









# Meta learning for mutant HLA class I epitope immunogenicity prediction to accelerate cancer clinical immunotherapy

Long Xu , Qiang Yang <sup>1,2</sup>, Weihe Dong <sup>3</sup>, Xiaokun Li <sup>1,4,5,6,\*</sup>, Kuanquan Wang <sup>1,\*</sup>, Suyu Dong<sup>3</sup>, Xianyu Zhang <sup>7</sup>, Tiansong Yang<sup>8</sup>, Gongning Luo <sup>1,9,\*</sup>, Xingyu Liao<sup>10</sup>, Xin Gao <sup>9</sup>, Guohua Wang <sup>1,3,\*</sup>

<sup>1</sup>School of Computer Science and Technology, Harbin Institute of Technology, West DaZhi Street, 150001 Harbin, China

<sup>2</sup>School of Medicine and Health, Harbin Institute of Technology, Yikuang Street, 150000 Harbin, China

<sup>3</sup>College of Computer and Control Engineering, Northeast Forestry University, Hexing Road, 150040 Harbin, China

<sup>4</sup>School of Computer Science and Technology, Heilongjiang University, Xuefu Road, 150080 Harbin, China

<sup>5</sup>Postdoctoral Program of Heilongjiang Hengxun Technology Co., Ltd., Xuefu Road, 150090 Harbin, China

<sup>6</sup>Shandong Hengxun Technology Co., Ltd., Miaoling Road, 266100 Qingdao, China

<sup>7</sup>Department of Breast Surgery, Harbin Medical University Cancer Hospital, Haping Road, 150081 Harbin, China

<sup>8</sup>Department of Rehabilitation, The First Affiliated Hospital of Heilongjiang University of Traditional Chinese Medicine, Xuefu Road, 150040 Harbin, China

<sup>9</sup>Computer, Electrical and Mathematical Sciences & Engineering Division, King Abdullah University of Science and Technology, Thuwal 23955, 4700 KAUST Saudi, Arabia

<sup>10</sup>School of Computer Science, Northwestern Polytechnical University, 710072 Xian, China.

\*Corresponding authors. Xiaokun Li, E-mail: [lixiaokun@hlju.edu.cn](mailto:lixiaokun@hlju.edu.cn); Kuanquan Wang, E-mail: [wangkq@hit.edu.cn](mailto:wangkq@hit.edu.cn); Gongning Luo, E-mail: [luogongning@hit.edu.cn](mailto:luogongning@hit.edu.cn); Guohua Wang, E-mail: [ghwang@nefu.edu.cn](mailto:ghwang@nefu.edu.cn)

## Abstract

Accurate prediction of binding between human leukocyte antigen (HLA) class I molecules and antigenic peptide segments is a challenging task and a key bottleneck in personalized immunotherapy for cancer. Although existing prediction tools have demonstrated significant results using established datasets, most can only predict the binding affinity of antigenic peptides to HLA and do not enable the immunogenic interpretation of new antigenic epitopes. This limitation results from the training data for the computational models relying heavily on a large amount of peptide-HLA (pHLA) eluting ligand data, in which most of the candidate epitopes lack immunogenicity. Here, we propose an adaptive immunogenicity prediction model, named MHLAPre, which is trained on the large-scale MS-derived HLA I eluted ligandome (mostly presented by epitopes) that are immunogenic. Allele-specific and pan-allelic prediction models are also provided for endogenous peptide presentation. Using a meta-learning strategy, MHLAPre rapidly assessed HLA class I peptide affinities across the whole pHLA pairs and accurately identified tumor-associated endogenous antigens. During the process of adaptive immune response of T-cells, pHLA-specific binding in the antigen presentation is only a pre-task for CD8<sup>+</sup> T-cell recognition. The key factor in activating the immune response is the interaction between pHLA complexes and T-cell receptors (TCRs). Therefore, we performed transfer learning on the pHLA model using the pHLA-TCR dataset. In pHLA binding task, MHLAPre demonstrated significant improvement in identifying neopeptide immunogenicity compared with five state-of-the-art models, proving its effectiveness and robustness. After transfer learning of the pHLA-TCR data, MHLAPre also exhibited relatively superior performance in revealing the mechanism of immunotherapy. MHLAPre is a powerful tool to identify neopeptides that can interact with TCR and induce immune responses. We believe that the proposed method will greatly contribute to clinical immunotherapy, such as anti-tumor immunity, tumor-specific T-cell engineering, and personalized tumor vaccine.

**Keywords:** epitope specificity; HLA genotyping; immunoinformatics; deep learning; transfer learning

## Introduction

T cell-based immunotherapy has achieved significant progress in the field of clinical oncology [1, 2]. Current evidence suggests that tumor immunotherapy relies on T-cell recognition of tumor-specific neoantigens that generated by nonsynonymous mutations in tumor tissues, thereby selectively eliminating corresponding tumor cells. In tumor immunotherapy, class I Human Leukocyte Antigen (HLA-I) plays a central role in recognizing and binding to these neoantigens [3–6]. HLA alleles are primarily

distributed on the short arm of chromosome 6 [7]. Short peptides bound to HLA-I molecules mainly originate from intracellular proteins, which are proteolytically cleaved by proteasomes and peptidases before loading and expression. HLA-I molecules present antigens to CD8<sup>+</sup> T-cell receptors (TCRs), which elicits the clonal expansion of cytotoxic T-cells to achieve the elimination of cancer cells and precise therapy [8]. Therefore, predicting the affinity of HLA I antigen peptides has become a key factor in neoantigen immunotherapy and may serve as a breakthrough [9]. However,

Received: March 11, 2024. Revised: September 18, 2024. Accepted: November 14, 2024

© The Author(s) 2024. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact [journals.permissions@oup.com](mailto:journals.permissions@oup.com)

though many computational methods have been developed for predicting binding affinities of HLA I bound epitopes, deficiencies still remain in indicating immunogenicity.

Computational models for predicting the binding affinity between peptides and HLA alleles, particularly for HLA-A and HLA-B alleles, have been greatly improved. Based on decades of laboratory work and accumulated data on HLA binding sequences, more than 30 tools are now available for predicting HLA-antigen interactions. However, even widely used algorithms, such as NetMHCpan [10, 11], show an increase in incorrectly predicted binding entities when the half-maximal inhibitory concentration (IC<sub>50</sub>) exceeds 100 mol. Most of the predictors, such as NetMHCpan, MHCnuggets [12, 13], and MHCflurry [14] only focus on the binding between HLA and peptide segments, whereas further experimental verification is needed to determine the immunogenicity of HLA I epitopes. Besides, some tools, such as MixMHCpred2.2 [6, 15, 16] and HLAthena [17, 18], rely on randomly generated negative samples or utilize transfer learning with HLA eluted peptide data to adapt the model to immunogenicity.

Several computational models are available for predicting peptide-HLA (pHLA)-TCR binding specificity. They are mainly divided by two categories: (1) peptide-specific TCR binding prediction models, including TCRGP [19], MixTCRPred [20], and NetTCR2.0 [21]; (2) predictive models for peptide-TCR binding using peptide-nonspecific peptides that are trained based on known binding TCRs, including pMTnet [22], DLpTCR [23], ERGO2 [24], PanPep [25]. Unfortunately, the models in the first category are restricted to specific peptides with limited utility, and the second category to develop general pHLA-TCR binding prediction models.

Generally, predicting pHLA binding affinity or antigenic-TCR interactions are usually treated as two separate tasks. However, these tasks are highly repetitive and correlated with the data used for model training as well as human physiological processes. This inspired us to adapt the previously learned parameters from the predicted antigenic pHLA immunogenicity model to improve the T-cell receptor interaction task using transfer learning. As shown in Fig. 1(a), available training data mainly come from eluting bound pHLA, detecting bound peptide sequences using mass spectrometry, and detecting HLA allele types by gene expression profiling. However, such data only represent whether the candidate antigenic peptide can be presented by HLA, but cannot explain whether the peptide can elicit an immune response from CD8+ T-cells. Therefore, we used laboratory-validated pHLA data capable of eliciting an immune response to simulate this physiological process based on a deep learning model, aiming to accurately predict whether a candidate antigen is immunogenic.

In this paper, we propose two MHLAPre models, namely MHLAPre IM and MHLAPre TT. To predict the immunogenicity of neoantigens, MHLAPre IM is trained on experimentally confirmed class I HLA-bound epitopes (pHLA pairs) that possess immunogenicity. Subsequently, based on MHLAPre IM, transfer learning is directly performed on pHLA-TCR data to build the MHLAPre TT model. Although transfer learning changes the scope of the original task, it has highly contextual relevance in terms of data representation because of the temporal order of tasks throughout the immune response process. Specifically, we propose the MHLAPre IM model to predict the epitope immunogenicity by Model-Agnostic Meta-Learning (MAML) [26], which allows more sensitive learning of data feature information on immunogenic samples with insufficient HLA eluted peptide data. For amino acid sequence encoding, the Blosom62 (Block substitution Matrix 62) are employed to encode antigenic peptides

and HLA alleles [27, 28]. The complementary decision region 3 (CDR3 $\beta$ ) of the TCRs  $\beta$  chain is considered as one of the inputs at the beginning of transfer learning. The representation of HLA is derived from the polymorphic residues of alleles A, B and C. A pseudosequence consisting of the 34 most important amino acid residues is used to represent each allele. In our experiments, we compared MHLAPre IM with five state-of-the-art models of HLA-antigen binding prediction and four models for pHLA-TCR binding prediction and discussed the results based on the comparison. In order to dig deeper into MHLAPre, we further compared it with pHLA-specific models (NetTCR2.0, MixTCRPred), and further research directions in the field afterwards are discussed.

## Results

### Framework development

MHLAPre is a generalized framework designed to predict whether epitopes of pHLA pairs are immunogenic. The development of MHLAPre was inspired by the effective integration of meta-learning [29–33] strategies and powerful learning ability of transfer learning [34]. The overall architecture of our model is shown in Fig. 2. MHLAPre includes a meta-learning strategy, a backbone network, and a transfer learning module. We use a Model-Agnostic Meta-Learning framework, which enables the model to be trained on multiple tasks, optimizes the initialization of the model, and allows the model to adapt quickly when encountering a new task. Additionally, since the training data are in a long-tailed distribution, we introduce a dynamic sampling technique to dynamically extract the support and query sets of pHLA or pHLA-TCR based on peptide types. Furthermore, it enables quickly respond to whether a neoantigen can trigger the immune reactivity of CD8+ T-cells by learning a priori immunogenicity data. To construct the MHLAPre IM model, we introduce the encoder module of the Transformer [28, 29] to process the encoded amino acid sequences. Transformer has achieved outstanding results in the encoding and processing of textual data. The Transformer encoder takes the pHLA matrix as an input to obtain a matrix in the same format as the input of the backbone network. The backbone network adopts TextCNN [30, 31], while taking the support set and query set data as input. We combined the knowledge of antigenic peptide sequences and HLA alleles in a biologically meaningful way as the input data, vectorized it, and enriched the potential features of the data by multiple coding rules. After obtaining the standardized output of the TextCNN network, we constructed a tri-layer fully-connected network to perform the dimensionality compression, and finally obtained the scoring model. Furthermore, we developed the MHLAPre TT model by introducing a transfer learning module. The immunogenic pHLA-TCR data were fed into the optimized backbone network. As shown in Fig. 1(c), we cleaned and divided the raw immunogenicity data collated from the IEDB [32] into different datasets. As shown in Fig. 1(b) and 1(d), we statistically characterised the sequence features of the antigenic peptides and the distribution of HLA alleles in the training set to better observe the feature information of the training data, which we investigate in detail later. Of note, the changes between the input data of IM and TT model were generated. In order to unify the input data format as well as to retain the features of the pHLA data, we spliced the antigenic peptides that have been vectorized and quantized with the TCR CDR3 sequence data to construct the neural network to unify the vector dimensions. Thus, MHLAPre TT is capable of predicting the immunogenicity of a pHLA

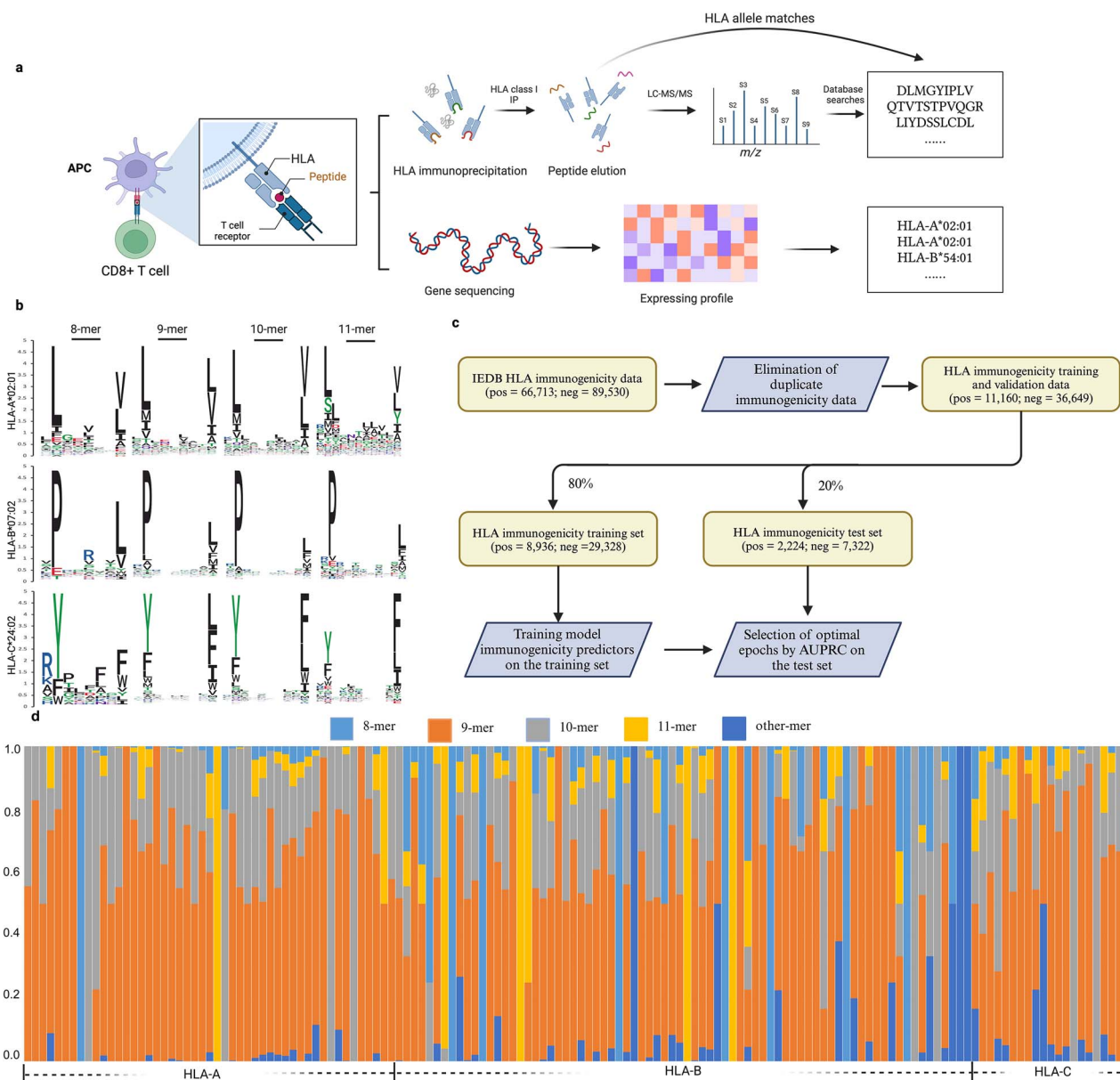


Figure 1. Statistical information on immunogenicity data. (a) Sequence determination by mass spectrometry methods after elution of antigenic peptides; allele type determination by gene expression. (b) Amino acid site frequency plots of three HLA alleles binding antigenic peptides classified according to peptide length (HLA-A\*02:01, HLA-B\*07:02, HLA-C\*24:02). (c) Raw data cleaning process, division ratio between training and test sets; (d) Observed peptide length frequencies across alleles. HLA-A alleles bind longer peptides more frequently compared with HLA-B and -C alleles, which tend to bind shorter peptides. Panel **a** created with BioReader.com.

complex and corresponding pHLA-TCR pair to determine whether a neoantigen trigger an immune response with the cytotoxic T-cells. The specific model structure information is provided in [Supplementary Fig. 2](#) and [Supplementary Table S6](#).

### Epitope immunogenicity prediction of pHLA complex

The vast majority of neoepitopes are not immunogenic, and it is inevitable that errors may occur when learning features in different training process. Specifically, MHLAPre learned large-scale eluted peptides with HLA I epitope data first, and then learned a pHLA-TCR dataset by transfer learning. Thus, directly using immunogenicity data for training and mining immunogenicity features in limited samples by meta-learning may avoid bias in weights during pre-training and knowledge transferring.

We evaluated with NetMHCpan [10, 11], MHCnuggetl [12, 13], MHCflurry [14], MixMHCpred2.2 [15, 16], and HLathena [17, 18] on the same epitope immunogenicity data, which collate from IEDB. The area under the Receiver Operating Characteristic (ROC) curve (AUROC) and the area under the Precision-Recall (PR) curve (AUPRC) were used as evaluation criteria. Based on the data characteristics of pHLA pairs, the epitope immunogenicity task was evaluated in two different modes, 1) allele-specific mode to study the binding preference of three HLA loci, and 2) length-specific modes to survey the peptide length influence the binding performance [6].

As shown in [Fig. 3\(a-b\)](#), we stratified the results according to the subdivided categories of HLA and demonstrated the distribution of AUROC and AUPRC for each classifier on HLA loci. MHLAPre showed significantly compelling performance for all

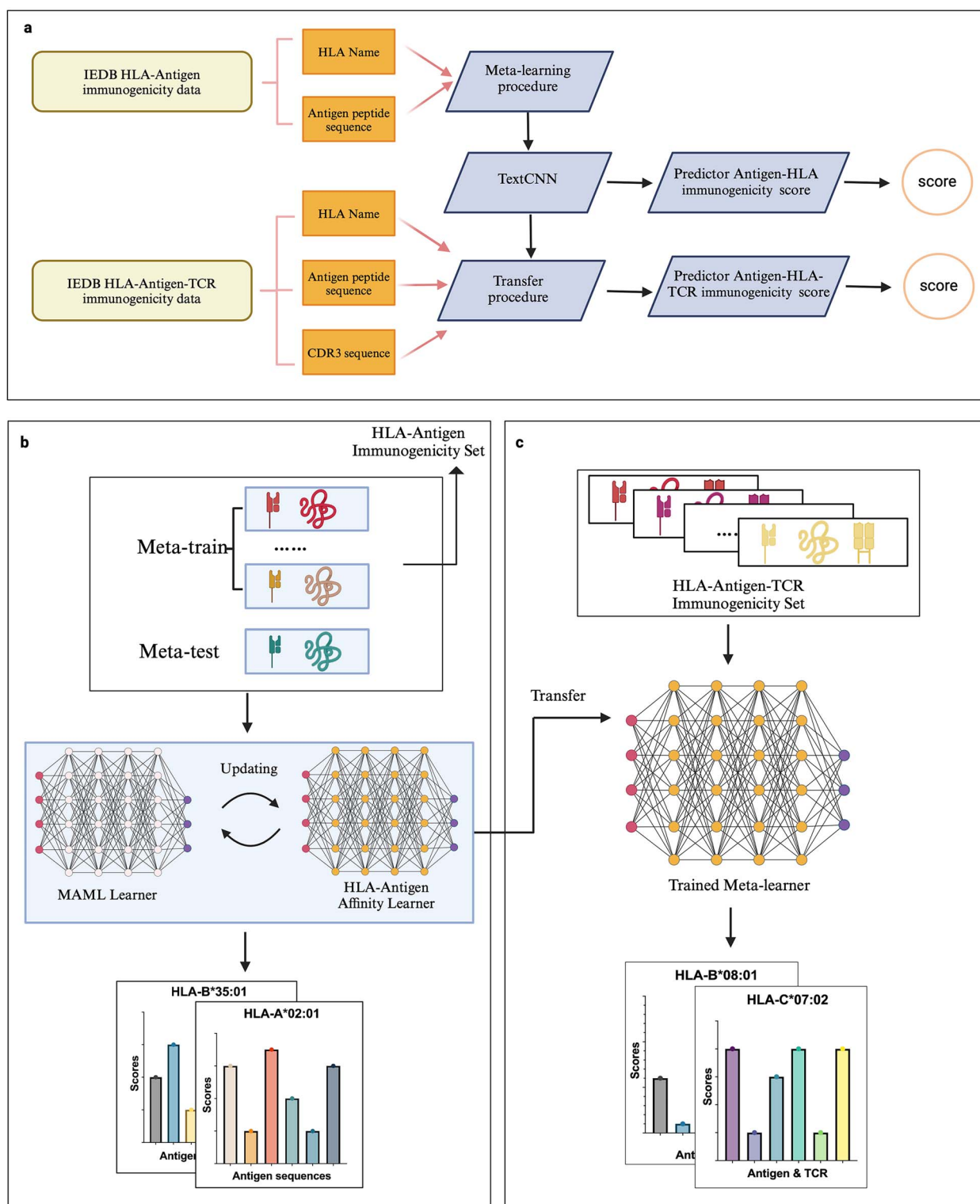


Figure 2. Workflow diagram of the model. (a) Data input structure of the MHLAPre model and model workflow. (b) Training process of MHLAPre IM against pHLA complex epitope immunogenicity. (c) Transfer learning and prediction process of MHLAPre TT in the pHLA-TCR scenario. Panels **a-c** created with BioReader.com.

HLA loci, whereas other models obtained a slight decrease, especially on HLA-C. Under the application of full data evaluation, MHLAPre obtained an average AUROC and AUPRC values of 0.9041 and 0.8462, respectively. In comparison, the best available priori method, HLATHENA, obtained an average AUROC and AUPRC of

0.8601 and 0.7878. The data of other methods are shown in [Supplementary Table S1](#).

As shown in [Fig. 3\(c\)](#), we stratified according to HLA and antigenic peptide length. After applying a more refined stratification, MHLAPre obtained an average AUROC and AUPRC of 0.8542 and



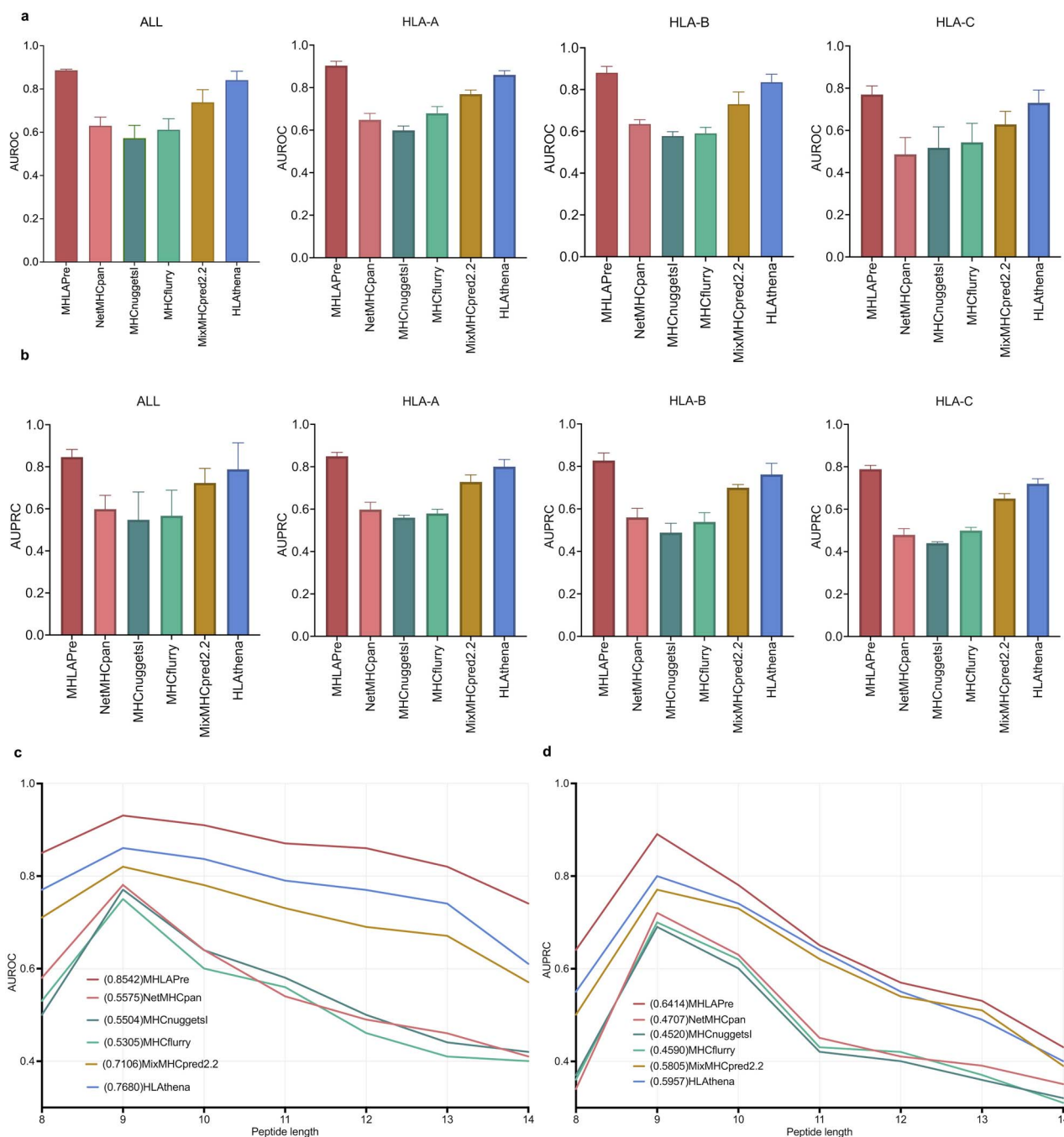


Figure 3. Performance evaluation of different pHLA antigen affinity prediction algorithms. (a) The figure presents a comparative analysis of the predictive accuracy of several antigen presentation prediction algorithms. The top panel displays the AUROC for each algorithm, whereas the bottom panel shows the AUPRC. Each bar represents the average performance score of the respective algorithms—MHCflurry, NetMHCpan, MHCnuggets, MixMHCpred 2.2, and HLAthena—across all HLA types (ALL) as well as individually stratified for HLA-A, HLA-B, and HLA-C alleles. The error bars correspond to the standard deviation of the performance scores, encapsulating the variability of the algorithm's predictive power. This comprehensive assessment underscores the varying degrees of efficacy that these computational tools exhibit when tasked with predicting the presentation of antigens by different HLA molecules. (b) The left panel shows the AUROC for each algorithm and the right panel shows the AUPRC. Each line represents the average performance score of different algorithms for different lengths of antigenic peptides.

0.6414, respectively, compared to an average AUROC and AUPRC of 0.7680 and 0.5957 for HLAthena. MHLAPre was most effective for peptides of length 9, which is also the most common peptide length for HLA-I classes. Although the predictive power of MHLAPre for immunogenic pHLA decreases with increasing antigenic peptide length, it still outperforms existing advanced methods to a certain extent in cross-sectional comparisons of all peptide lengths. In conclusion, MHLAPre can accurately predict

the epitope immunogenicity of pHLA pairs in both allele-specific and length-specific modes (shown in [Supplementary Tables S2 and S3](#)).

Compared to other models, MHLAPre IM demonstrates outstanding performance advantages in predicting candidate immunogenic antigens. This is because MHLAPre is trained on immunogenic pHLA data, whereas NetMHCpan, MHCnuggets, MHCflurry, and MixMHCpred2.2 are trained on eluted peptide

data after mass spectrometry determination. The predicted results of these models represent the affinity of candidate peptides and HLA, which significantly differs from our model's prediction performance. Notably, HLAthena achieves performance close to our model because it transfers the parameters from the eluted peptide data to the immunogenicity data through transfer learning. However, since not all candidate antigens presented via HLA can elicit an immune response from CD8+ T-cells, HLAthena's reliance on eluted peptide data and the parameters learned through transfer learning based on small-scale immunogenicity data still result in some bias in learning the overall features of the immunogenicity data. Therefore, HLAthena is still insufficient compared to MHLAPre.

### Prediction of TCR binding specificity with pHLA

Since the immunogenicity candidate antigen results predicted by MHLAPre IM can be used as a precursor task for pHLA interaction with TCRs, this provides a basis for transferring from the antigen-HLA affinity prediction task to the TCR-pHLA interaction prediction task. Therefore, we proposed the MHLAPre TT model based on the MHLAPre IM model (shown in Fig. 2(a-c), and Supplementary Fig. 2) to enable the prediction of the TCR-pHLA interaction task. We collated the TCR-pHLA dataset from the IEDB database as a training set for the transfer learning process. As shown in Fig. 4(a), the loss of MHLAPre TT model decreased rapidly and tended to be stable when the epochs were approximately 50 during the 241 transferring process. Since the limited number of experimental validated pHLA-TCR data, it is necessary to construct a candidate list of immunogenic candidate neoepitopes to support validation in clinical settings. Therefore, only the top prediction results in practical applications can be further proven by experiments, and the prediction results must maintain a high accuracy. To evaluate the accuracy, it is usually necessary to use Positive Predictive Value (PPV) and more dimensional performance indicators to evaluate models. We suggested that the mean value of the 2% PPV among the highest scoring peptides is a more appropriate measurement to assess predictors of TCR binding specificity. Because it is important to assess the true positive rate at the actual 2% prevalence of the conjugate. AUROC and AUPRC were also used as evaluation indicators. To summarize the performance comparison with the prior method, the average PPV was plotted in Fig. 4(c-d). As shown, the average PPV of MHLAPre TT was 0.7397, which is better than the current best prior method pMTnet with an average PPVn of 0.7105. In mean AUROC and AUPRC, MHLAPre TT yielded values of 0.7996 and 0.7623, and pMTnet was 0.7603 and 0.7899 (shown in Supplementary Table S4). We evaluated MHLAPre TT according to different HLA genotypes. MHLAPre performed well in different genotyped samples (shown in Fig. 4e-f). The results demonstrated the effectiveness of transfer learning in the transfer of epitope immunogenicity data to Tcell binding specificity data to reveal the significantly clinical benefits for cancer immunotherapy.

### Network architecture study

We further discussed how the meta-learning framework affected the performance of MHLAPre. To investigate this, we ablated the meta-learning framework and retrained the model. When stratified by allele and peptide length, the MAML framework improved the IM AUROC by 0.0256 and AUPRC by 0.0158, particularly for longer peptides (11–15 mers). Similarly, the TT model's AUROC was improved by 0.0014 and AUPRC was improved by 0.0058,

which was not as significant as on the IM data, but still a modest improvement.

### MHLAPre IM reveals how to predict unseen pHLA data mechanism

To investigate the prediction performance of MHLAPre IM on rare alleles and antigenic peptides not present in the training set, we integrated class I human-derived HLA-peptide pairs from the VDJdb (Score $\geq$ 1) and McPAS-TCR databases. These were then compared and filtered against the previously compiled IEDB data to identify pHLA pairs not present in the IEDB dataset. We finally collected 69 pairs of positive pHLA samples that had not been seen before (VDJdb: 43, McPAS-TCR: 26). We searched to obtain 19 peptides that did not appear in the training set. HLA alleles with less than 30 occurrences in the training set were defined as rare alleles, and 36 rare alleles were obtained. We input the unseen positive HLA-peptide dataset into MHLAPre IM and obtained predictions (Fig. 5(a), Supplementary Table S12). We observed that pHLA pairs with high-frequency HLA alleles and antigenic peptides that had appeared in the IEDB dataset received higher scores (Supplementary Figs 16 and 17). Notably, the unseen antigenic peptides (FLGKIWPESHK\_HLA-A\*02:01 and GPEPLPQQQLTAY\_HLA-B\*07:02) also obtained excellent prediction scores due to the very high frequency of HLA-A\*02:01 (Frequency: 9 741) and HLA-B\*07:02 (Frequency: 2 402) in the training set, and these two HLA alleles were sufficiently trained in the MHLAPre IM training set. MHLAPre learned enough binding features of the related genes. The novel pHLA data samples (WLDNFELCL\_HLA-B\*51:193, TYDTVHRHL\_HLA-A\*08:01) had the lowest scores, suggesting that the model did not sufficiently learn the information from this sample. We found that HLA sequences of HLA alleles with high sequence similarity to the high frequency alleles performed excellently in the prediction results (e.g. HLA-A\*02:01 is similar to the sequences of HLA-A\*02:11, HLA-A\*02:09, HLA-A\*02:14, and HLA-A\*02:10).

Overall, our results show that predicting unseen pHLA samples and candidate antigen immunogenicity is highly dependent on the training set for the current scale of pHLA sequence data. Successful prediction is only possible when the unseen pHLA is associated with the presence of HLA alleles or antigens in the MHLAPre IM training set, or when the HLA allele sequences have a high level of sequence similarity to the HLA alleles in the training set. These findings further demonstrate the expansion potential of the MHLAPre IM model. Future directions could involve addressing the shortage of pHLA data by mining the sequence similarity features of high-frequency alleles and rare allele sequences to achieve more accurate prediction of antigen immunogenicity with large-scale training.

### MHLAPre TT exhibits high potential performance in peptide-specific TCR binding prediction

To further validate the performance and potential of MHLAPre in pHLA-TCR data, we compared the prediction performance with the pHLA-specific models NetTCR2.0 and MixTCRPred. We therefore further validated the antigen-specific binding prediction performance of MHLAPre TT against antigens restricted to the HLA-A\*02:01 allele. We compared this with the current performance in peptide-TCR binding prediction models NetTCR2.0 and MixTCRPred. The NetTCR2.0 model restricts pMHC-TCR interaction prediction to the HLA-A\*02:01 allele, and supports prediction using both paired CDR3 $\alpha/\beta$  versus CDR3 $\beta$  only on TCR inputs. MixTCRPred is a pMHC-specific predictor, where a separate model is trained for each epitope. It is based on the pMHC classes, 113

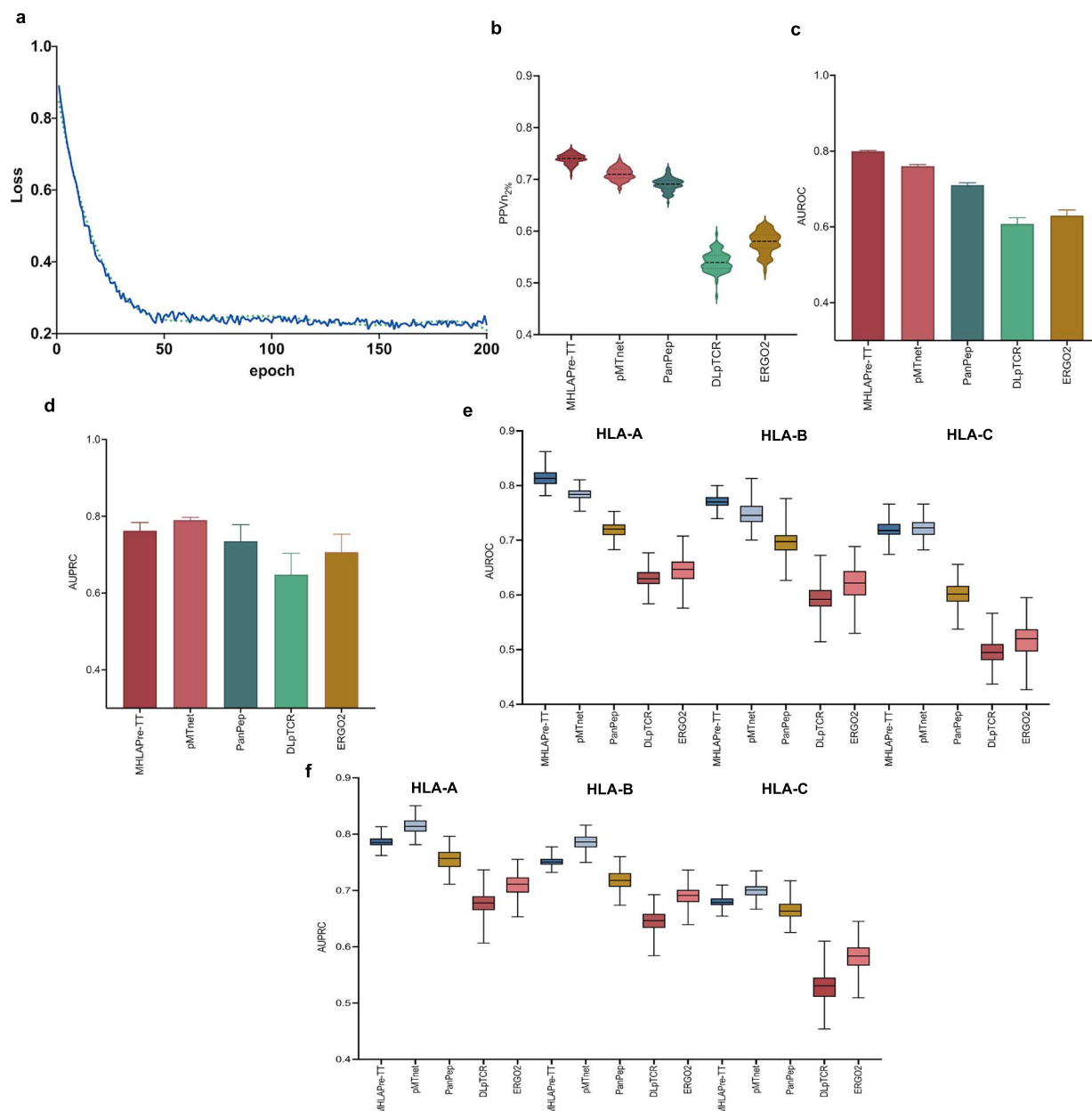
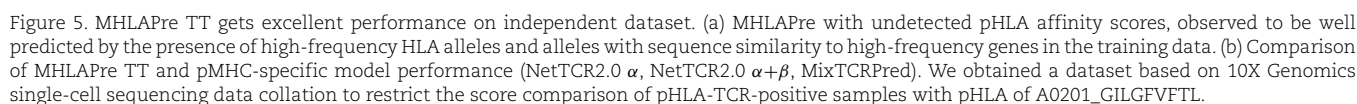


Figure 4. Performance comparison of MHLAPre TT with different pHLA-TCR interaction prediction models. (a) Trend of loss function loss for the MHLAPre TT transfer learning process, where it can be observed that the MHLAPre TT model learnt new environmental knowledge and fitted quickly. (b) PPV of the top 2% for the comparison of different models. Each bar or violin represents the average performance score of the respective algorithms pMTnet, PanPep, DLpTCR, ERGO2 across all HLA types (ALL) as well as individually stratified for HLA-A, HLA-B, and HLA-C alleles. (c, d) Mean AUROC and AUPRC for different model comparisons. (e, f) Mean AUROC and AUPRC plots grouped by HLA-A, -B, -C alleles.

pMHC-specific predictor models are trained on human source data for different pMHCs. The focus of this study is on data from the 10X Genomics Chromium Single Cell Immunoassay Profiling platform, which applies feature barcoding technology to generate single-cell 5' libraries and V(D)J-enriched libraries for TCR sequences identified using highly multiplexed pMHC multimer reagents. We examined a single-cell dataset containing 44 pMHC complex-unspecific CD8+ T-cell profiles from a healthy donor who had not been virally infected. We studied their clonal Tcell expansions to screen for TCRs capable of interacting with pMHC binding and counted them according to the original unique molecular identifiers (UMIs), when UMIs  $\geq 10$ , we determined that the pMHC-TCR pair could interact with T cells and counted it as a positive sample. Since NetTCR 2.0 and MixTCRPred have

specificity requirements for pMHC, and NetTCR is even more restricted to the input of HLA-A\*02:01 with specific three peptides (GILGFVFTL, NLVPMVATV, GLCTLVAML), we finally used the HLA-A\*02:01 and GILGFVFTL combination for screening. We finally screened 2 083 TCR-pHLA pairs as a test dataset, and since the results of the test set were all positive data, we performed statistics based on the average of the final scores of each model. The results are shown in Fig. 5(b). The average scores of the four models on our collated TCR-pHLA positive test set are MHLAPre TT (0.8953), NetTCR2.0 $\beta$  (0.7608), NetTCR2.0 $\alpha+\beta$  (0.9404), MixTCRPred (0.8046). MHLAPre scores over TCRNet2.0 $\beta$  and MixTCRPred and below TCRNet2.0 $\alpha+\beta$  for CDR3 $\beta$ -only inputs. The results suggest that migrating the a priori learned parameters of MHLAPre IM in immunogenicity data to the pHLA-TCR environment has the



As a pan-peptide model, MHLAPre is able to achieve superior performance with pHLA-specific predictors. It is shown that MHLAPre has great potential for transferring from immunogenicity models to the TCR-pHLA environment. In future studies, we will use CDR3 $\alpha$  and CDR1 and CDR2 as inputs to improve the predictive and generalisation performance of the model [33].

In this study, we proposed a deep learning framework based on meta-learning and transfer learning to predict the immunogenicity of candidate antigens. An immunogenicity metadata dataset was constructed, which contained 146 HLA alleles and 47 831 samples from the IEDB database [32]. The Transformer [28, 29] Encoder module was employed to context-preprocess the encoded pHLA matrix. The MHLAPre IM model was proposed and the MAML [26, 34–36] framework was used to generate the task learning of TextCNN [30, 31]. Additionally, the pHLA-TCR dataset was introduced. We performed transfer learning [37] to construct the projection of pHLA to pHLA-TCR and evaluated it with corresponding metrics. The results demonstrated that meta-learning can effectively improve the performance and generalization in epitope immunogenicity of pHLA complexes, which is further confirmed through ablation experiments. In addition, meta-learning enabled the MHLAPre TT to fit the training pattern more efficiently on a new dataset, achieving better results despite the fact that the data was compressed to lose some of its features in order to maintain its coherence during the transfer learning process. This is because during the training of the MHLAPre IM model, the transferred environment is physiologically consistent and the transferring process carries a large number of a priori parameters achieving faster adaptation to the new environment. With the complexity of the TCR sequence system, which a single TCR is able to recognise more than a million peptides [33], using random pairing would generate numerous false-negative samples, reducing the model’s accuracy and generalization ability. Therefore, for the selection of negative samples, we used experimentally confirmed negative samples from the IEDB database. The ability of the model to

Despite the excellent performance of our model, several limitations should be considered. Firstly, a meta-learning paradigm based on prototype networks was used in the framework. In the future, more advanced meta-learning paradigms may become available. In addition, to prevent the errors of immunogenicity data from different databases causing the contamination for the training samples, we used the IEDB dataset alone, and only discussed HLA-A, -B and -C subtypes. In further studies, we will expand the abundance of HLA types and peptide types to provide more peptide-specific tasks for meta-learning tasks, which will greatly improve the scalability and adaptability of the model to introduce new data. The TextCNN model is used in the backbone module, and it can be replaced by a more advanced deep learning model to fully learn the biological features of the peptide-HLA-TCR triplets. This may bring better performance to a certain extent. For the pHLA-TCR data, The reason why MHLAPre TT showed a humble but not excellent performance may due to two main factors. On the one hand, to ensure the consistency with the immunogenicity data structure, the sequence length of TCRs and peptides were compressed, which interfered the learning efficiency of pHLA complexes' biological features to a certain extent. On the other hand, the immunogenicity characteristics between pHLA data and pHLA-TCR data have a natural distribution shift, bring more difficulties to comprehensively learn the latent features of epitope-TCR in transfer learning.

The hardware and software configurations for MHLAPre development were 2×24GB RAM NVIDIA RTX 3090TI GPUs, running



on Python 3.8.1 and Pytorch 1.12.1 [38] with Compute Unified Device Architecture (CUDA) version 11.7. The dataset was randomly divided into training set and test set at a ratio of 4:1. The training process used the Adam optimizer and the initial learning rate was set to  $5 \times 10^{-4}$ , and the training epochs and the batch size were set to 350 and 128 (Supplementary Table S6). A rigorous grid search was conducted to find out the suitable parameters (e.g. the number of layers, attention mechanism specifics, and loss functions) of the model with the highest predictive performance.

### Immunogenicity epitope-HLA dataset

Since our work focuses on the immunogenicity of neoantigens presented by HLA-I, our raw data are derived from experimental data on human samples. Immunogenicity epitope-HLA data for neoepitopes were obtained from the IEDB [32] (tcell\_table.xlsx file, downloaded on 26 November 2023) database, which consists 156 244 samples in total. Of these, there were 66 714 positive samples and 89 531 negative samples. The data were cleaned to remove duplicates and low-quality samples. Additionally, to better assess the impact of peptide length on the immunogenicity prediction of the model, the cleaning process retained antigen sequences with a length of up to 15 amino acids. The final dataset included 146 alleles, comprising 47 810 samples (999 (8-mers), 29 301 (9-mers), 14 339 (10-mers), 2 244 (11-mers), 98 (12-mers), 202 (13-mers), 129 (14-mers), 298 (15-mers)). The immunogenicity HLA-antigen peptide database consisted of 36 651 negative samples and 11 159 positive samples after the cleaning process.

### TCR binding specificity dataset

The TCR binding specificity dataset that record the interaction of pHLA-TCR pairs was also obtained from the IEDB database [32] (receptor\_table.xlsx file), with a total of 121 689 positive samples. After deleting samples with TCR sequence lengths greater than 30, antigen length distributions that did not satisfy the 8–15 mer range, duplicates and samples with missing data, 33 517 samples spanning 32 alleles were retained. As the overall sequence of TCR is very complex and a single TCR can recognise more than one million peptides [39]. In view of the fact that the traditional manufacturing of negative samples is done by randomly pairing the positive samples, which is prone to the emergence of a large number of false-negative samples, which affects the accuracy of the training and leads to the degradation of the model's generalisation and prediction performance. We screened in the IEDB, retained pHLA-TCR-negative samples that were laboratory-confirmed. 49 988 pairs of pHLA-TCR-negative samples were retained after screening. Since the complexity of the CDR3 sequence of the TCR, we found regular repeats of the CDR3 sequence by calculating the frequency of CDR3 sequence sites, which were visualized as a logo plot (Supplementary Fig. 11).

We acquired single-cell sequencing data from the 10X Genomics Chromium single-cell immunoassay platform and processed these data with Cell Ranger to obtain structured data. We collected data from expanded T cells consistently labelled with HLA-A\*02:01 and GILGFVFTL reagents and screened and sorted these cells for the TCR CDR3 region. A sample dataset containing 2083 TCR-pHLA pairs was obtained.

### Input embedding

In this study, we tested three different methods for encoding peptide sequences, including one-hot encoding, Blosum62 matrix [19], and Atchley factor [25]. The three amino acid coding ways

were illustrated in Supplementary Fig. 1(a–c). One-hot encoding is a numerical method to convert categorical data into binary vectors. It generates a new column for each category and specifies a value of 1 or 0 to indicate the presence or absence of that category. One-hot encoding is often used to represent the statements in a state machine, thereby eliminating the need for a decoder. Atchley factors, proposed by William R. Atchley in 2005, also known as Atchley parameters or Atchley amino acid descriptors, is a set of numerical values used to quantitatively represent various biochemical properties of amino acids. Blosum62 is a protein substitution matrix for protein sequence alignment. This matrix was derived from Protein Data Bank [40] analysis and measures substitution probabilities between different amino acids. Specifically, each element in the Blosum62 matrix represents the substitution probability between two different amino acids in a protein family, as determined by variant analysis of actual protein sequences.

We selected the Blosum62 matrix as the coding matrix for amino acid sequences. Due to the Blosum62 matrix has more information about amino acid features than the other two coding methods, and it is easier to identify relevant interaction features in the high-dimensional space. By integrating the coding for each HLA I epitope sequence, Blosum62 enriches the features between data points. This is advantageous in model training, because it enabled deeper representations to be explored in a higher dimensional space. Additionally, it maintains the generalization ability of the model to a certain extent.

### Performance evaluation

In this paper, samples that had been experimentally verified and proven to be immunogenic are regarded as positive samples, and that were not immunogenic or unable to generate immune responses were defined as negative samples. Therefore, we used the same training data and test data to validate our method and existing advanced models. We calculated the true and false positive rates and plotted the ROC curves. TPR and FPR are defined as follows, where TP and TN are the number of positive and negative samples successfully identified, respectively.

$$TPR = \frac{TP}{TP + FN}, FPR = \frac{FP}{TN + FP} \quad (1)$$

Where FN (FP) is the number of negative (positive) samples that were incorrectly identified. AUROC was used to evaluate the predictive performance of the model. Since the number of negative samples is much larger than the number of positive samples, the AUPRC is more representative in assessing the overall performance of prediction methods in this case. Precision is the percentage of correctly identified positive samples relative to those judged to be positive, PPV has the same meaning as precision and Recall is the same as TPR. Precision and Recall were defined as follows:

$$Precision = \frac{TP}{TP + FP}, Recall = \frac{TP}{TP + FN} \quad (2)$$

All evaluation metrics were computed by the scikit-learn python package [41].

## HLA sequence residue binding frequency

The polymorphic HLA sequence from the HLA-A, -B, and -C alleles is encoded by a pseudosequence consisting of 34 amino acid residues that are most likely to contact with the peptide. Since the central antigenic peptide residues can interact with different subsets of residues within the binding groove, there are multiple possible binding conformations, and the residues that make up these conformations are included in the pseudosequence [11]. The interaction mapping between the peptide and the HLA sequences is shown in [Supplementary Fig. 1d](#). As shown in [Fig. 3](#) of the Supplementary Material, we hierarchically analyzed the frequency of each amino acid at the 34-residue position for the HLA alleles. According to the differences of the HLA allele types in the available immunogenicity data, we can find that the binding grooves of the three HLA loci to antigenic peptide residues have a relatively consistent distribution. Generally, for class I adaptive immune, the epitope immunogenicity of pHLA complexes have a regular distribution to support the further use of neural networks to explore the high-level characteristics of immunogenic pHLA pairs.

## MHLAPre access HLA anchor sites at the molecular level of antigen

The MHLAPre model uses pseudocoding sequences in the HLA coding stage. The 34-bit pseudocoding sequence contains most variety anchor sites in binding groove for all HLA I antigenic peptides. This enables MHLAPre to be more sensitive to class I HLA molecules and antigens at the molecular level rather than one-hot coding directly for HLA names as in most studies. Therefore, MHLAPre can make full use of the positional traits of HLA allelic anchor sites when studying pan-peptide binding to HLA, and can combine with high-throughput sequencing techniques to realize in-depth exploration of spatial morphology of large-flux data.

## HLA alleles were consistent in the distribution of antigens with different lengths

The study of antigen-binding HLA allele-specific preferences plays a crucial role in antigen presentation. Considering this factor, We compare the frequency of antigenic sequences with different length distributions at the binding site of the same allele ([Fig. 1d](#)). We selected three HLA genes in the HLA-A, -B, and -C allele classes (HLA-A\*02:01, HLA-B\*07:02, HLA-C\*24:02), generation of binding antigen signature maps for HLA-I using ggseqlog [42] ([Fig. 1b](#)). Based on the logo plot generated from our three HLA allele binding antigen data, it can be shown that antigenic sequences of different lengths bound to the same HLA allele are highly similar and consistent.

To validate this idea even further, we performed experimental analyses for all existing HLA allele types and the antigenic sequences it binds with different length distributions, and generated antigenic sequence signature maps with different length distributions ([Supplementary Figs 4–10](#)). Experimental results confirm this conjecture. And we found that some of the homologous alleles have a certain preference for some of the binding sites when binding to antigens of the same length distribution. For example, at antigen lengths of 9-mer, there is a much higher frequency of Leucine at amino acid position 2 and Lysine, Arginine, and Valine at amino acid position 9 in the antigenic sequences bound by the HLA-A class allele than the other amino acids.

## HLA attention

The MHLAPre network architecture uses multi-head self-attention mechanism with a head number of 10 in the preprocessing of the data using the Transformer-Encoder. The input to the preprocessing is a concatenated HLA-peptide encoding, and both the output and input data are in the dimension of  $50 \times 21$ , where the first  $34 \times 21$  matrix represents the HLA sequence encoding. We output the pHLA attentional weights from the preprocessing process, and intercepted the first 34 columns of weight information representing the HLA sequences, with the attentional weight data of  $34 \times 50$  for each HLA sequence. The columns representing each amino acid position in the HLA were then averaged to obtain the attention-related weights of the HLA alleles in this sample. In this way, we obtained the HLA sequence of each sample, and then according to the positive or negative classification of the sample, the same HLA allele was averaged for the amino acid attention at the same position and plotted as a heat map. The average attention for each pseudosequence position of each allele in the IEDB dataset is shown in [Supplementary Fig. 14](#). It can be clearly seen that the first half of the pseudosequence of the positive samples gained more attention weight, and the overall attention weight averages are greater than those of the negative samples. Therefore, we are able to obtain a visual interpretation of the amino acid residue position information, which has an important impact on sample classification, based on the attention weight information.

## Meta-learning module

Different from the traditional supervised deep learning method of training, we introduce a meta-learning strategy to optimize the learning parameters for the prediction of epitope immunogenicity. The Model-Agnostic Meta-Learning framework [43] is utilized to adapt the peptide-specific task distribution  $p(Q)$  after optimal training of the meta-learning architecture, where  $Q$  is a sampled peptide-specific task. MAML is independent of the structure and form of supervised deep learning models. It allows quick convergence on the current task after a small number of iterations, and can be parameterized only for the assumption that the model is composed of a parameter vector  $\theta$ . Supervised deep learning model with MAML can be thought of as a function  $f_\theta$  with parameter  $\theta$ . The main objective of Meta-learner is to learn the parameter  $\theta_{meta}$  in the meta information, to quickly adapt the peptide task distribution according to  $p(Q)$ . During meta-training, the model samples a task  $Q_i = \{S_i, T_i\} \subset D_{train}$ , where  $S_i$  denotes the support set and  $T_i$  denotes the query set, the set of tasks obtained from each sampling is trained as meta-learning training samples. Based on the support set  $S_i$ , the model computes predictions for the query sequence  $T_i$  using the generated prototype.

During the training process of the meta-learning module, we construct a peptide dataset for each outer loop peptide-specific task. Since the distribution of the data obeys the long-tailed distribution, each peptide binds a different number of HLAs. Thus, a large amount of binding information is lost when building each peptide-specific pHLA. PanPep's dynamic sampling approach provides us with inspiration, i.e., the support and query sets for each peptide-associated pHLA or pHLA-TCR are randomly re-selected during the training process. This approach mitigates the impact of unbalanced data on model training. We choose the cross-entropy loss function to calculate the loss value:

$$\Psi = - \sum_{i=1}^n [y_i \ln \hat{y}_i + (1 - y_i) (\ln(1 - \hat{y}_i))] \quad (3)$$

where  $y_i$  represents the label of samples and  $\hat{y}_i$  is the prediction score of the corresponding samples.

$$\theta_i^j = \theta_i^{j-1} - \varphi \cdot \nabla \theta_i^{j-1} \Psi_{hla'_i, y_i \sim S_i} [f_{\theta_i^{j-1}}(IM_i, hla'_i), y'_i] \quad (4)$$

where  $\theta_i^j$  represents the parameter of the peptide-specific learner after three inner loops on task  $Q_i$ , and  $\varphi$  is the learning rate of the inner loop.  $j$  represents the number of inner loops. When  $j=0$ ,  $\theta_i^0 = \theta_{meta}$ .  $IM_i$  is the immunogenic antigenic peptide  $i$  embedding. The peptide-specific learner uses  $(IM_i, hla'_i)$  as input, which represents the pair of peptide  $IM_i$  and HLA  $i$  encoding of task  $Q_i$ .  $y'_i$  is the label indicating whether HLA  $i$  binds to peptide  $i$ . The model parameters are trained by optimizing for the performance of  $f_{\theta_i^{j-1}}$  with respect to  $\theta$  across tasks sampled from  $p(Q)$ . More concretely, the meta-objective is as follows:

$$\begin{aligned} \min_{\theta} \sum_{Q_i \sim p(Q)} \Psi_{hla'_i, y_i \sim S_i} f_{\theta_i^j} \\ = \sum_{Q_i \sim p(Q)} \Psi_{hla'_i, y_i \sim S_i} (f_{\theta_i^{j-1}} - \varphi \cdot \nabla \theta \Psi_{hla'_i, y_i \sim S_i} (f_{\theta_i^{j-1}})) \end{aligned} \quad (5)$$

The meta-learner parameter  $\theta_{meta}$  is trained by the  $f_{\theta_i}$  on the query set  $T_i$  with respect to across tasks from  $p(Q)$ . Meta-learner optimizes the parameter  $\theta_{meta}$  to adapt to the new task through the gradient steps of the inner loop of the new task. The update of cross-task meta-learner parameter is an outer loop. In this study, we use the Adam to optimize the meta-learner, and the next update of the meta-learning parameter  $\theta_{meta}$  is as follows:

$$\begin{aligned} \theta_{meta} &\leftarrow \theta_{meta} - \\ \varphi' \cdot \nabla \theta_{meta} \sum_{Q_i \sim p(Q)} &\left[ \Psi_{hla'_i, y_i \sim S_i} [f_{\theta_i^j}(IM_i, hla'_i), y_i] \right] \end{aligned} \quad (6)$$

where  $\varphi'$  is the learning rate of the outer loop. Therefore, the updates of the meta-learner parameter involve the high-order derivative of the gradient. Algorithm 1 describes the complete procedure of meta-training.

---

#### Algorithm 1. Meta-training of MHLAPre

---

**Input:** Peptide-HLA/TCR-pHLA task samples, denoted by:

$$Q = \{Q_1, Q_2, Q_3, \dots\}$$

**Output:** The meta parameters  $\theta_i^j$

- 1: Initialize the parameter and learning-rate:  $\theta_{meta}, \varphi, \varphi'$ .
  - 2: **while** not done **do**:
  - 3:   Sample batch of task  $Q_i \in p(Q)$
  - 4:   **for all** **do**:
  - 5:     Obtaining the parameters specific to the task  $Q_i$  by the process of gradient descent:  

$$\theta_i^j = \theta_i^{j-1} - \varphi \cdot \nabla \theta_i^{j-1} \Psi_{hla'_i, y_i \sim S_i} [f_{\theta_i^{j-1}}(IM_i, hla'_i), y'_i]$$
  - 6:   **end for**
  - 7:   Update the meta parameters via gradient descent:  

$$\theta_{meta} \leftarrow \theta_{meta} - \varphi' \cdot \nabla \theta_{meta} \sum_{Q_i \sim p(Q)} [\Psi_{hla'_i, y_i \sim S_i} [f_{\theta_i^j}(IM_i, hla'_i), y_i]]$$
  - 8: **end while**
  - 9: **output**  $\theta_i^j$
- 

### Detailed meta-learner architecture of MHLAPre

MHLAPre used pHLA pairs as input and calculates the final binding probability. Each pHLA pair was encoded by a Blosom62

matrix into a  $50 \times 21$  matrix, denoted as pHLA matrix, considering the HLA sequence and typing information and antigenic peptides. Specifically, antigenic peptides and HLA sequences were converted into  $16 \times 21$  matrices and  $34 \times 21$  matrices, and reserving  $30 \times 21$  matrix space in advance for the TCR CDR3 sequences that are later transfer learning into the pHLA-TCR data. We use sinusoidal positional encoding to add positional information for each amino acid in the matrix. This encoding helps the model understand the relative distance and sequence information between amino acid positions within a protein, and can more clearly direct the attention of the model based on position. The pHLA matrix, enriched with positional information, was then fed into the Transformer encoder as input. Here, we set the number of heads in the multi-head self-attention mechanism as 10 to capture the global relationship between the matrices. After three layers of the Transformer encoder, a matrix with the same structure as the original pHLA is obtained, and each amino acid has global information from the context. The resulting matrix was used as input for the following TextCNN, which was composed of three one-dimensional convolutional neural networks with ReLU activation function and one-dimensional maximum pooling to mine the latent biological features. The output of TextCNN was fed into a tri-layer fully connected neural network with a 0.2 dropout to prevent overfitting. Finally, the softmax function was applied to calculate the binding probability, representing the likelihood of a specific pHLA pair forming a stable complex.

#### Key Points

- In this study, we proposed a deep learning framework based on a meta-learning framework, named MHLAPre, to predict the epitope immunogenicity of tumor-associated antigens that can be present by HLA molecules and elicit immune responses. The results show that our model outperforms existing SOTA models in predicting immunogenic HLA antigen affinity and still exhibits excellent performance after transfer learning based on pHLA-TCR data, which demonstrates its effectiveness and robustness.
- Based on the prediction tasks of peptide-HLA affinity and T-cell receptor interaction, in which physiological processes and training data are highly repetitive and correlated, we connected the two tasks using transfer learning. According to the experimental comparison, the proposed MHLAPre is superior to those of the existing advanced predictors.
- Based on the model-independent MAML architecture, the proposed method achieved robust performance using TextCNN deep learning models, and better performance can be expected with more advanced learning model in the future.
- MHLAPre is the first proposal and preliminary demonstration of feasibility for transfer learning from pHLA data to pHLA-TCR data.

### Acknowledgments

The authors thank the reviewers for their valuable comments and suggestions.

## Supplementary data

Supplementary data is available at Briefings in Bioinformatics online.

## Funding

The authors acknowledge support from the National Key R and D Program of China (2022YFF1202100), National Natural Science Foundation of China (Nos. 62202092, 62272135, and 62372135), and the King Abdullah University of Science and Technology (KAUST) Office of Research Administration (ORA) under Award No REI/1/5234-01-01, REI/1/5414-01-01, REI/1/5289-01-01, REI/1/5404-01-01, REI/1/5992-01-01, URF/1/4663-01-01, Center of Excellence for Smart Health (KCSH), under award number 5932, and Center of Excellence on Generative AI, under award number 5940.

## Data availability

MHLAPre is available on github (<https://github.com/ChanganMakeYi/MHLAPre>), HLA allele immunogenicity data were obtained from the IEDB database (<https://www.iedb.org/>), VDJdb (<https://vdjdb.cdr3.net/>), and McPAS-TCR (<http://friedmanlab.weizmann.ac.il/McPAS-TCR/>). The detailed information of the 10x Genomics cohort is available at: (<https://www.10xgenomics.com/datasets/cd-8-plus-t-cells-of-healthy-donor-1-1-standard-3-0-2>).

## Author contributions statement

L.X. contributed to the methodology, modeling, data analysis, software, and writing the original draft. Q.Y., L.X., and W.D. contributed to modeling, data analysis and software, and writing the original draft. K.W. and G.L. contributed to experimental design, methodology, data analysis, and writing the original draft. S.D., T.Y., X.Z., X.L., X.G., and G.W. contributed to the data analysis and experimental design. All authors reviewed the final manuscript.

## References

- Roemer MG, Advani RH, Redd RA. et al. Classical Hodgkin lymphoma with reduced  $\beta$ 2M/MHC class I expression is associated with inferior outcome independent of 9p24.1 status. *Cancer Immunol Res* 2016;**4**:910–6. <https://doi.org/10.1158/2326-6066.CIR-16-0201>.
- Garrido F, Aptsiauri N. Cancer immune escape: MHC expression in primary tumours versus metastases. *Immunology* 2019;**158**: 255–66. <https://doi.org/10.1111/imm.13114>.
- Hu Y, Wang Z, Hu H. et al. ACME: pan-specific peptide-MHC class I binding prediction through attention-based deep neural networks. *Bioinformatics* 2019;**35**:4946–54. <https://doi.org/10.1093/bioinformatics/btz427>.
- Jensen KK, Andreatta M, Marcatili P. et al. Improved methods for predicting peptide binding affinity to MHC class II molecules. *Immunology* 2018;**154**:394–406. <https://doi.org/10.1111/imm.12889>.
- Bassani-Sternberg M, Chong C, Guillaume P. et al. Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allosteric regulating HLA specificity. *PLoS Comput Biol* 2017;**13**:e1005725. <https://doi.org/10.1371/journal.pcbi.1005725>.
- Gfeller D, Guillaume P, Michaux J. et al. The length distribution and multiple specificity of naturally presented HLA-I ligands. *J Immunol* 2018;**201**:3705–16. <https://doi.org/10.4049/jimmunol.1800914>.
- Peters B, Nielsen M, Sette A. T cell epitope predictions. *Annu Rev Immunol* 2020;**38**:123–45. <https://doi.org/10.1146/annurev-immunol-082119-124838>.
- He Q, Jiang X, Zhou X. et al. Targeting cancers through TCR-peptide/MHC interactions. *J Hematol Oncol* 2019;**12**:139. <https://doi.org/10.1186/s13045-019-0812-8>.
- Yamamoto TN, Kishton RJ, Restifo NP. Developing neoantigen-targeted T cell-based treatments for solid tumors. *Nat Med* 2019;**25**:1488–99. <https://doi.org/10.1038/s41591-019-0596-y>.
- Reynisson B, Alvarez B, Paul S. et al. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res* 2020;**48**:W449–54. <https://doi.org/10.1093/nar/gkaa379>.
- Nielsen M, Lundegaard C, Blicher T. et al. NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. *PLoS One* 2007;**2**:e796. <https://doi.org/10.1371/journal.pone.0000796>.
- Bhattacharya R, Sivakumar A, Tokheim C. et al. Evaluation of machine learning methods to predict peptide binding to MHC class I proteins. *bioRxiv*. 2017;154757.
- Shao XM, Bhattacharya R, Huang J. et al. High-throughput prediction of MHC class I and II Neoantigens with MHCnuggets. *Cancer Immunol Res* 2020;**8**:396–408. <https://doi.org/10.1158/2326-6066.CIR-19-0464>.
- O'Donnell TJ, Rubinsteyn A, Bonsack M. et al. MHCflurry: open-source class I MHC binding affinity prediction. *Cell Syst* 2018;**7**:129–132.e4. <https://doi.org/10.1016/j.cels.2018.05.014>.
- Gfeller D, Schmidt J, Croce G. et al. Improved predictions of antigen presentation and TCR recognition with MixMHCpred2.2 and PRIME2.0 reveal potent SARS-CoV-2 CD8+ T-cell epitopes. *Cell Syst* 2023;**14**:72–83.e5. <https://doi.org/10.1016/j.cels.2022.12.002>.
- Bassani-Sternberg M, Chong C, Guillaume P. et al. Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allosteric regulating HLA specificity. *PLoS Comput Biol* 2017;**13**:e1005725. <https://doi.org/10.1371/journal.pcbi.1005725>.
- Schmidt J, Smith AR, Magnin M. et al. Prediction of neo-epitope immunogenicity reveals TCR recognition determinants and provides insight into immunoediting. *Cell Rep Med* 2021;**2**:100194. <https://doi.org/10.1016/j.xcrm.2021.100194>.
- Sarkizova S, Klaeger S, Le PM. et al. A large peptidome dataset improves HLA class I epitope prediction across most of the human population. *Nat Biotechnol* 2020;**38**:199–209. <https://doi.org/10.1038/s41587-019-0322-9>.
- Jokinen E, Huuhtanen J, Mustjoki S. et al. Predicting recognition between T cell receptors and epitopes with TCRGP. *PLoS Comput Biol* 2021;**17**:e1008814. <https://doi.org/10.1371/journal.pcbi.1008814>.
- Giancarlo C, Sara B, Dana M. et al. Deep learning predictions of TCR-epitope interactions reveal epitope-specific chains in dual alpha T cells. *Nature Communications* 2024;**15**:3211.
- Montemurro A, Schuster V, Povlsen HR. et al. NetTCR-2.0 enables accurate prediction of TCR-peptide binding by using paired TCR $\alpha$  and  $\beta$  sequence data. *Commun Biol* 2021;**4**:1060. <https://doi.org/10.1038/s42003-021-02610-3>.
- Lu T, Zhang Z, Zhu J. et al. Deep learning-based prediction of the T cell receptor-antigen binding specificity. *Nat Mach Intell* 2021;**3**: 864–75. <https://doi.org/10.1038/s42256-021-00383-2>.



23. Xu Z, Luo M, Lin W. et al. DLpTCR: An ensemble deep learning framework for predicting immunogenic peptide recognized by T cell receptor. *Brief Bioinform* 2021;**22**:bbab335. <https://doi.org/10.1093/bib/bbab335>.
24. Springer I, Besser H, Tickotsky-Moskovitz N. et al. Prediction of specific TCR-peptide binding from large dictionaries of TCR-peptide pairs. *Front Immunol* 2020;**11**:1803. <https://doi.org/10.3389/fimmu.2020.01803>.
25. Gao Y, Gao Y, Fan Y. et al. Pan-peptide meta learning for T-cell receptor–antigen binding recognition. *Nat Mach Intell* 2023;**5**: 236–49. <https://doi.org/10.1038/s42256-023-00619-3>.
26. Santoro A, Bartunov S, Botvinick M. et al. Meta-learning with memory-augmented neural networks. In: *International Conference on Machine Learning* 2016;**48**:1842–50.
27. Chu Y, Zhang Y, Wang Q. et al. A transformer-based model to predict peptide–HLA class I binding and optimize mutated peptides for vaccine design. *Nat Mach Intell* 2022;**4**:300–11. <https://doi.org/10.1038/s42256-022-00459-7>.
28. Vaswani A, Shazeer N, Parmar N. et al. Attention is all you need. *Adv Neural Inf Process Syst* 2017;**30**:1–11.
29. Dong W, Yang Q, Wang J. et al. Multi-modality attribute learning-based method for drug-protein interaction prediction based on deep neural network. *Brief Bioinform* 2023;**24**:bbad161. <https://doi.org/10.1093/bib/bbad161>.
30. Li Z, Jin J, He W. et al. CoraL: interpretable contrastive meta-learning for the prediction of cancer-associated ncRNA-encoded small peptides. *Brief Bioinform* 2023;**24**:bbad352. <https://doi.org/10.1093/bib/bbad352>.
31. Guo B, Zhang C, Liu J, et al. Improving text classification with weighted word embeddings via a multi-channel TextCNN model[J]. *Neurocomputing* 2019;**363**: 366–374.
32. Vita R, Mahajan S, Overton JA. et al. The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res* 2019;**47**: D339–43. <https://doi.org/10.1093/nar/gky1006>.
33. Lien L, Steve L, Bruno F. et al. Challenges in neoantigen-directed therapeutics. *Cancer Cell* 2023;**41**:15–40. <https://doi.org/10.1016/j.ccell.2022.10.013>.
34. Lv Q, Chen G, Yang Z. et al. Meta learning with graph attention networks for low-data drug discovery. *IEEE Trans Neural Netw Learn Syst* 2023;**35**:11218–30. <https://doi.org/10.1109/TNNLS.2023.3250324>.
35. Brbić M, Zitnik M, Wang S. et al. MARS: discovering novel cell types across heterogeneous single-cell experiments. *Nat Methods* 2020;**17**:1200–6. <https://doi.org/10.1038/s41592-020-00979-3>.
36. Yaqing W, Quanming Y, Kwok James T. et al. Generalizing from a few examples: a survey on few-shot learning. *ACM Comput Surv (CSUR)* 2020;**53**:1–34.
37. Albert BA, Yang Y, Shao XM. et al. Deep neural networks predict class I major histocompatibility complex epitope presentation and transfer learn neoepitope immunogenicity. *Nat Mach Intell* 2023;**5**:861–72. <https://doi.org/10.1038/s42256-023-00694-6>.
38. Adam P, Sam G, Francisco M. et al. PyTorch: an imperative style, high-performance deep learning library. *Adv Neural Inf Process Syst* 2019;**32**:1–12.
39. Wooldridge L, Ekeruche-Makinde J, van den Berg HA. et al. A single autoimmune T cell receptor recognizes more than a million different peptides. *J Biol Chem* 2012;**287**:1168–77. <https://doi.org/10.1074/jbc.M111.289488>.
40. wwPDB consortium. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res* 2019;**47**:D520–8. <https://doi.org/10.1093/nar/gky949>.
41. Fabian P, Gal V, Alexandre G. et al. ggseqlogo: a versatile R package for drawing sequence logos. *J Mach Learn Res* 2011;**12**: 2825–30.
42. Wagih O. ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics* 2017;**33**:3645–7. <https://doi.org/10.1093/bioinformatics/btx469>.
43. Lien L, Steve L, Bruno F. et al. Challenges in neoantigen-directed therapeutics. *Cancer Cell* 2023;**41**:15–40.