are all generated based on the input sequence S.

The calculation process of the self-attention mechanism can be summarized into three processes (Figure 6): (1) Calculate *Q*, *K*, and *V* according to input embedding *X* of sequence S; (2) The similarity or correlation between *Q* and *K* is calculated to generate the weight of *V* (i.e. attention score); (3) The attention (i.e. output) is calculated by the weighted sum of *V*.

Multi-head attention uses multiple *Q*, *K*, and *V* to calculate multiple attentions in parallel, where each attention mechanism focuses on a different information pattern of the input sequence. Suppose *h* attentions are to be calculated, that is, the number of heads is *h*. The *Q*, *K*, and *V* obtained by linear mapping from *X* are divided into *h* parts, and then the attention is calculated for each part. Concatenate these attentions, and then perform linear mapping again to obtain the final output, whose dimension is the same as *X*. The detailed

process can be seen in Figure 6. The whole process is expressed by the formula as follows:

$$Attention\ score_i = \mathrm{softmax}\left(\frac{Q_i {K_i}^T}{\sqrt{d_K}}\right) \quad (6)$$

$$Attention\ score_i = \mathrm{softmax}\left(\frac{Q_i {K_i}^T}{\sqrt{d_K}}\right) \quad (7)$$

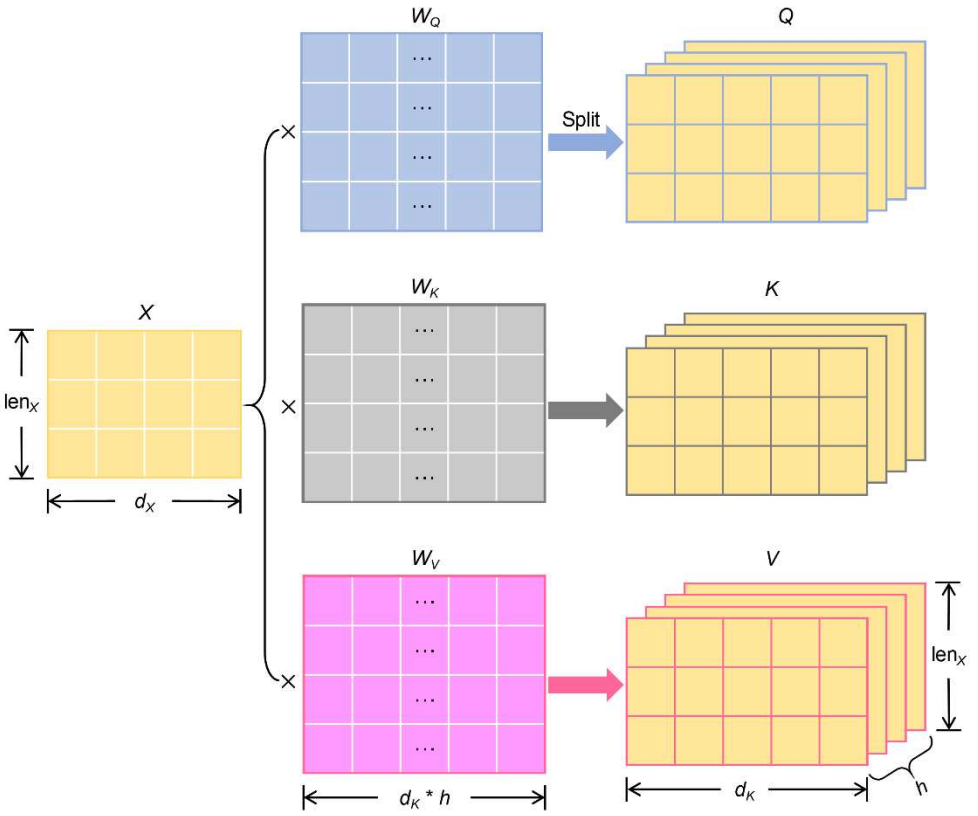
$$Attention_i = Attention\ score_i V_i \quad (8)$$

$$Output = Concat(Attention_1, \ldots, Attention_h)W \quad (9)$$

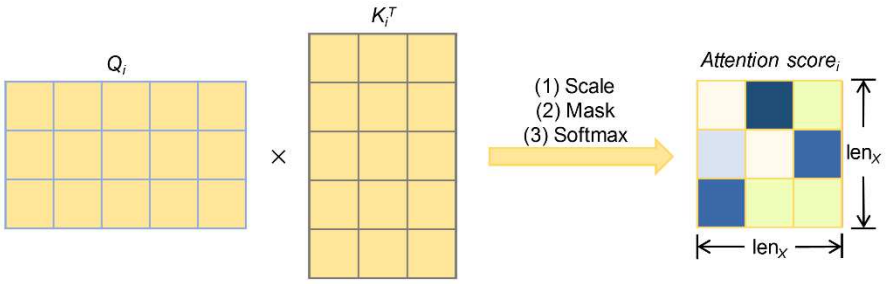
where *i* represents the *i*-th head attention, d_K is the dimension of K_i, and $\sqrt{d_K}$ is the scaled factor to prevent the large dot products.

It is worth noting that this study introduced the mask operation when calculating the attention. For peptide or HLA sequences whose length is less than the corresponding maximum length, non-amino acid characters should not be considered for the model training. Therefore, we use 10^{-9} that is very close to 0 as their attention scores, so that non-amino acid characters do not play a role in calculating the attention.

(a) Generate Q, K, V matrices according to the embedding X of sequence S.



(b) Calculate attention scores.



(c) Calculate the output of multi-head self-attention.

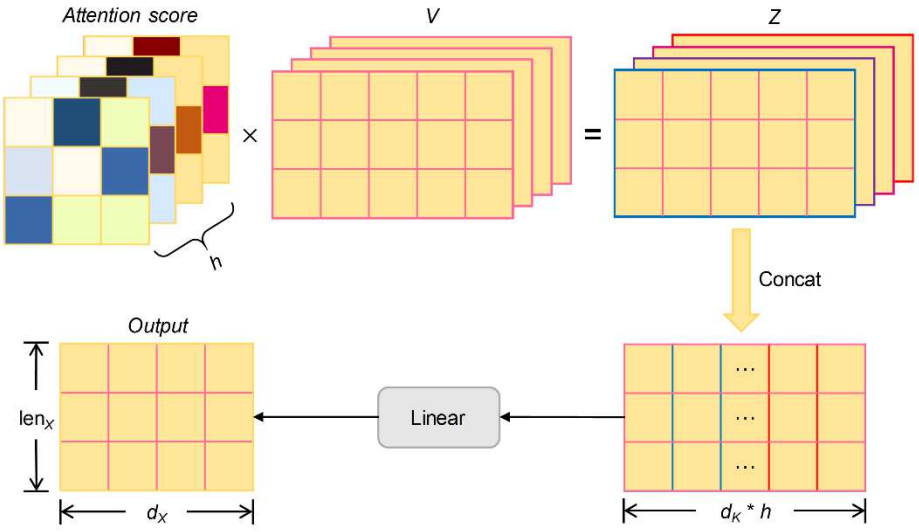


Figure 6. The calculation process of masked multi-head self-attention mechanism, where *i* represents the *i*-th head attention, *h* is the number of heads, len_X is the length of sequence S, d_X and d_K are the dimensions of X and K_i, and $\sqrt{d_K}$ is the scaled factor to prevent the large dot products.

AOMP program

In this study, we developed an AOMP program for the first time. This program aims to search for higher affinity mutant peptides based on the specific important peptide with weak affinity for a specific HLA allele. For example, the specific important peptides can be E6 and E7 peptide from HPV, neoantigen, and TNF epitope.

The program designed four directed mutation strategies based on the attention score obtained by TransPHLA (Figure 2). According to the analysis in Section 2.2, the attention score not only represents the pattern of pHLA binding, but also reveals that the key amino acid sites on the peptide sequence which are important for binding or non-binding to the target HLA. On the other hand, for effective vaccine design, we also considered the homology of the mutant peptide and the source peptide. The homology between the mutant peptide and the source peptide is calculated by sequence similarity, and experiments show that the similarity is very close to the blast result. The homology of 1, 2, 3, and 4 amino acid positions were mutated on average 90%, 80%, 70%, and 61%,

respectively. Therefore, we limit the number of mutations in the amino acid site of the source peptide to no more than 4.

For each of the 366 HLA-peptide length combinations, we have established a binding contribution matrix of 20 amino acids at each peptide position. To adapt to the new or unknown HLA-peptide length combination, a general binding contribution matrix is established. We provide these 367 contribution matrices and their visual heatmap figures on the webserver. On the other hand, when predicting relatively weak affinity pHLA, the attention score obtained by TransPHLA is used to calculate the contribution matrix of each amino acid site on the peptide. We also provide an attention score heatmap of the pHLA if the user needs it.

Subsequently, four optimization strategies are designed, with details as follows. We calculated two contribution rate matrices based on the above two contribution matrices. The larger the element value in the contribution matrix, means that the corresponding amino acid site is more critical for binding or non-binding. Intuitively, since the amino acid site contributes more to non-binding prediction, if we replace them with other amino acids which contribute more to binding prediction, the mutated peptide is more likely to have a better affinity with the target HLA. Based on the above four matrices, we designed four strategies to generate mutant peptides. Their main idea is to compare the amino acid sites on the source peptide that have a large impact on weak affinity and the amino acid sites on the target HLA-peptide length that contribute

significantly to the high affinity. Then, the corresponding amino acid substitutions are made according to the comparison results. The process is as follows: (1) Predict the binding score for source peptide and target HLA; (2) Find some most important amino acid sites based on the self-attention mechanism; (3) Replace these important sites of weak affinity pHLA with some amino acids which may contribute more on binding prediction; (4) Select some best mutation candidates for evaluation.

For the source peptide and the target HLA (i.e. the specific pHLA), the mutant peptides generated by the four strategies will be merged, and the duplicates will be removed. Then, TransPHLA will screen and retain mutant peptides that can bind to the target HLA. Excitingly, the original intention of this program is for non-binding pHLA, and we found it can also find mutant peptides with a stronger affinity for binding pHLA.

Figure 2 visualizes the process of the AOMP program, and shows the automatic mutation of strategy 2 for source peptide DLLPETPW and target HLA-B*51:01 as an example.

Availability

The webserver is at https://issubmission.sjtu.edu.cn/TransPHLA-AOMP/index.html.

The data and code are available at https://github.com/a96123155/TransPHLA-AOMP.

The attention score and heatmap of amino acid types and position of peptides for specific HLA and peptide binder length can be downloaded from https://issubmission.sjtu.edu.cn/TransPHLA-AOMP/download.html.

Acknowledgements

This work was supported by the grants from the National Science Foundation of China [grant numbers 32070662, 61832019, 32030063], the Key Research Area Grant [2016YFA0501703] of the Ministry of Science and Technology of China, the Science and Technology Commission of Shanghai Municipality [grant number 19430750600], as well as SJTU JiRLMDS Joint Research Fund and Joint Research Funds for Medical and Engineering and Scientific Research at Shanghai Jiao Tong University [YG2021ZD02]. The computations were partially performed at the Pengcheng Lab. and the Center for High-Performance Computing, Shanghai Jiao Tong University.

Author contributions

Yanyi Chu and Yan Zhang conceived the original ideas of this study, designed and performed the experiments, and co-wrote the manuscript. Qiankun Wang helped with molecular dynamics. Lingfeng Zhang helped with the model designed and manuscript revised. Xuhong Wang participated in the design of the initial frame of the model. Yanjing Wang helped on the webserver building. Jianmin Wang proposed some ideas in drawing figures. Xue Jiang

participated in the discussion of the framework of the manuscript. Dennis Russell Salahub, Yi Xiong and Dong-Qing Wei guided the work. All the authors discussed the results and commented on the manuscript.

Competing interests

The authors declare no competing financial interests.

References

1 Yewdell, J. W. & Bennink, J. R. Immunodominance in major histocompatibility complex class I–restricted T lymphocyte responses. *Annual review of immunology* **17**, 51-88 (1999).

2 Huppa, J. B. *et al.* TCR–peptide–MHC interactions in situ show accelerated kinetics and increased affinity. *Nature* **463**, 963-967 (2010).

3 Jensen, P. E. Recent advances in antigen processing and presentation. *Nature immunology* **8**, 1041-1048 (2007).

4 Chang, S.-C., Momburg, F., Bhutani, N. & Goldberg, A. L. The ER aminopeptidase, ERAP1, trims precursors to lengths of MHC class I peptides by a “molecular ruler” mechanism. *Proceedings of the National Academy of Sciences* **102**, 17107-17112 (2005).

5 Kloetzel, P. M. Generation of major histocompatibility complex class I antigens: functional interplay between proteasomes and TPPII. *Nature immunology* **5**, 661-669 (2004).

6 Unanue, E. R. From antigen processing to peptide-MHC binding. *Nature immunology* **7**, 1277-1279 (2006).

7 Trowsdale, J. HLA genomics in the third millennium. *Current opinion in Immunology* **17**, 498-504 (2005).

8 Mei, S. *et al.* A comprehensive review and performance evaluation of bioinformatics tools for HLA class I peptide-binding prediction. *Briefings in bioinformatics* **21**, 1119-1135 (2020).

9 Reynisson, B., Alvarez, B., Paul, S., Peters, B. & Nielsen, M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic acids research* **48**, W449-W454 (2020).

10 Zhang, L., Udaka, K., Mamitsuka, H. & Zhu, S. Toward more accurate pan-specific MHC-peptide binding prediction: a review of current methods and tools. *Briefings in bioinformatics* **13**, 350-364 (2012).

11 Govindarajan, K. R., Kangueane, P., Tan, T. W. & Ranganathan, S. MPID: MHC-Peptide Interaction Database for sequence-structure-function information on peptides binding to MHC molecules. *Bioinformatics* **19**, 309-310 (2003).
12 Purcell, A. W., McCluskey, J. & Rossjohn, J. More than one reason to rethink the use of peptides in vaccine design. *Nature reviews Drug discovery* **6**, 404-414 (2007).
13 Koşaloğlu-Yalçın, Z. *et al.* Predicting T cell recognition of MHC class I restricted neoepitopes. *Oncoimmunology* **7**, e1492508 (2018).
14 Rizvi, N. A. *et al.* Mutational landscape determines sensitivity to PD-1 blockade in non–small cell lung cancer. *Science* **348**, 124-128 (2015).
15 Snyder, A. *et al.* Genetic basis for clinical response to CTLA-4 blockade in melanoma. *New England Journal of Medicine* **371**, 2189-2199 (2014).
16 Ott, P. A. *et al.* An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature* **547**, 217-221 (2017).
17 Sahin, U. *et al.* Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. *Nature* **547**, 222-226 (2017).
18 Slingluff Jr, C. L. The present and future of peptide vaccines for cancer: single or multiple, long or short, alone or in combination? *Cancer journal (Sudbury, Mass.)* **17**, 343 (2011).
19 Gfeller, D., Bassani-Sternberg, M., Schmidt, J. & Luescher, I. F. Current tools for predicting cancer-specific T cell immunity. *Oncoimmunology* **5**, e1177691 (2016).
20 Linnemann, C. *et al.* High-throughput epitope discovery reveals frequent recognition of neo-antigens by CD4+ T cells in human melanoma. *Nature medicine* **21**, 81-85 (2015).
21 Bentzen, A. K. & Hadrup, S. R. Evolution of MHC-based technologies used for detection of antigen-responsive T cells. *Cancer immunology, immunotherapy* **66**, 657-666 (2017).
22 Vaswani, A. *et al.* Attention is all you need. *arXiv preprint arXiv:1706.03762* (2017).
23 Mei, S. *et al.* Anthem: a user customised tool for fast and accurate prediction of binding between peptides and HLA class I molecules. *Briefings in Bioinformatics* (2021).
24 Andreatta, M. & Nielsen, M. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics* **32**, 511-517 (2016).
25 Peters, B. & Sette, A. Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method. *BMC Bioinformatics* **6**, 132, doi:10.1186/1471-2105-6-132 (2005).
26 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* **10**, 1-11 (2009).
27 Sidney, J. *et al.* Quantitative peptide binding motifs for 19 human and mouse MHC class I molecules derived using positional scanning combinatorial peptide libraries. *Immunome research* **4**, 1-14 (2008).
28 Moutaftsi, M. *et al.* A consensus epitope prediction approach identifies the breadth of murine T CD8+-cell responses to vaccinia virus. *Nature biotechnology* **24**, 817-819 (2006).
29 Karosiene, E., Lundegaard, C., Lund, O. & Nielsen, M. NetMHCcons: a consensus method for the major histocompatibility complex class I predictions. *Immunogenetics* **64**, 177-186 (2012).
30 Zhang, H., Lund, O. & Nielsen, M. The PickPocket method for predicting binding specificities for receptors based on receptor pocket similarities: application to MHC-peptide binding. *Bioinformatics* **25**, 1293-1299 (2009).
31 Rasmussen, M. *et al.* Pan-specific prediction of peptide–MHC class I complex stability, a correlate of T cell immunogenicity. *The Journal of Immunology* **197**, 1517-1524 (2016).
32 Wells, D. K. *et al.* Key parameters of tumor epitope immunogenicity revealed through a consortium approach improve neoantigen prediction. *Cell* **183**, 818-834. e813 (2020).
33 Wang, G. *et al.* INeo-Epp: A novel T-cell HLA class-I Immunogenicity or neoantigenic epitope prediction method based on sequence-related amino acid features. *BioMed research international* **2020** (2020).
34 Bonsack, M. *et al.* Performance evaluation of MHC class-I binding prediction tools based on an experimentally validated MHC–peptide binding data set. *Cancer immunology research* **7**, 719-736 (2019).
35 Ebrahimi, S., Mohabatkar, H. & Behbahani, M. Predicting promiscuous T cell epitopes for designing a vaccine against Streptococcus pyogenes. *Applied biochemistry and biotechnology* **187**, 90-100 (2019).
36 Zhavoronkov, A. *et al.* Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nature biotechnology* **37**, 1038-1040 (2019).
37 Feldmann, M. & Maini, R. N. TNF defined as a therapeutic target for rheumatoid arthritis and other autoimmune diseases. *Nature medicine* **9**, 1245-1250 (2003).
38 Le Buanec, H. *et al.* TNFα kinoid vaccination-induced neutralizing antibodies to TNFα protect mice from autologous TNFα-driven chronic and acute inflammation. *Proceedings of the National Academy of Sciences* **103**, 19442-19447 (2006).
39 Lundegaard, C. *et al.* NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8–11. *Nucleic acids research* **36**, W509-W512 (2008).
40 Vitiello, A. & Zanetti, M. Neoantigen prediction and the need for validation. *Nature biotechnology* **35**, 815-817 (2017).

41 Hathaway, J. K. HPV: diagnosis, prevention, and treatment. *Clinical obstetrics and gynecology* **55**, 671-680 (2012).
42 Yang, A., Farmer, E., Wu, T. & Hung, C.-F. Perspectives for therapeutic HPV vaccine development. *Journal of biomedical science* **23**, 1-19 (2016).
43 Chowell, D. *et al.* TCR contact residue hydrophobicity is a hallmark of immunogenic CD8+ T cell epitopes. *Proceedings of the National Academy of Sciences* **112**, E1754-E1762 (2015).
44 Hu, Y. *et al.* ACME: pan-specific peptide–MHC class I binding prediction through attention-based deep neural networks. *Bioinformatics* **35**, 4946-4954 (2019).
45 Yusim, K. *et al.* Hiv molecular immunology 2015. (Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2016).
46 Stewart-Jones, G. B. *et al.* Structures of three HIV-1 HLA-B* 5703-peptide complexes and identification of related HLAs potentially associated with long-term nonprogression. *The Journal of Immunology* **175**, 2459-2468 (2005).
47 Niu, L. *et al.* Structural basis for the differential classification of HLA-A* 6802 and HLA-A* 6801 into the A2 and A3 supertypes. *Molecular immunology* **55**, 381-392 (2013).
48 Macdonald, W. A. *et al.* A naturally selected dimorphism within the HLA-B44 supertype alters class I structure, peptide repertoire, and T cell recognition. *The Journal of experimental medicine* **198**, 679-691 (2003).
49 Jurtz, V. *et al.* NetMHCpan-4.0: improved peptide–MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. *The Journal of Immunology* **199**, 3360-3368 (2017).
50 Larsen, M. V. *et al.* An integrative approach to CTL epitope prediction: a combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions. *European journal of immunology* **35**, 2295-2303 (2005).
51 Dhanda, S. K. *et al.* IEDB-AR: immune epitope database—analysis resource in 2019. *Nucleic acids research* **47**, W502-W506 (2019).
52 Reche, P. A., Zhang, H., Glutting, J.-P. & Reinherz, E. L. EPIMHC: a curated database of MHC-binding peptides for customized computational vaccinology. *Bioinformatics* **21**, 2140-2141 (2005).
53 Lata, S., Bhasin, M. & Raghava, G. P. MHCBN 4.0: A database of MHC/TAP binding peptides and T-cell epitopes. *BMC research notes* **2**, 1-6 (2009).
54 Rammensee, H.-G., Bachmann, J., Emmerich, N. P. N., Bachor, O. A. & Stevanović, S. SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics* **50**, 213-219 (1999).
55 Mommen, G. P. *et al.* Expanding the detectable HLA peptide repertoire using electron-transfer/higher-energy collision dissociation (EThcD). *Proceedings of the National Academy of Sciences* **111**, 4507-4512

(2014).
56 Hassan, C. *et al.* Naturally processed non-canonical HLA-A* 02: 01 presented peptides. *Journal of Biological Chemistry* **290**, 2593-2603 (2015).
57 Marcilla, M. *et al.* Increased diversity of the HLA-B40 ligandome by the presentation of peptides phosphorylated at their main anchor residue. *Molecular & Cellular Proteomics* **13**, 462-474 (2014).
58 Mobbs, J. I. *et al.* The molecular basis for peptide repertoire selection in the human leukocyte antigen (HLA) C* 06: 02 molecule. *Journal of Biological Chemistry* **292**, 17203-17215 (2017).
59 Yair-Sabag, S. *et al.* The Peptide Repertoire of HLA-B27 may include Ligands with Lysine at P2 Anchor Position. *Proteomics* **18**, 1700249 (2018).
60 Müller, M., Gfeller, D., Coukos, G. & Bassani-Sternberg, M. 'Hotspots' of antigen presentation revealed by human leukocyte antigen ligandomics for neoantigen prioritization. *Frontiers in immunology* **8**, 1367 (2017).
61 Abelin, J. G. *et al.* Defining HLA-II ligand processing and binding rules with mass spectrometry enhances cancer epitope prediction. *Immunity* **51**, 766-779. e717 (2019).
62 Kalaora, S. *et al.* Use of HLA peptidomics and whole exome sequencing to identify human immunogenic neo-antigens. *Oncotarget* **7**, 5110 (2016).
63 Faridi, P., Purcell, A. W. & Croft, N. P. In immunopeptidomics we need a sniper instead of a shotgun. *Proteomics* **18**, 1700464 (2018).
64 Schellens, I. M. *et al.* Comprehensive analysis of the naturally processed peptide repertoire: differences between HLA-A and B in the immunopeptidome. *PloS one* **10**, e0136417 (2015).
65 Abelin, J. G. *et al.* Mass spectrometry profiling of HLA-associated peptidomes in mono-allelic cells enables more accurate epitope prediction. *Immunity* **46**, 315-326 (2017).
66 Schittenhelm, R. B., Sian, T. C. L. K., Wilmann, P. G., Dudek, N. L. & Purcell, A. W. Revisiting the arthritogenic peptide theory: quantitative not qualitative changes in the peptide repertoire of HLA–B27 allotypes. *Arthritis & rheumatology* **67**, 702-713 (2015).
67 Illing, P. T. *et al.* HLA-B57 micropolymorphism defines the sequence and conformational breadth of the immunopeptidome. *Nature communications* **9**, 1-13 (2018).
68 Marcilla, M. *et al.* Comparative analysis of the endogenous peptidomes displayed by HLA-B* 27 and Mamu-B* 08: two MHC class I alleles associated with elite control of HIV/SIV infection. *Journal of proteome research* **15**, 1059-1069 (2016).
69 Hillen, N. *et al.* Essential differences in ligand presentation and T cell epitope recognition among HLA molecules of the HLA-B44 supertype.

European journal of immunology **38**, 2993-3003 (2008).
70 Kaur, G. *et al.* Structural and regulatory diversity shape HLA-C protein expression levels. *Nature communications* **8**, 1-12 (2017).
71 Liu, G. *et al.* PSSMHCpan: a novel PSSM-based software for predicting class I peptide-HLA binding affinity. *Giga Science* **6**, gix017 (2017).
72 O'Donnell, T. J. *et al.* MHCflurry: open-source class I MHC binding affinity prediction. *Cell systems* **7**, 129-132. e124 (2018).
73 Liu, Z. *et al.* DeepSeqPan, a novel deep convolutional neural network model for pan-specific class I HLA-peptide binding affinity prediction. *Scientific reports* **9**, 1-10 (2019).
74 Phloyphisut, P., Pornputtapong, N., Sriswasdi, S. & Chuangsuwanich, E. MHCSeqNet: a deep neural network model for universal MHC binding prediction. *BMC bioinformatics* **20**, 1-10 (2019).
75 Boehm, K. M., Bhinder, B., Raja, V. J., Dephoure, N. & Elemento, O. Predicting peptide presentation by major histocompatibility complex class I: an improved machine learning approach to the immunopeptidome. *BMC bioinformatics* **20**, 1-11 (2019).
76 Alvarez, B. *et al.* NNAlign_MA; MHC peptidome deconvolution for accurate MHC binding motif characterization and improved T-cell epitope predictions. *Molecular & Cellular Proteomics* **18**, 2459-2477 (2019).
77 Stranzl, T., Larsen, M. V., Lundegaard, C. & Nielsen, M. NetCTLpan: pan-specific MHC class I pathway epitope predictions. *Immunogenetics* **62**, 357-368 (2010).
78 Vang, Y. S. & Xie, X. HLA class I binding prediction via convolutional neural networks. *Bioinformatics* **33**, 2658-2665 (2017).
79 Nielsen, M. & Andreatta, M. NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. *Genome medicine* **8**, 1-9 (2016).
80 Han, Y. & Kim, D. Deep convolutional neural networks for pan-specific peptide-MHC class I binding prediction. *BMC bioinformatics* **18**, 1-9 (2017).
81 Singh, H. & Raghava, G. ProPred1: prediction of promiscuous MHC Class-I binding sites. *Bioinformatics* **19**, 1009-1014 (2003).
82 Shao, X. M. *et al.* High-throughput prediction of MHC class I and II neoantigens with MHCnuggets. *Cancer immunology research* **8**, 396-408 (2020).

Figure legends

Figure 1. The proposed TransPHLA model.

Figure 2. The workflow of automatically optimize mutated peptide (AOMP)

program for example peptide DLLPETPW and HLA-B*51:01. The number and the letter, for example 8I, indicate that the amino acid at the 8-th position of the peptide obtained at the previous level is replaced with the amino acid I.

Figure 3. The user input and output results of the web server for TransPHLA and AOMP program.

Figure 4. Comparison of the proposed TransPHLA method with 11 existing methods. For subfigures (a) and (c), matchable means that the data of methods in the subgraph is consistent, that is independent or external data. For subfigures (b) and (d), unmatchable means that the data of different methods in the subgraph is not the same, and they are different subsets of the provided data. For each method, the TransPHLA performs prediction and pairwise comparison on the corresponding subset. For subfigures (e) and (f), the predictable number represents the number of predictable peptide-HLA-I binders.

Figure 5. (a) Attention scores associated with all correctly predicted samples, correctly predicted positive samples and correctly predicted negative samples. (b) The contribution (i.e. accumulative attention score) of the amino acid types of peptides and peptide positions to peptide-HLA-I (pHLA) binding. (c) Accumulative attention scores for peptide binders associated with several well-characterized HLA-I alleles. Only 9-mer peptides are focused on here. The brighter residues are considered more important in pHLA binding.

Figure 6. The calculation process of masked multi-head self-attention

mechanism, where i represents the i-th head attention, h is the number of heads, len_X is the length of sequence S, d_X and d_K are the dimensions of X and K_i, and $\sqrt{d_K}$ is the scaled factor to prevent the large dot products.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterials.docx