ARTICLES
https://doi.org/10.1038/s42256-022-00459-7

nature machine intelligence

Check for updates

A transformer-based model to predict peptide–HLA class I binding and optimize mutated peptides for vaccine design

Yanyi Chu[1,2,9], Yan Zhang[3,9], Qiankun Wang[1], Lingfeng Zhang[4], Xuhong Wang[5], Yanjing Wang[1], Dennis Russell Salahub[2], Qin Xu[1], Jianmin Wang[6], Xue Jiang[1], Yi Xiong[1,7] ✉ and Dong-Qing Wei[1,8] ✉

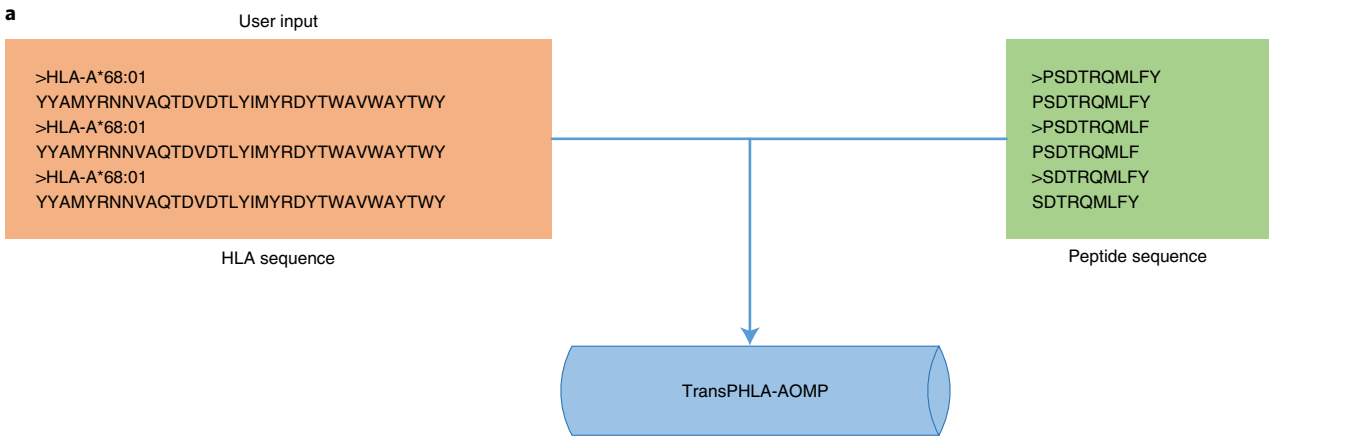
Human leukocyte antigen (HLA) can recognize and bind foreign peptides to present them to specialized immune cells, then initiate an immune response. Computational prediction of the peptide and HLA (pHLA) binding can speed up immunogenic peptide screening and facilitate vaccine design. However, there is a lack of an automatic program to optimize mutated peptides with higher affinity to the target HLA allele. Here, to fill this gap, we develop the TransMut framework—composed of TransPHLA for pHLA binding prediction and an automatically optimized mutated peptides (AOMP) program—which can be generalized to any binding and mutation task of biomolecules. First, TransPHLA is developed by constructing a transformer-based model to predict pHLA binding, which is superior to 14 previous methods on pHLA binding prediction and neoantigen and human papilloma virus vaccine identification. For vaccine design, the AOMP program is then developed by exploiting the attention scores generated by TransPHLA to automatically optimize mutated peptides with higher affinity to the target HLA allele and with high homology to the source peptide. The proposed framework may automatically generate potential peptide vaccines for experimentalists.

The binding of peptides with human leukocyte antigen (HLA) 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]. HLAs are generally divided into two categories: HLA class I (HLA-I) and HLA class II (HLA-II). HLA-I is encoded by three I loci and expressed on the surface of all nucleated cells, whereas HLA-II can only be expressed in professional antigen-presenting cells[3]. In this Article we focus on HLA-I molecules (hereafter referred to as HLA). HLA mainly binds short peptides with a length of 8–10 amino acids, because both ends of the binding groove are blocked by conserved tyrosine residues[4,5], of which 9-mer peptides are the most common[6]. Then, some of these pHLAs are presented on the cell surface for recognition by $CD8^+$ T cells[7,8]. Peptide binders with 11–14 amino acids have been identified[9,10]. Considering the comprehensive applicability of the method, peptides with lengths of 8–14 amino acids are included in this study.

Because HLA molecules are highly specific and polymorphic in the human population[11], only 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[3,12]. The first step towards this goal is to verify the affinity between peptides and HLA alleles. Given that the affinity between a peptide and its binding HLA allele is closely related to whether it can be presented, many in silico methods have been developed to predict the affinity between peptides and HLA alleles (Supplementary Section 1 summarizes the work related to this). Existing methods are mainly based on using machine learning models, especially neural networks, to predict the binding affinity between peptides and HLA alleles[13]. Although the accuracy is as high as 90% for peptides with nine amino acids[14], the prediction capabilities for peptides of other lengths are still not satisfactory[13]. This can be explained by the fact that the 9-mer peptides bind more easily with HLA alleles, as they have more pHLA binding data for training[15] than peptides of lengths 13 and 14. Moreover, both allele-specific and pan-specific models have been developed for pHLA binding prediction[16]. The former cannot be applied in HLA alleles or for peptide lengths that do not exist in the training data, whereas the latter are trained on multi-allele data, which can accurately predict pHLA binding, especially for rare HLAs and peptide lengths[16].

It is attractive to synthesize short peptides to elicit highly targeted immune responses. Understanding the interactions of pHLAs can facilitate peptide vaccine design[17] and play an important role in the development of candidate vaccines for various diseases[18,19]. Several studies[20,21] have demonstrated that neoantigens produced by non-synonymous mutations in cancer cells play a key role in the anti-tumour immune response. Moreover, vaccines for neoantigens have proven to be beneficial to clinical outcomes[22,23].

[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, P. R. China. [2]Department of Chemistry, CMS (Centre for Molecular Simulation) and IQST (Institute for Quantum Science and Technology), University of Calgary, Calgary, Alberta, Canada. [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, Ottawa, Ontario, Canada. [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, P. R. China. [6]Integrative Biotechnology & Translational Medicine, Yonsei University, Incheon, Republic of Korea. [7]Shanghai Artificial Intelligence Laboratory, Shanghai, P.R. China. [8]Peng Cheng Laboratory, Shenzhen, Guangdong, P. R. China. [9]These authors contributed equally: Yanyi Chu, Yan Zhang. ✉e-mail: xiongyi@sjtu.edu.cn; dqwei@sjtu.edu.cn

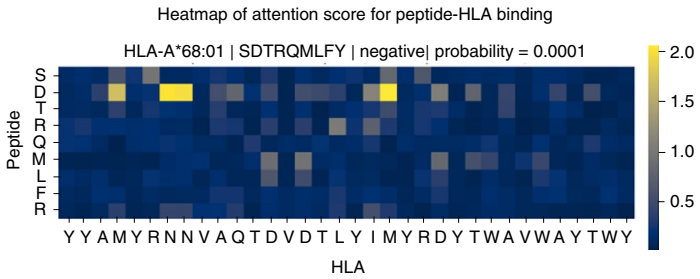


b

TransPHLA for peptide-HLA binding

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	Mutated 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

Fig. 1 | TransPHLA and the AOMP program. a–**c**, The workflow of the proposed TransPHLA and AOMP program, including the user input (**a**) and the output results (**b**,**c**) of the freely available webserver.

Peptide vaccines have many advantages over traditional vaccines[18,24]. The principle of peptide vaccines design is that antigen peptides bind to a specific HLA to form peptide–HLA–TCR complexes to elicit T-cell immune responses[25]. Theoretically, the antigen peptide should selectively bind to a specific HLA allele with high affinity. The process of identifying neoantigens is as follows[13]. First, high-throughput sequencing technologies and bioinformatics pipelines are established to characterize the non-synonymous mutations of the primary tumour, then computational methods are developed to reliably predict the binding probability of the mutant peptide and the HLA allele[26]. With these two stages, the number of candidate mutant peptides can be reduced greatly, thus speeding up the process of experimental validation[27,28]. However, the above-mentioned process is relatively complicated. Therefore, the development of an automatically optimized mutated peptides (AOMP) program would represent a huge breakthrough in the neoantigen design field.

In this Article we describe the design of a transformer-based model[29] for pHLA binding prediction (TransPHLA) and the AOMP program for mutant peptide optimization (Fig. 1 shows the entire workflow). TransPHLA is a pan-specific method[16] that achieves improved performance and can be applied to rare and unseen HLA alleles (Fig. 2). The core idea of the TransPHLA model is to apply self-attention[29] to peptides, HLAs and pHLA pairs to obtain the binding score. Some techniques are used to construct and optimize the model, which consists of four major sub-modules: (1) the embedding block (besides the encoding of amino acids in the sequence, we added positional embedding to describe the position information of the sequence); (2) the encoder block (multiple self-attentions are applied to focus on different components of the sequences, and padding positions of the sequence are masked to prevent misleading the model); (3) the 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 achieve better feature representation); (4) the projection block (multiple fully connected layers are used to predict the final pHLA binding score). The proposed TransPHLA model was compared to 14 previous pHLA binding prediction methods, including the state-of-the-art method[30], the Immune Epitope Database (IEDB) recommended method[14], nine IEDB baseline methods[14,15,31–37] and three recent attention-based methods[38–40]. TransPHLA not only achieves better performance with higher efficiency, but also solves the limitations of many methods with HLA alleles and peptides with variable lengths. We also conducted two types of case study to demonstrate the usability and validity of the TransPHLA method. TransPHLA shows better performance than 14 previously published methods for neoantigen identification[41,42]

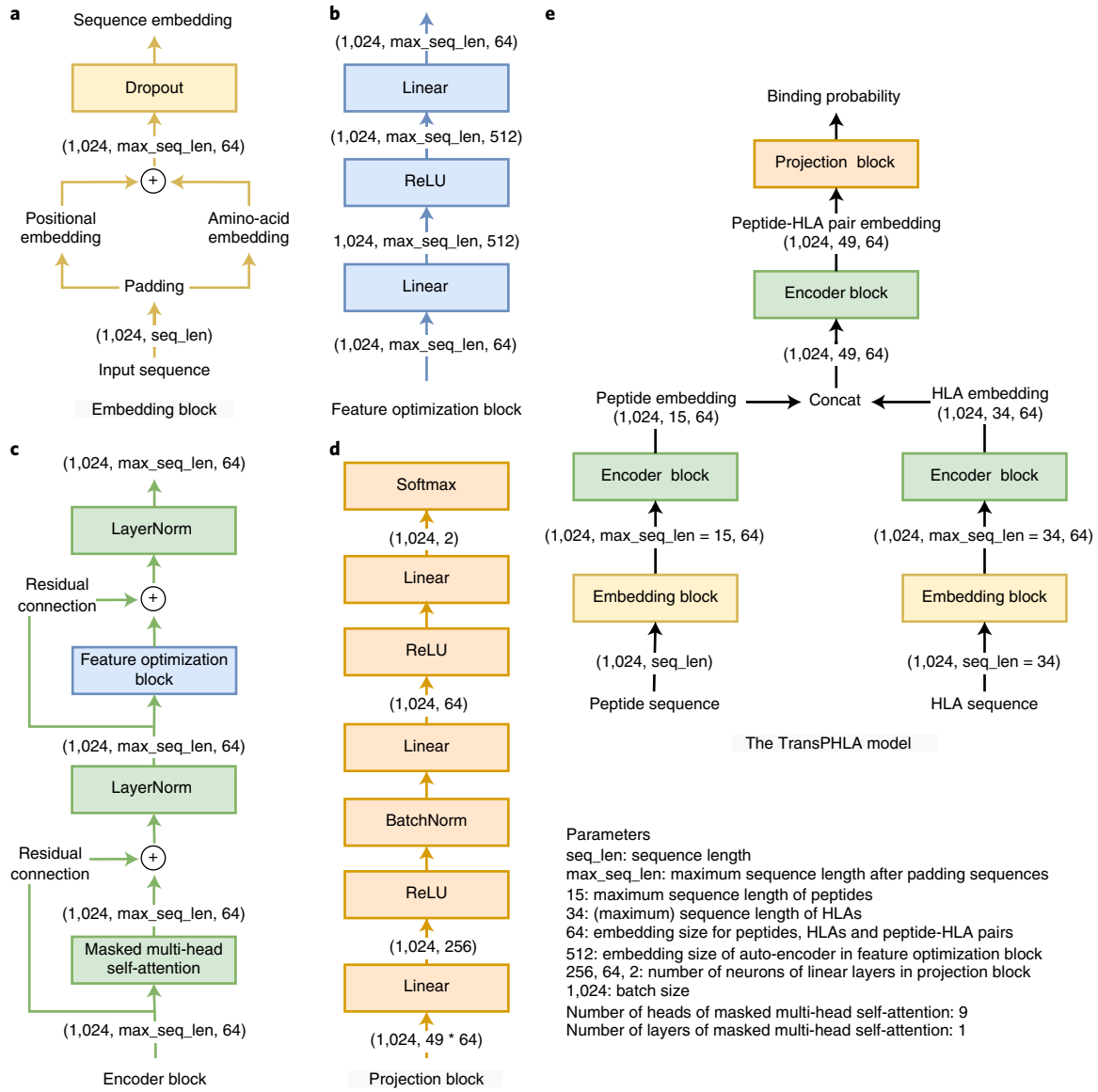


Fig. 2 | Sub-modules of the proposed TransPHLA model. a–e, The proposed TransPHLA model (**e**) is composed of four major sub-modules (**a–d**).

and achieves a positive screening rate of 96%. Although the positive screening rate is not very high for human papilloma virus (HPV) vaccine identification[43] due to the inconsistent threshold, it is superior to the other 14 methods.

We also develop an AOMP program (Fig. 3) for peptide vaccine design based on the attention mechanism, obtained by TransPHLA. When the user provides a pair comprising a source peptide and a target HLA allele, the AOMP program can search for mutant peptides with higher affinity for the target HLA allele and no more than four mutation positions. This program not only guarantees the affinity between the mutant peptide and the target HLA allele, but also ensures the homology of the mutant peptide and the source peptide to trigger cross-immunization. We tested all 366 combinations of the different HLAs and peptide binder lengths using two strategies. The first strategy randomly selects ten negative pHLAs correctly predicted by TransPHLA for each combination, and a total of 3,660 true negative pHLAs are selected. The other strategy only considers the negative pHLAs predicted by TransPHLA and does not consider the ground-truth label. With the two strategies, the 3,633 and 3,635 source peptides successfully found the optimized mutant peptide binding to HLA alleles, and 93.4% and 93.7% of them were verified by the method recommended by IEDB[14], confirming the usability of our program. Furthermore, 88.8% of 3,633 and 89.5% of 3,635 optimized mutant peptides have homology of more than 80% (1–2 mutated sites) with their source peptides, which is promising for vaccine design.

The TransPHLA and AOMP program jointly form the TransMut framework, which applies the transformer to the field of biomolecular binding and mutations. This framework can be applied to any biomolecular mutation task, such as epitope optimization[44] or drug design[45], and is useful for vaccine development in particular. For example, the tumour-necrosis factor-α (TNF-α) targeted vaccine, because of the biological activity of TNF-α, will cause inflammation in the body, and long-term medication holds the risk of causing autoimmune disease[46]. The core problem of TNF-α vaccine development is how to reduce the biological activity of TNF-α while maintaining sufficient immunogenicity[47]. The AOMP program is suited to this task. The transformer-derived model is first deployed to train the mutation direction data of the biomolecules, then the attention score in the mutation direction is obtained. Based on the attention score, the AOMP program will find a better mutant.

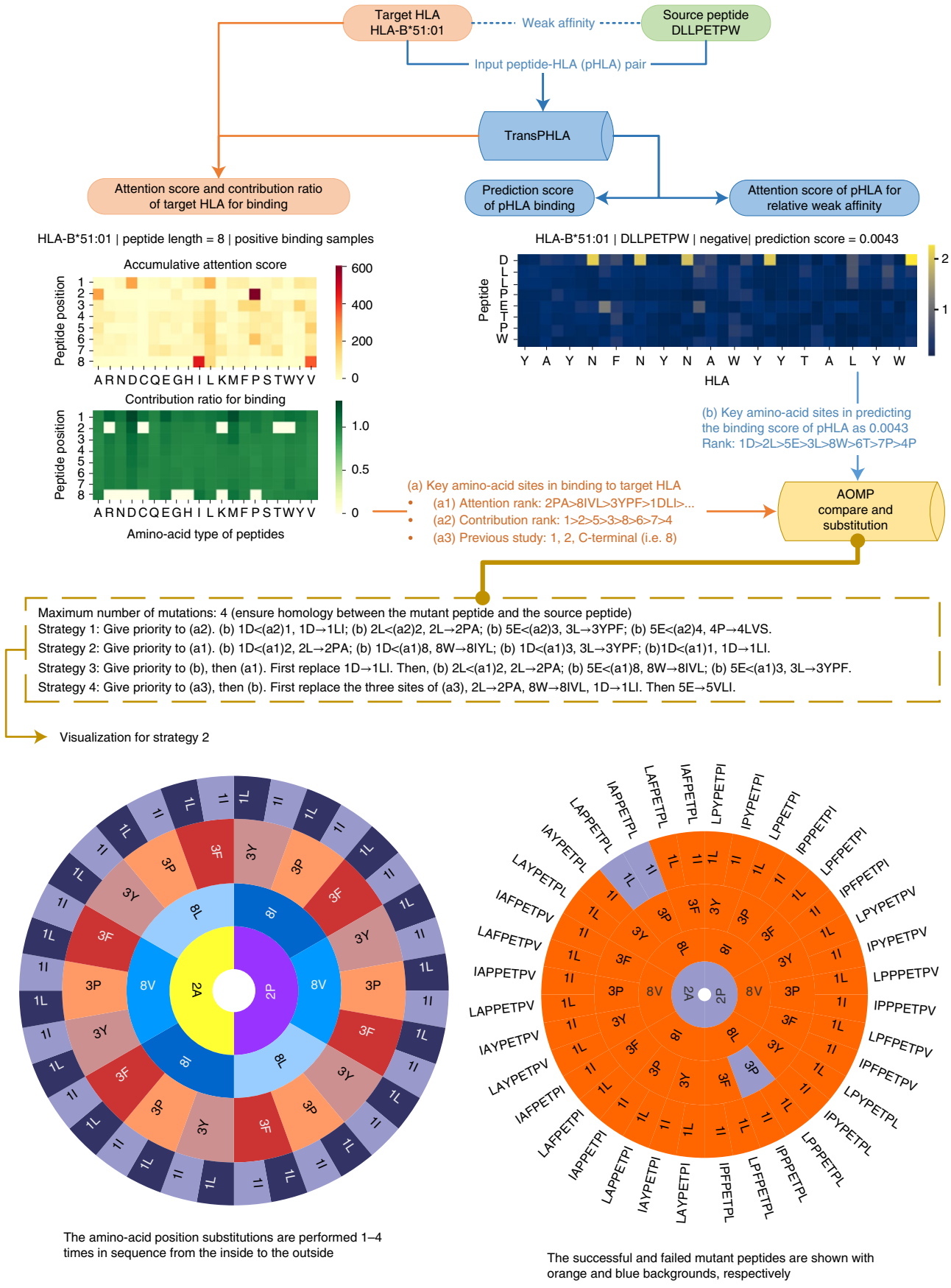


Fig. 3 | Workflow of the AOMP program. The workflow of the AOMP program for example peptide DLLPETPW and target HLA HLA-B*51:01. The number and letter—for example, 8I—indicate that the amino acid at the eighth position of the peptide obtained at the previous level is replaced with amino acid I.

Results

Comparison of TransPHLA with existing methods. To verify the effectiveness of TransPHLA, we compared it with nine baseline methods from IEDB, the recommended method from IEDB (NetMHCpan_EL[14]), the state-of-the-art method published in 2021 (Anthem[30]) and three attention-based methods published recently (ACME[38], DeepNetBim[40] and DeepAttentionPan[39]). The baseline methods are ANN[15], Consensus[34], NetMHCcons[35], NetMHCpan_BA[14], NetMHCstabpan[37], PickPocket[36], CombLib[33], SMM[31] and SMMPMBEC[32], which can be obtained from http://tools.iedb.org/main/tools-api/. The different methods use different scoring methods to determine whether pHLA can bind, such as the predicted half-maximum inhibitory concentration (IC_{50}), predicted score and percentile rank. We used the predicted IC_{50} and predicted score as the criteria for the regression and classification tasks, respectively (Consensus only provides percentile rank as the criterion). Supplementary Table 1 lists details of the criteria strategies for the different methods[14,30,34,48].

It is worth noting that not every method is compatible with every HLA allele and peptide of every length. Except for NetMHCpan_BA, NetMHCpan_EL and our method, the methods have different limitations. For example, SMM and SMMPMBEC only support peptides with lengths in the range of 8–11, and DeepNetBim and CombLib only support peptides with a fixed length of 9. In summary, with the same data, not every method can predict all the samples provided by the user.

The comparison was performed on a pHLA independent test, a pHLA external test, neoantigen identification and HPV vaccine identification (Fig. 4).

Figure 4 reveals two perspectives on the pHLA test set: (1) the methods can predict all the provided data (Fig. 4a,b, matchable) or (2) the methods can only predict part of the provided data as a result of their limitations (Fig. 4c,d, unmatchable). In Fig. 4a,b, the data used for the performance comparison of the different methods are all consistent, so the prediction performance can be compared fairly. In Fig. 4c,d, the HLA alleles and peptide lengths that can be predicted by the methods differ. Therefore, for each method in these subfigures, the data used for performance comparison are a subset of the provided data. To make the performance comparison fairer and more reasonable, the proposed TransPHLA performs a pairwise comparison with each method on the corresponding subset data. On both independent and external data, the proposed method is superior to the other methods, except Anthem. Anthem shows slightly inferior performance than TransPHLA on the independent data and competitive performance on the external data. However, it cannot be extended to some unknown HLA alleles or peptide lengths because of its limited published data, whereas TransPHLA does not have this limitation. A more detailed comparison between TransPHLA and Anthem is presented in Supplementary Section 2.3. Moreover, although NetMHCpan_EL achieves good performance on external data, its performance on independent data is greatly reduced. The independent data contain 112 types of HLA alleles, whereas the external data contain only five HLA alleles. As we mentioned before, those two types of test data are complementary in the performance comparison of the methods, so only a method that works well on both types of data can demonstrate its superiority.

We also discuss the performance of each method for each peptide length on the independent and external data. Supplementary Figs. 1–8 present violin plots for the distributions of the area under the curve (AUC), accuracy, Matthews correlation coefficient (MCC) and F_1 for the 15 methods when used on the independent and external data. These results indicate the superiority of TransPHLA over the other 14 methods, as follows: (1) TransPHLA is not restricted by HLA allotype or peptide length; (2) for any peptide length, TransPHLA shows superior performance on all metrics; (3) TransPHLA shows a tight distribution on four metrics, especially for peptide length 9, reflecting the potential of TransPHLA to increase the performance as the amount of training data increases, and, if pHLA data of other peptide lengths or HLAs increase, TransPHLA also achieves better results; (4) the MCC results show that TransPHLA is effective for any HLAs of any length; (5) when performing predictions on ~170,000 pHLAs, TransPHLA requires 28 s on a GeForce RTX 3080 GPU and 2 min on the CPU (the other methods are not as fast). Supplementary Sections 2.1 and 2.2 provided a detailed analysis of the results.

The primary determinant of neoantigen screening is the binding of a peptide and an autologous specific HLA molecule[49]. For neoantigen identification, we collected neoantigen data from non-small-cell lung cancer, melanoma, ovarian cancer and pancreatic cancer from recent works[41,42], including 221 experimentally verified pHLA binders. The comparison results for the different methods on these data are shown in Fig. 4e. These show that TransPHLA was able to screen out 96.4% of neoantigens. Although CombLib achieved 100% accuracy, it only supports 9-mer peptides, which limits its application. The remaining ten methods have lower performance than TransPHLA and may be limited by predictable HLAs or peptide lengths.

The 221 neoantigen samples consist of 62 combinations of HLA alleles and peptide lengths. Among these, ten samples of eight combinations are not included in the training data. In these ten samples, TransPHLA only mispredicts three samples, indicating the generalization ability of TransPHLA.

HPV is the most common sexually transmitted disease[50] and there are some preventive HPV vaccines. However, the therapeutic effect of these vaccines is limited and the use rate very low[51]. It is thus critical to develop therapeutic vaccines to treat HPV infections and diseases. A previous study[43] presented 278 experimentally verified pHLA binders from HPV16 proteins E6 and E7, consisting of 8–11-mer peptides. The comparison results for use of the different methods on these data are shown in Fig. 4f. Although TransPHLA only shows a screening rate of 68%, it still achieves higher performance than the other methods.

According to the source reference[43] for the HPV vaccine data, the data are identified as 'binder' according to $IC_{50} < 100\,\mu M$, which is 200 times the common threshold of 500 nM. The value of 500 nM is the threshold for the data used for the 15 prediction methods. Thus, peptides with IC_{50} values over 500 nM are negative samples in these prediction methods. This is the reason why the HPV vaccine data show poorer performance than other datasets.

We also evaluated the performances of the methods on samples with $IC_{50} \leq 500\,nM$. The results are shown in Extended Data Fig. 1 and Supplementary Section 10. Based on the results, TransPHLA only mispredicts three samples (that is, a total of 18 samples), and achieves performance superior to those of the other 14 methods.

TransPHLA uncovers the underlying patterns of pHLA binding. The attention mechanism of TransPHLA provides biological interpretability for the model. In this section, we explore the binding rules of pHLA by means of the attention scores. The evidence shows that the C-terminal, N-terminal and anchor sites[52] of the peptide are critical for binding to HLA and are always located at the first, last and second positions of the peptide sequence. The attention scores of these positions were confirmed, as shown in Fig. 5a.

We next analysed the contributions of the amino-acid types on the positive and negative samples to binding and non-binding at different peptide positions (Fig. 5b). It was found that the binding and non-binding of pHLAs are affected by different components of the peptides. In addition, we analysed 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

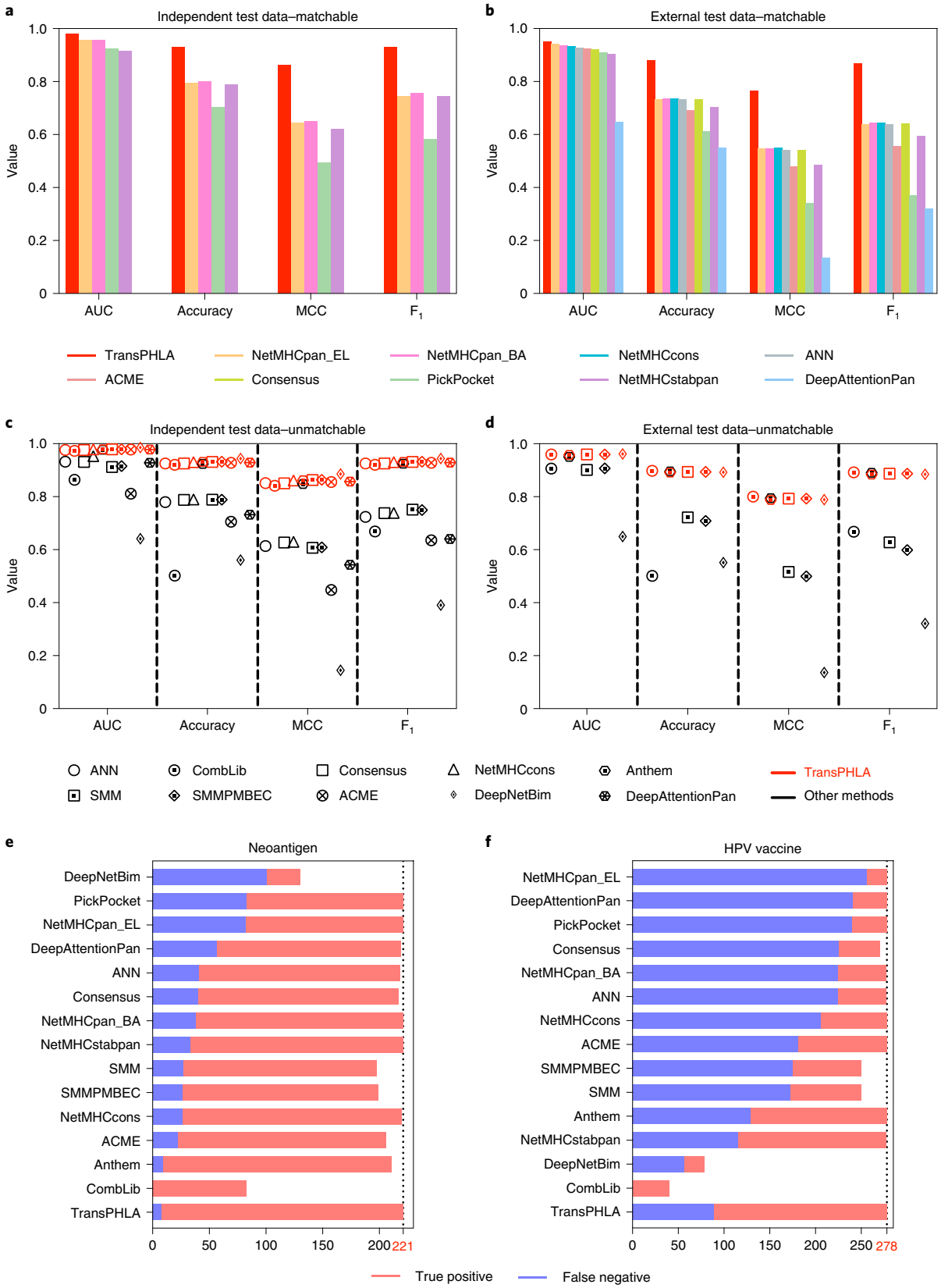


Fig. 4 | Comparison of the proposed TransPHLA method with 14 existing methods. a–f, Comparison of the methods on a pHLA independent test (**a**,**c**), a pHLA external test (**b**,**d**), neoantigen identification (**e**) and HPV vaccine identification (**f**). In **a** and **b**, matchable means that the data for the methods in the graphs are consistent (that is, independent or external data). In **c** and **d**, unmatchable means that the data for the different methods in the graphs are not the same, indicating different subsets of the data. For each method, the TransPHLA performs prediction and pairwise comparison on the corresponding subset. For **e** and **f**, the predictable number of pHLA binders is the sum of true positive and false negative.

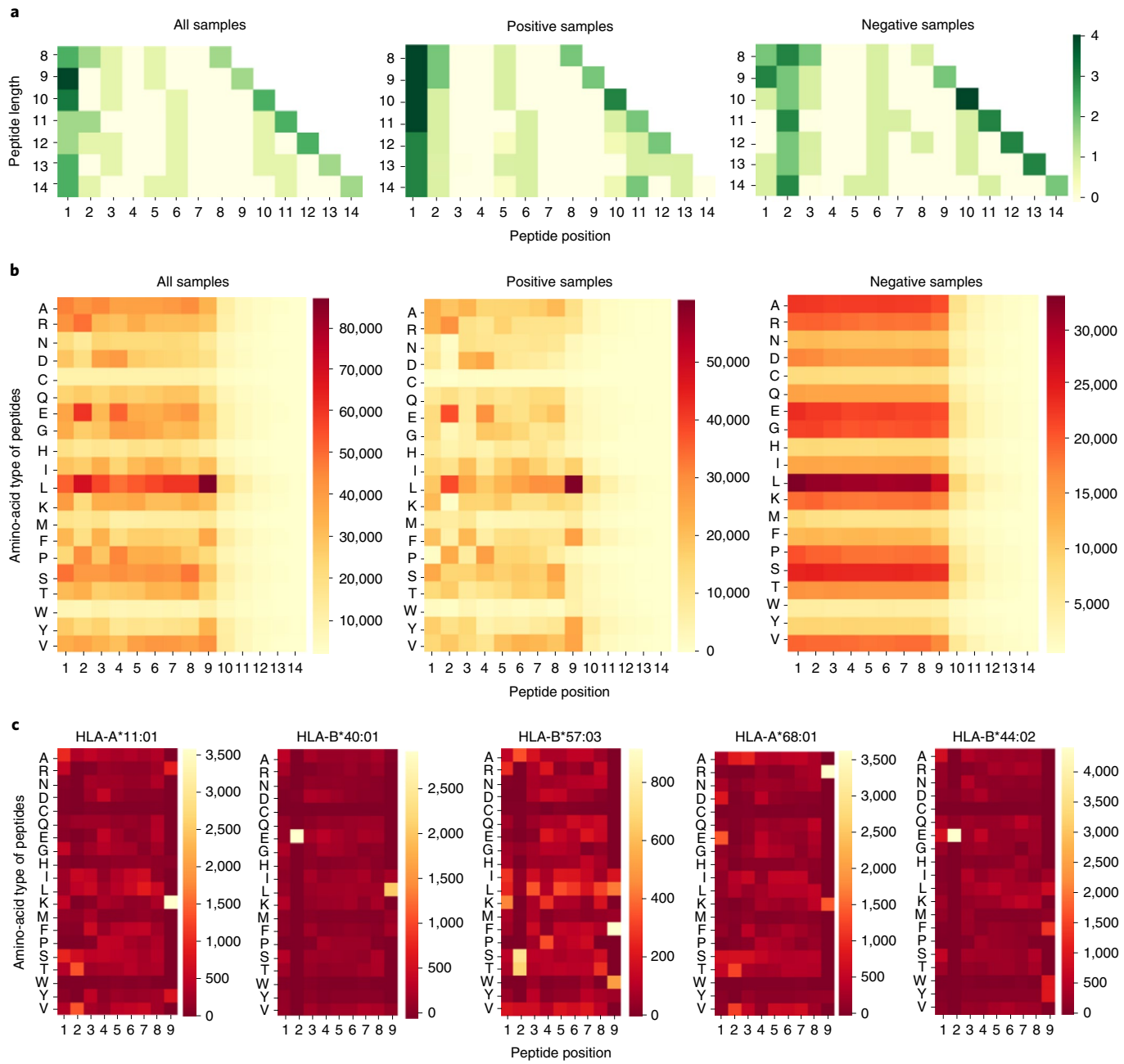


Fig. 5 | Attention scores. a, Heatmap of attention scores associated with all correctly predicted samples, correctly predicted positive samples and correctly predicted negative samples. **b**, The contribution (that is, accumulative attention score) of the amino-acid types of peptides and peptide positions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

pHLA binding, but can also be used for vaccine design, as shown in the sections AOMP program in the Results and AOMP program in the Methods.

In addition, because the attention score represents the pattern of pHLA binding, it implies that the key amino-acid sites on the peptide sequence are important for binding or non-binding to the target HLA. We thus visualized the binding pattern of five HLA alleles according to ACME[38] (Fig. 5c). As expected, TransPHLA found a similar pattern for amino-acid types at different peptide positions to the previous studies[38,53]. For HLA-A*11:01, TransPHLA recognizes the anchor residue for the peptides with K (Lys) at position 9 (ninth K). For HLA-B*40:01, the key residues—the second E (Glu) and ninth 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 the ninth L, ninth F (Phe) and ninth W (Trp), which is consistent with the structures in PDB 2BVP[54]. For HLA-A*68:01, 4HWZ[55] demonstrates that the ninth K and ninth R (Arg) residues of the peptide greatly contribute to the binding. For HLA-B*44:02, the key role of the second E has been proved by 1M6O[56]. All these results have been supported by previous studies and demonstrate the effectiveness of our methods.

AOMP program. It is proposed to search for mutant peptides with higher affinity if the source peptide under consideration has weak binding affinity with its specific HLA allele. Figure 3 visualizes the process of AOMP and the automatic mutation of the sec-

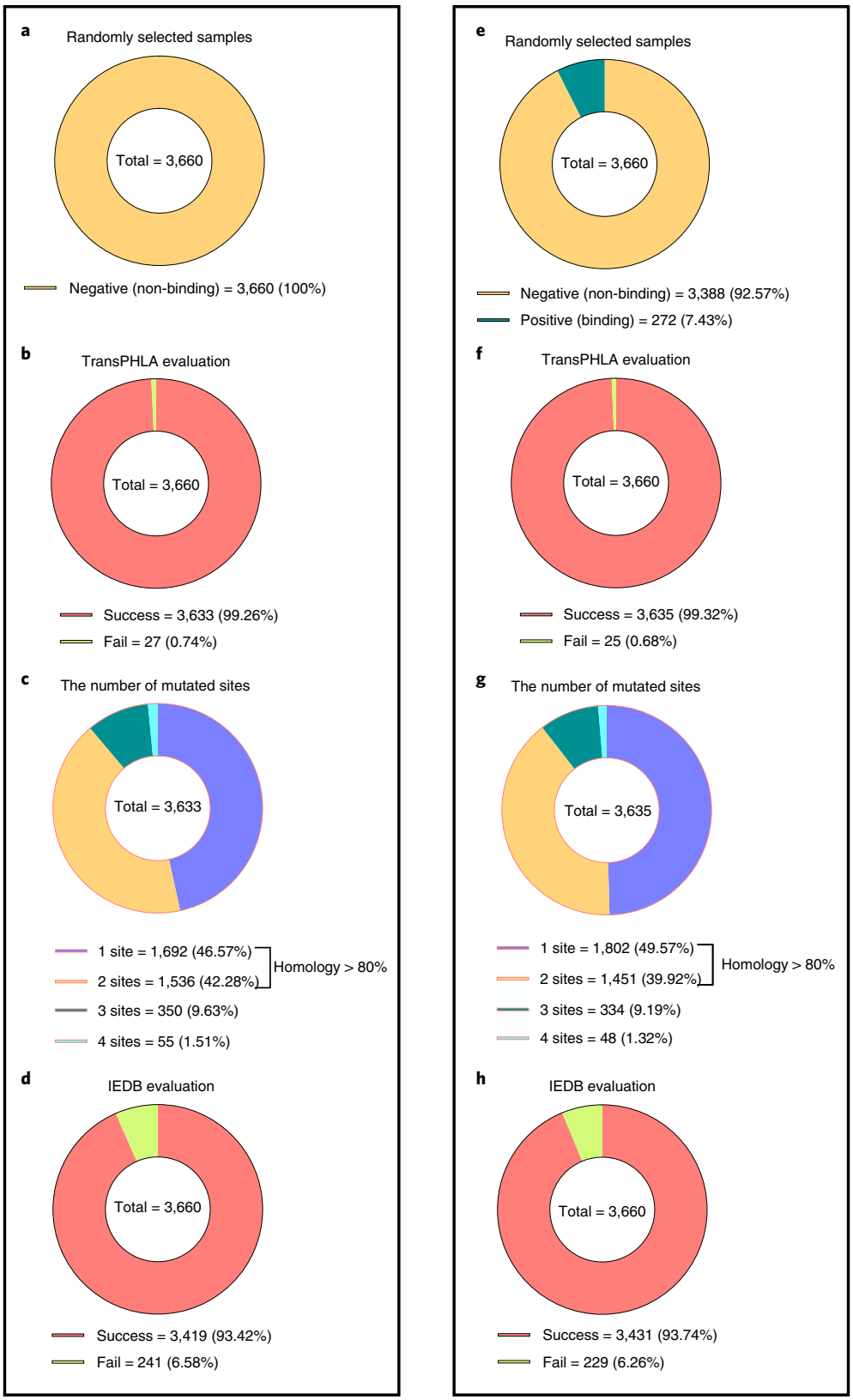


Fig. 6 | Summary of the two random selection strategies of negative pHLAs for AOMP evaluation. a–**h**, Randomly selected negative samples correctly predicted by TransPHLA (**a**–**d**) and randomly selected negative samples predicted by TransPHLA (**e**–**h**): the randomly selected samples (**a**,**e**); TransPHLA evaluation (**b**,**f**); the number of mutated sites (**c**,**g**); IEDB evaluation (**d**,**h**).

ond strategy for the example of source peptide DLLPETPW and target HLA-B*51:01.

To demonstrate the effectiveness of the AOMP program we proposed two strategies for testing all 366 HLA–peptide length combinations in this study. The first strategy selects the non-binding pHLAs correctly predicted by TransPHLA; that is, both the ground-truth-labelled and prediction results are non-binding. For the second strategy, only the prediction results of TransPHLA are considered, and the ground-truth label is not considered. In short, the evaluation samples are selected from the non-binding pHLAs predicted by TransPHLA. After random selection, the proportion of true negative samples is 92.57% in the second strategy (Fig. 6e). The AOMP program was then used to search for mutant peptides for 3,660 negative pHLAs with the two strategies.

To verify the authenticity and usability of the mutation results, we used the NetMHCpan_BA[14] recommended by IEDB to validate the mutation results for 3,660 pHLAs under the two strategies. The results are shown in Fig. 6d,h, showing success rates of 93.42% and 93.74% with the two strategies, respectively.

The second strategy shows slightly better performance than the first, because the evaluation samples of the second strategy contain binding pHLAs, and AOMP can more easily generate binding mutation pHLAs for them. The first strategy can more accurately evaluate the probability of successful mutation of AOMP for the non-binding pHLAs, whereas the second strategy can better reveal the successful mutation rate of AOMP in actual situations, because the ground-truth label is unknown in practice.

We also used molecular dynamics (MD) simulations to verify the effectiveness of AOMP. We used HLA-A*02:01 as the target HLA and YKLVVVGAG as the source peptide. Eight mutated peptides were chosen for the simulations and compared with the source peptide. According to the results, (1) the attention mechanism obtained by the proposed TransPHLA is consistent with the structure of the pHLA complex and (2) the prediction results of TransPHLA are consistent with the results of the MD simulation and NetMHCpan_BA. On the other hand, some mutated peptides produced by AOMP have been experimentally verified that can bind to the corresponding HLA allele. More details are provided in Supplementary Section 11.

Discussion

pHLA binding and interaction are critical to epitope presentation and a prerequisite for the T-cell recognition that initiates an effective immune response. As a first step, epitope screening and identification depend on the affinity of pHLA, especially in the neoepitope-based immunotherapy that is recognized as the most promising cancer treatment. The primary determinant of neoantigen screening is the affinity of peptides and specific autologous HLA molecules. Accurate pHLA binding prediction is thus essential for the identification of immunotherapy targets, epitope screening and vaccine design. Peptide vaccine design is another important field for the treatment of diseases. However, the current vaccine design method is in its infancy and cannot yet be automated.

First, we have proposed a TransPHLA method for pHLA binding prediction based on the transformer model, which is a generalized pan-specific model that is not restricted by HLA alleles or peptide length. We conducted two types of independent test and two types of case study (neoantigen and HPV vaccine identification). Compared with the state-of-the-art method (Anthem), the IEDB recommended method (NetMHCpan_EL), nine IEDB baseline methods and three attention-based methods published recently, TransPHLA achieves superior performance for all four experiments.

Based on TransPHLA, we have also developed an AOMP program by using the attention scores generated by TransPHLA to search for mutant peptides with higher affinity to the target HLA allele and high homology with the source peptide. For two evaluation strategies for the AOMP program, among 7,320 pHLAs for different HLA alleles and peptide lengths, 7,268 samples were successfully found for the binding mutant peptide–HLA; 94% were verified by the method recommended by IEDB, and 89% with a homology of more than 80%, which is useful for vaccine design.

This is the first attempt to propose a transformer-based TransMut framework in the field of automatic mutation of biomolecules that has the potential to be applied to other binding prediction and mutation tasks for biomolecules.

Methods

Dataset. In this study, the pHLA binding data (positive data) were obtained from Anthem[30], which can be downloaded from https://github.com/17shutao/Anthem/tree/master/Dataset. The negative data were generated in a similar way to previous studies[13,14,57]. 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 immunopeptidomes. Although false negative peptides may be generated, the possibility and proportion of such peptides are very low[1,58] and can be ignored. This strategy of constructing negative samples guarantees that the dataset is balanced (Supplementary Table 2).

To fairly compare our method with previous methods, we followed the training and evaluation strategy of Anthem[30], which is the state-of-the-art pHLA binding prediction method. There were three types of dataset with 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 data sources for the training and independent test set are the same: (1) four public HLA binders databases (IEDB[59], EPIMHC[60], MHCBN[61] and SYFPEITHI[62]), (2) allotype-specific HLA ligands identified by mass spectrometry in previously published studies[63–78] and (3) peptide binders from training datasets of other pHLA binding prediction tools[38,48,59,79–89]. The external test set was experimentally verified by Anthem[30].

We also checked and deleted some error or duplicate samples; for example, 'HLA-B*07:01'-related samples are ignored because its sequence contains errors. The statistics of the three types of dataset are listed in Supplementary Table 2. The number of pHLA binders for each peptide length of each HLA allele spans a large range, from 10^1 to 10^5 (for details see Supplementary Fig. 12). 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 Extended Data Fig. 2 the number of 9-mer peptides is very large, whereas there are very few 13- and 14-mer peptides. This leads to differences in the performance of the method for different peptide binder lengths (Extended Data Fig. 2).

Experiment settings. To follow previous studies[13,30] for pHLA binding prediction, we conducted fivefold cross-validation (CV) and independent testing. Because the source of the independent test set and the training set are the same, the data distributions for the training set and independent test set are very similar (Supplementary Figs. 11 and 12). When the model is tested on data with a similar distribution to the training data, it is easier to obtain a better test performance than on a model that is trained with a different distribution to the test data. In other words, our proposed method and Anthem[30] may have an advantage over the other methods on the independent test set. We therefore set up an external test to perform a fairer comparison of the different methods.

The fivefold CV was used in this study for model evaluation to optimize the model at the training stage. It divides the training set into five equal parts, 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 five times to ensure that each part of the data participates four times for model training and once for model evaluation. Finally, the average result of the five model evaluations is used as the final evaluation result. Usually, the use of CV can avoid, to a certain extent, overfitting of the model.

The independent test is a popular strategy to evaluate the generalization ability of the considered method for unseen data. Independent test data does not have any overlap with training data, but follows the same distribution as the training dataset. It also provides common data independent from the training data so as to fairly evaluate the performance of different methods.

To enable a fair comparison, we used experimental data as the external test data to eliminate possible deviations as a result of there being the same data distribution. According to Supplementary Figs. 11 and 12, 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, the following metrics were calculated:

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

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

$$\mathrm{F_1\ score} = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}},\ \text{where Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \text{ and Recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \quad (3)$$

where TP is true positive, FP is false positive, FN is false negative and TN is true negative. In addition, we adopt AUC, that is, the area under the receiver operating characteristic curve, as the other performance evaluation metric.

Other than MCC, which ranges from −1 to 1, the other metrics range from 0 to 1. The higher the value of the metric, the better the model or method. It is worthwhile noting that MCC cannot be calculated when two of TN, TP, FN, FP are 0, because the denominator is 0. 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 allele, this implies that the method is invalid for this HLA allele with this peptide length.

TransPHLA. The core idea of TransPHLA is the application of the self-attention mechanism[29]. TransPHLA is composed of the following four blocks (Fig. 2). The embedding block adds positional embedding to the amino-acid embedding to generate the sequence embedding, and then applies a dropout technology to enhance the robustness. Through the embedding block, TransPHLA generates the embeddings for peptides and HLA alleles, respectively. Next, these embeddings are taken as input into the encoder block, 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 improves the feature representation obtained by the attention mechanism, mainly because more layers are added. The output feature representations of the peptide and HLA allele are then concatenated as the embedding of a pHLA pair. After pHLA pair embedding passes through the encoder block, the projection block is used to predict the pHLA binding score.

Model training is conducted on the CentOS Linux release 7.7.1908 (Core) system. The CPU is an Intel(R) Xeon(R) Gold 6230 CPU @ 2.10 GHz, with 80 logical CPUs. The GPU is a GeForce RTX 3080. The memory is 92G. The model is trained on the GPU, the code language is Python 3.7.8, and the model is built using PyTorch 1.7.0. The training consists of 50 epochs, with each epoch lasting 72 s. Among the 50 epochs, the model with the best performance on the fivefold CV is the final model. In the code environment (for example, random, numpy and torch), the random seed is set to 19,961,231.

Sequence embedding in TransPHLA. First, the peptide and HLA allele sequences are padded to the maximum length of 15 and 34, respectively, to handle the variable input length. The character embedding model is then used to create a unique embedding for each amino acid, with the dimension of the embedding defined as d_X. Taking the peptide SDKYGLGY as an example, it has a length of 8. From Supplementary Fig. 13a, embeddings of six 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 and function of the peptide and HLA allele sequence, but the above embedding method does not consider it. We thus apply positional embedding to encode the position of the amino acid in the sequence. Given the position p in the sequence, the positional embedding encoded as a d_X-dimensional vector, and the value of the ith element of this vector being $PE(p)_i$, then

$$\mathrm{PE}(p)_{2i} = \sin\left(p/10,000^{2i/d_X}\right) \quad (4)$$

$$\mathrm{PE}(p)_{2i+1} = \cos\left(p/10,000^{2i/d_X}\right) \quad (5)$$

where $2i$ represents the even dimensions and $2i+1$ the odd ones. This position embedding method can reflect not only the absolute position information of the amino acid but also the relative position information. We visualize positional embedding in Supplementary Fig. 13b. It is worth noting that, for any peptide or HLA allele, positional embedding is the same. We also conducted the ablation experiment for positional embedding and demonstrated its validity for TransPHLA (more details are provided in Supplementary Section 5).

Finally, the amino-acid embedding and positional embedding are summed to obtain the sequence embedding (shown in Supplementary Fig. 13c).

Masked multi-head self-attention mechanism in TransPHLA. The attention mechanism is the core of the transformer. It can 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 obtaining an output, where K-V pairs are the form of storing sequence elements in memory. This reflects the attention score (that is, the weight) according to the correlation or similarity of Q and K. The attention score represents the importance of information (that is, V). The larger the attention score, the more focused the corresponding information.

Compared with recurrent neural networks (RNNs), transformer realizes parallelization and solves the long-term dependencies problem, so it can process the data faster than RNNs. Compared with convolutional neural networks (CNNs), which extract local information commendably, transformer extracts more global information, which is suitable for the information exploration of the whole sequence of peptides and HLA alleles. In experiments (Supplementary Section 9), transformer has better performance than RNNs and CNNs as the encoder block in TransPHLA.

The self-attention mechanism belongs to a variant of the attention mechanism that captures the internal correlation of a sequence and reduces the dependence on external information. It is worth noting that this study introduced the mask operation when calculating the attention. For peptide or HLA allele equences with lengths less than the corresponding maximum length, non-amino-acid characters should not be considered for the model training. We thus use 10^{-9}, which is very close to zero, as their attention scores, so that non-amino-acid characters do not play a role in calculating the attention. The calculation process for the self-attention mechanism is shown in Extended Data Fig. 3 and Supplementary Section 6.

Model selection is carried out on the layer and head of the multi-head attention mechanism, and the final parameters are the attention of one layer and nine heads. The results indicate that our model is not overfitting (as shown in Supplementary Fig. 16 and Supplementary Section 7).

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

The program designed four directed mutation strategies based on the attention score obtained by TransPHLA (Fig. 3). The attention score not only represents the pattern of pHLA binding, but also reveals the key amino-acid sites on the peptide sequence that are important for binding or non-binding to the target HLA allele. 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 calculated with the difflib module in Python is very close to the blast result. The homologies of one, two, three and four amino-acid positions were mutated on average 90%, 80%, 70% and 61%, respectively. Therefore, we limited the number of mutations in the amino-acid site of the source peptide to no more than four.

For each of the 366 HLA–peptide length combinations, we established a binding contribution matrix of 20 amino acids at each peptide position. To adapt to a new or unknown HLA–peptide length combination, a general binding contribution matrix is established. We provide these 367 contribution matrices and their visual heatmaps on the webserver. On the other hand, when predicting a 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 calculate two contribution rate matrices based on the above two contribution matrices. The larger the element value in the contribution matrix, the more critical the corresponding amino-acid site for binding or non-binding. Intuitively, because the amino-acid site contributes more to non-binding prediction, if we replace them with other amino acids that contribute more to binding prediction, the mutated peptide is more likely to have a higher affinity with the target HLA allele. Based on the above four matrices, we designed four strategies to generate mutant peptides. The 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 greatly to the high affinity. The corresponding amino-acid substitutions are then made according to the comparison results. The process is as follows: (1) predict the binding score for the source peptide and target HLA; (2) find some of the most important amino-acid sites based on the self-attention mechanism; (3) replace these important sites of a weak-affinity pHLA with some amino acids that may contribute more to binding prediction; (4) select some of the best mutation candidates for evaluation.

For the source peptide and the target HLA allele (the specific pHLA), the mutant peptides generated by the four strategies are merged and the duplicates removed. TransPHLA then screens and retains mutant peptides that can bind to the target HLA allele. Excitingly, the original target of this program was non-binding pHLA, but we found that it can also find mutant peptides with stronger affinity for binding pHLA.

Figure 3 visualizes the process of the AOMP program and shows the automatic mutation of the second strategy for the source peptide DLLPETPW and target HLA-B*51:01 as an example. Supplementary Section 8 describes, in detail, the implementation process for the four AOMP strategies in this example. Supplementary Section 11 describes some AOMP instances according to experimentally verified literature and MD simulations.

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

Data availability

The datasets are available at https://github.com/a96123155/TransPHLA-AOMP/tree/master/Dataset, which contains the training data, independent test data, external test data, neoantigen data and HPV vaccine data. The statistics of these data are provided in Supplementary Section 3. In addition, the attention scores and heatmaps of amino-acid types and position of peptides for specific HLA alleles and peptide binder lengths can be downloaded from https://issubmission.sjtu.edu.cn/TransPHLA-AOMP/download.html. Source data are provided with this paper.

Code availability

The code is freely available at https://github.com/a96123155/TransPHLA-AOMP with GNU General Public Licence Version 3. This web page contains the code dependencies, operating environment, instructions and some interaction between code and results (file with ipynb suffix). The DOI is https://doi.org/10.5281/zenodo.5715479.

Received: 5 August 2021; Accepted: 14 February 2022;
Published online: 23 March 2022

References

1. Yewdell, J. W. & Bennink, J. R. Immunodominance in major histocompatibility complex class I—restricted T lymphocyte responses. *Annu. Rev. Immunol.* **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. *Nat. Immunol.* **8**, 1041–1048 (2007).
4. Bouvier, M. & Wiley, D. C. Importance of peptide amino and carboxyl termini to the stability of MHC class I molecules. *Science* **265**, 398–402 (1994).
5. Zacharias, M. & Springer, S. Conformational flexibility of the MHC class I α1-α2 domain in peptide bound and free states: a molecular dynamics simulation study. *Biophys. J.* **87**, 2203–2214 (2004).
6. 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. *Proc. Natl Acad. Sci. USA* **102**, 17107–17112 (2005).
7. Kloetzel, P. M. Generation of major histocompatibility complex class I antigens: functional interplay between proteasomes and TPPII. *Nat. Immunol.* **5**, 661–669 (2004).
8. Unanue, E. R. From antigen processing to peptide-MHC binding. *Nat. Immunol.* **7**, 1277–1279 (2006).
9. Park, B., Lee, S., Kim, E. & Ahn, K. A single polymorphic residue within the peptide-binding cleft of MHC class I molecules determines spectrum of tapasin dependence. *J. Immunol.* **170**, 961–968 (2003).
10. Burrows, S. R., Rossjohn, J. & McCluskey, J. Have we cut ourselves too short in mapping CTL epitopes? *Trends Immunol.* **27**, 11–16 (2006).
11. Trowsdale, J. HLA genomics in the third millennium. *Curr. Opin. Immunol.* **17**, 498–504 (2005).
12. Neefjes, J. & Ovaa, H. A peptide's perspective on antigen presentation to the immune system. *Nat. Chem. Biol.* **9**, 769–775 (2013).
13. Mei, S. et al. A comprehensive review and performance evaluation of bioinformatics tools for HLA class I peptide-binding prediction. *Briefings Bioinform.* **21**, 1119–1135 (2020).
14. Reynisson, B., Alvarez, B., Paul, S., Peters, B. & Nielsen, M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res.* **48**, W449–W454 (2020).
15. Andreatta, M. & Nielsen, M. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics* **32**, 511–517 (2016).
16. 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 Bioinform.* **13**, 350–364 (2012).
17. 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).
18. Purcell, A. W., McCluskey, J. & Rossjohn, J. More than one reason to rethink the use of peptides in vaccine design. *Nat. Rev. Drug Discov.* **6**, 404–414 (2007).
19. Koşaloğlu-Yalçın, Z. et al. Predicting T-cell recognition of MHC class I restricted neoepitopes. *Oncoimmunology* **7**, e1492508 (2018).
20. Rizvi, N. A. et al. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* **348**, 124–128 (2015).
21. Snyder, A. et al. Genetic basis for clinical response to CTLA-4 blockade in melanoma. *New Engl. J. Med.* **371**, 2189–2199 (2014).
22. Ott, P. A. et al. An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature* **547**, 217–221 (2017).
23. Sahin, U. et al. Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. *Nature* **547**, 222–226 (2017).
24. Slingluff, C. L. Jr The present and future of peptide vaccines for cancer: single or multiple, long or short, alone or in combination? *Cancer J.* **17**, 343–350 (2011).
25. Lu, T. et al. Deep learning-based prediction of the T-cell receptor–antigen binding specificity. *Nat. Mach. Intell.* **3**, 864–875 (2021).
26. Gfeller, D., Bassani-Sternberg, M., Schmidt, J. & Luescher, I. F. Current tools for predicting cancer-specific T-cell immunity. *Oncoimmunology* **5**, e1177691 (2016).
27. Linnemann, C. et al. High-throughput epitope discovery reveals frequent recognition of neo-antigens by CD4$^+$ T cells in human melanoma. *Nat. Med.* **21**, 81–85 (2015).
28. Bentzen, A. K. & Hadrup, S. R. Evolution of MHC-based technologies used for detection of antigen-responsive T cells. *Cancer Immunol. Immunother.* **66**, 657–666 (2017).
29. Vaswani, A. et al. Attention is all you need. In *Advances in Neural Information Processing Systems* 5998–6008 (2017).
30. Mei, S. & Li, F. Anthem: a user customised tool for fast and accurate prediction of binding between peptides and HLA class I molecules. *Brief Bioinform.* **22**, bbaa415 (2021).
31. Peters, B. & Sette, A. Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method. *BMC Bioinformatics* **6**, 132 (2005).
32. 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).
33. 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 Res.* **4**, 1–14 (2008).
34. Moutaftsi, M. et al. A consensus epitope prediction approach identifies the breadth of murine T CD8$^+$-cell responses to Vaccinia virus. *Nat. Biotechnol.* **24**, 817–819 (2006).
35. 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).
36. 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).
37. Rasmussen, M. et al. Pan-specific prediction of peptide-MHC class I complex stability, a correlate of T cell immunogenicity. *J. Immunol.* **197**, 1517–1524 (2016).
38. Hu, Y. et al. ACME: pan-specific peptide-MHC class I binding prediction through attention-based deep neural networks. *Bioinformatics* **35**, 4946–4954 (2019).
39. Jin, J. et al. Deep learning pan-specific model for interpretable MHC-I peptide binding prediction with improved attention mechanism. *Proteins* **89**, 866–883 (2021).
40. Yang, X., Zhao, L., Wei, F. & Li, J. DeepNetBim: deep learning model for predicting HLA-epitope interactions based on network analysis by harnessing binding and immunogenicity information. *BMC Bioinformatics* **22**, 231 (2021).
41. Wells, D. K. et al. Key parameters of tumor epitope immunogenicity revealed through a consortium approach improve neoantigen prediction. *Cell* **183**, 818–834 (2020).
42. 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 Res. Int.* **2020**, 5798356 (2020).
43. Bonsack, M. et al. Performance evaluation of MHC class-I binding prediction tools based on an experimentally validated MHC–peptide binding data set. *Cancer Immunol. Res.* **7**, 719–736 (2019).
44. Ebrahimi, S., Mohabatkar, H. & Behbahani, M. Predicting promiscuous T cell epitopes for designing a vaccine against *Streptococcus pyogenes*. *Appl. Biochem. Biotechnol.* **187**, 90–100 (2019).
45. Zhavoronkov, A. et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. Biotechnol.* **37**, 1038–1040 (2019).
46. Feldmann, M. & Maini, R. N. TNF defined as a therapeutic target for rheumatoid arthritis and other autoimmune diseases. *Nat. Med.* **9**, 1245–1250 (2003).
47. Le Buanec, H. et al. TNFα kinoid vaccination-induced neutralizing antibodies to TNFα protect mice from autologous TNFα-driven chronic and acute inflammation. *Proc. Natl Acad. Sci. USA* **103**, 19442–19447 (2006).
48. 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 Med.* **8**, 33 (2016).
49. Vitiello, A. & Zanetti, M. Neoantigen prediction and the need for validation. *Nat. Biotechnol.* **35**, 815–817 (2017).
50. Hathaway, J. K. HPV: diagnosis, prevention and treatment. *Clin. Obstet. Gynecol.* **55**, 671–680 (2012).

51. Yang, A., Farmer, E., Wu, T. & Hung, C.-F. Perspectives for therapeutic HPV vaccine development. *J. Biomed. Sci.* **23**, 75 (2016).
52. Madden, D. R. The three-dimensional structure of peptide-MHC complexes. *Annu. Rev. Immunol.* **13**, 587–622 (1995).
53. Yusim, K. et al. *HIV Molecular Immunology 2015* (US Department of Energy, 2016).
54. 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. *J. Immunol.* **175**, 2459–2468 (2005).
55. Niu, L. et al. Structural basis for the differential classification of HLA-A* 6802 and HLA-A* 6801 into the A2 and A3 supertypes. *Mol. Immunol.* **55**, 381–392 (2013).
56. Macdonald, W. A. et al. A naturally selected dimorphism within the HLA-B44 supertype alters class I structure, peptide repertoire and T cell recognition. *J. Exp. Med.* **198**, 679–691 (2003).
57. Jurtz, V. et al. NetMHCpan-4.0: improved peptide-MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. *J. Immunol.* **199**, 3360–3368 (2017).
58. 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. *Eur. J. Immunol.* **35**, 2295–2303 (2005).
59. Dhanda, S. K. et al. IEDB-AR: immune epitope database—analysis resource in 2019. *Nucleic Acids Res.* **47**, W502–W506 (2019).
60. 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).
61. Lata, S., Bhasin, M. & Raghava, G. P. MHCBN 4.0: a database of MHC/TAP binding peptides and T-cell epitopes. *BMC Res. Notes* **2**, 61 (2009).
62. 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).
63. Mommen, G. P. et al. Expanding the detectable HLA peptide repertoire using electron-transfer/higher-energy collision dissociation (EThcD). *Proc. Natl Acad. Sci. USA* **111**, 4507–4512 (2014).
64. Hassan, C. et al. Naturally processed non-canonical HLA-A* 02: 01 presented peptides. *J. Biol. Chem.* **290**, 2593–2603 (2015).
65. Marcilla, M. et al. Increased diversity of the HLA-B40 ligandome by the presentation of peptides phosphorylated at their main anchor residue. *Mol. Cell. Proteomics* **13**, 462–474 (2014).
66. Mobbs, J. I. et al. The molecular basis for peptide repertoire selection in the human leukocyte antigen (HLA) C* 06: 02 molecule. *J. Biol. Chem.* **292**, 17203–17215 (2017).
67. Yair-Sabag, S. et al. The peptide repertoire of HLA-B27 may include ligands with lysine at P2 anchor position. *Proteomics* **18**, 1700249 (2018).
68. Müller, M., Gfeller, D., Coukos, G. & Bassani-Sternberg, M. 'Hotspots' of antigen presentation revealed by human leukocyte antigen ligandomics for neoantigen prioritization. *Front. Immunol.* **8**, 1367 (2017).
69. Abelin, J. G. et al. Defining HLA-II ligand processing and binding rules with mass spectrometry enhances cancer epitope prediction. *Immunity* **51**, 766–779 (2019).
70. Kalaora, S. et al. Use of HLA peptidomics and whole exome sequencing to identify human immunogenic neo-antigens. *Oncotarget* **7**, 5110–5117 (2016).
71. Faridi, P., Purcell, A. W. & Croft, N. P. In immunopeptidomics we need a sniper instead of a shotgun. *Proteomics* **18**, e1700464 (2018).
72. 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).
73. 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).
74. 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 Rheumatol.* **67**, 702–713 (2015).
75. Illing, P. T. et al. HLA-B57 micropolymorphism defines the sequence and conformational breadth of the immunopeptidome. *Nat. Commun.* **9**, 4693 (2018).
76. 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. *J. Proteome Res.* **15**, 1059–1069 (2016).
77. Hillen, N. et al. Essential differences in ligand presentation and T cell epitope recognition among HLA molecules of the HLA-B44 supertype. *Eur. J. Immunol.* **38**, 2993–3003 (2008).
78. Kaur, G. et al. Structural and regulatory diversity shape HLA-C protein expression levels. *Nat. Commun.* **8**, 15924 (2017).
79. Liu, G. et al. PSSMHCpan: a novel PSSM-based software for predicting class I peptide-HLA binding affinity. *Gigascience* **6**, gix017 (2017).
80. O'Donnell, T. J. et al. MHCflurry: open-source class I MHC binding affinity prediction. *Cell Syst.* **7**, 129–132 (2018).
81. Liu, Z. et al. DeepSeqPan, a novel deep convolutional neural network model for pan-specific class I HLA-peptide binding affinity prediction. *Sci. Rep.* **9**, 794 (2019).
82. Phloyphisut, P., Pornputtapong, N., Sriswasdi, S. & Chuangsuwanich, E. MHCSeqNet: a deep neural network model for universal MHC binding prediction. *BMC Bioinformatics* **20**, 270 (2019).
83. 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**, 7 (2019).
84. Alvarez, B. et al. NNAlign_MA; MHC peptidome deconvolution for accurate MHC binding motif characterization and improved T-cell epitope predictions. *Mol. Cell. Proteomics* **18**, 2459–2477 (2019).
85. Stranzl, T., Larsen, M. V., Lundegaard, C. & Nielsen, M. NetCTLpan: pan-specific MHC class I pathway epitope predictions. *Immunogenetics* **62**, 357–368 (2010).
86. Vang, Y. S. & Xie, X. HLA class I binding prediction via convolutional neural networks. *Bioinformatics* **33**, 2658–2665 (2017).
87. Han, Y. & Kim, D. Deep convolutional neural networks for pan-specific peptide-MHC class I binding prediction. *BMC Bioinformatics* **18**, 585 (2017).
88. Singh, H. & Raghava, G. ProPred1: prediction of promiscuous MHC Class-I binding sites. *Bioinformatics* **19**, 1009–1014 (2003).
89. Shao, X. M. et al. High-throughput prediction of MHC class I and II neoantigens with MHCnuggets. *Cancer Immunol. Res.* **8**, 396–408 (2020).

Acknowledgements

This work was supported in part by the National Science Foundation of China (grants nos. 61832019, 61872094, 32030063, 32070662 and 62172274), the National Key R&D Program of China (grant no. 2016YFA0501703), the Science and Technology Commission of Shanghai Municipality (grant no. 19430750600), as well as the SJTU JiRLMDS Joint Research Fund and Joint Research Funds for Medical and Engineering and Scientific Research at Shanghai Jiao Tong University (grant no. YG2021ZD02). The computational experiments were partially run at the Pengcheng Laboratory and the Center for High-Performance Computing, Shanghai Jiao Tong University.

Author contributions

Y.C. and Y.Z. conceived the original ideas for this study, designed and performed the experiments, and co-wrote the manuscript. Q.W. and Q.X. were responsible for MD simulations. Y.X., L.Z. and X.W. contributed to the model design and revision of the manuscript. Y.W., J.W. and X.J. helped to prepare figures. D.R.S., Y.X. and D.-Q.W. guided the work.

Competing interests

The authors declare no competing interests.

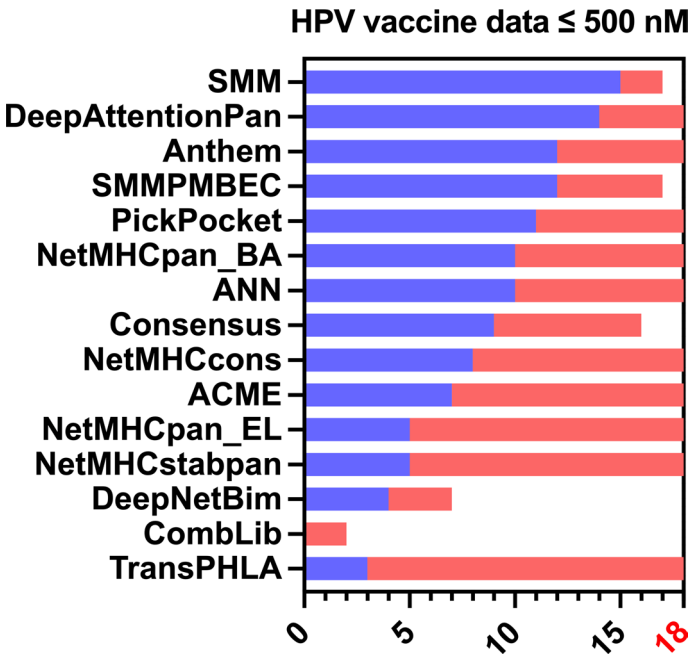
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s42256-022-00459-7.

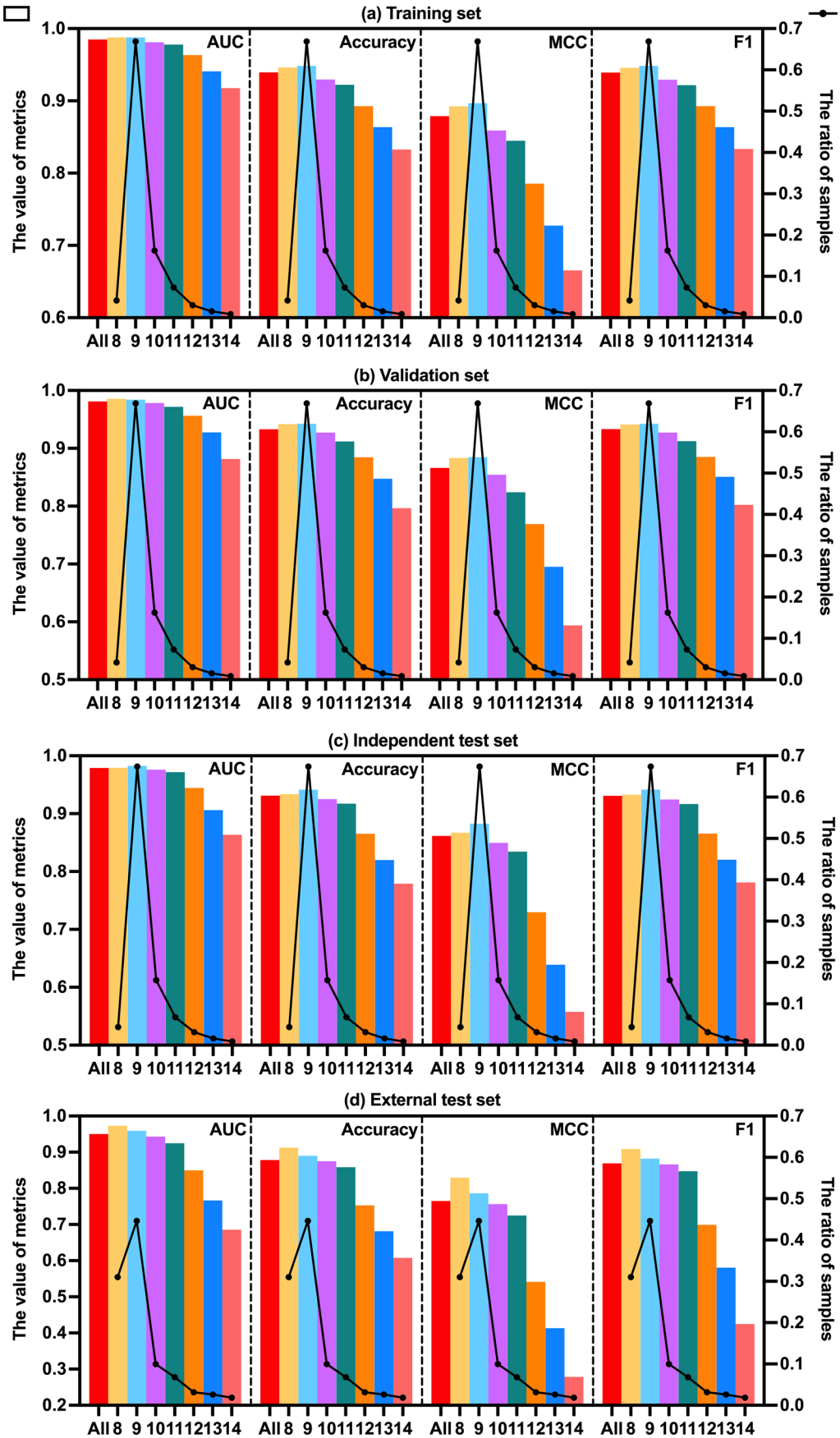
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-022-00459-7.

Correspondence and requests for materials should be addressed to Yi Xiong or Dong-Qing Wei.

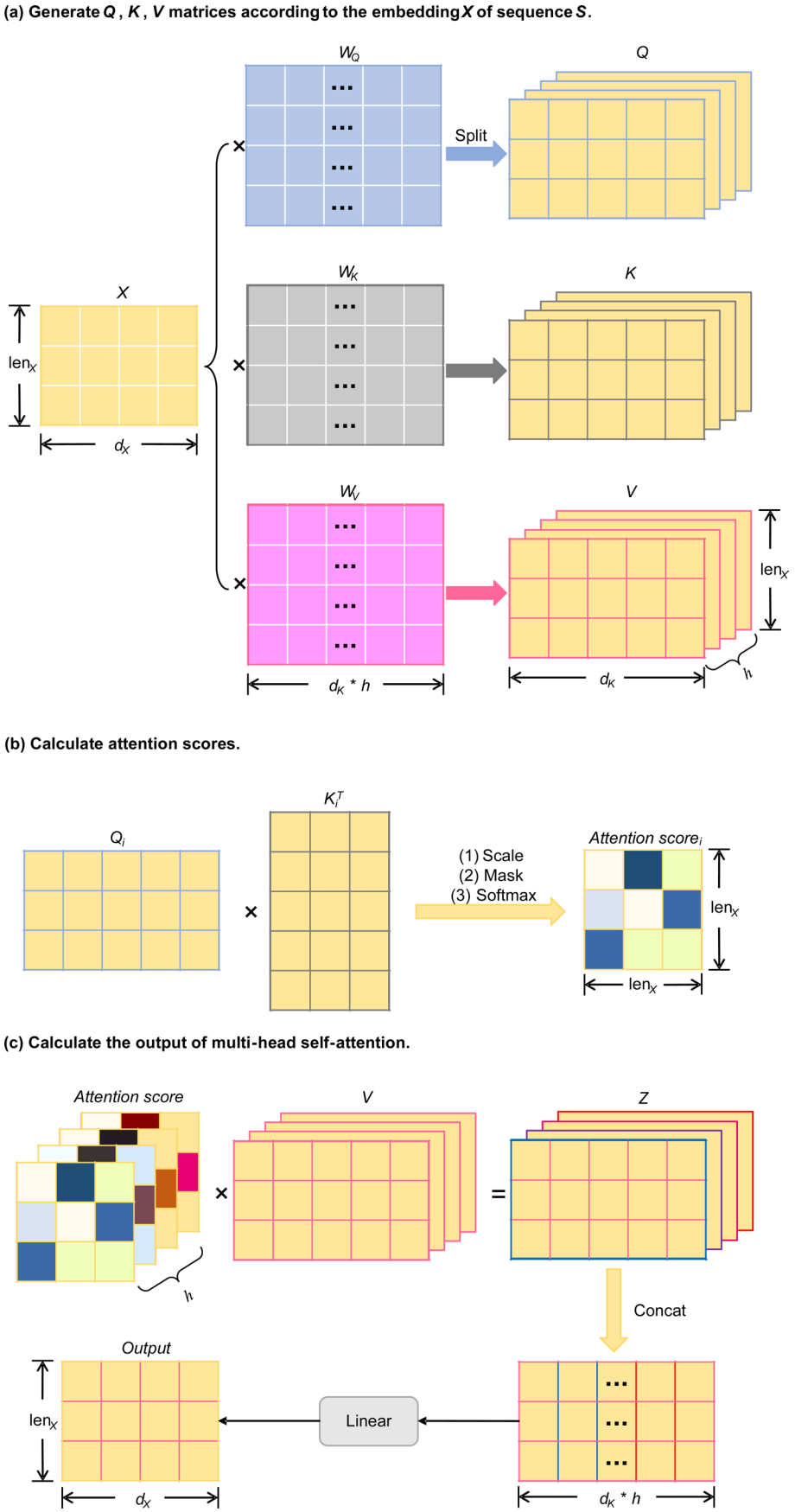
Peer review information *Nature Machine Intelligence* thanks Yuedong Yang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



Extended Data Fig. 1 | Comparison of the proposed TransPHLA method with 14 existing methods on HPV vaccine data with threshold 500 nM. The number of true positive and false negative are described, and the sum of true positive and false negative represents the number of predictable peptide-HLA-I binders.



Extended Data Fig. 2 | Correlation between the prediction performance and peptide length on various datasets. The correlation between the performance and peptide length on the (**a**) training set, (**b**) validation set, (**c**) independent test set, and (**d**) external test set. The performance is displayed in bar based on the left ordinate, and the distribution of the ratio of peptides with different lengths is displayed in dots and lines based on the right ordinate.



Extended Data Fig. 3 | Workflow to calculate masked multi-head self-attention. The workflow to implement masked multi-head self-attention by three steps: (**a**) Generation of matrics, (**b**) Calculation of attention scores, (**c**) Calculation of output, 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.