PLOS Computational Biology

RESEARCH ARTICLE

DSCA-HLAII: A dual-stream cross-attention model for predicting peptide–HLA class II interaction and presentation

Ke Yan[1,2], Hongjun Yu[1], Shutao Chen[1], Alexey K. Shaytan[3,4], Bin Liu[1,2,5]*, Youyu Wang[6]*

1 School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China, **2** Zhongguancun Academy, Beijing, China, **3** Department of Biology, Lomonosov Moscow State University, Moscow, Russia, **4** Centre for Biomedical Research and Technology, AI and Digital Sciences Institute, Faculty of Computer Science, HSE University, Moscow, Russia, **5** SMBU-MSU-BIT Joint Laboratory on Bioinformatics and Engineering Biology, Shenzhen MSU-BIT University, Shenzhen, Guangdong, China, **6** Department of Thoracic Surgery, Sichuan Academy of Medical Sciences and Sichuan Provincial People's Hospital, University of Electronic Science and Technology of China, Chengdu, Sichuan Province, China

* bliu@bliulab.net (BL); syywyy123@126.com (YW)

OPEN ACCESS

Citation: Yan K, Yu H, Chen S, Shaytan AK, Liu B, Wang Y (2026) DSCA-HLAII: A dual-stream cross-attention model for predicting peptide–HLA class II interaction and presentation. PLoS Comput Biol 22(1): e1013836. https://doi.org/10.1371/journal.pcbi.1013836

Editor: Lun Hu, Xinjiang Technical Institute of Physics and Chemistry, CHINA

Received: October 21, 2025

Accepted: December 12, 2025

Published: January 2, 2026

Data availability statement: The source code and data used to produce the results presented in this manuscript are available from the GitHub repository: https://github.com/cokeyk/DSCA-HLAII.

Funding: This research was supported by the Beijing Natural Science Foundation

Abstract

Motivation

The interaction between peptides and human leukocyte antigen class II (HLA-II) molecules plays a pivotal role in adaptive immune responses, as HLA-II mediates the recognition of exogenous antigens and initiates T cell activation through peptide presentation. Accurate prediction of peptide-HLA-II binding serves as a cornerstone for deciphering cellular immune responses, and is essential for guiding the optimization of antibody therapeutics. Researchers have developed several computational approaches to identify peptide-HLA-II interaction and presentation. However, most computational approaches exhibit inconsistent predictive performance, poor generalization ability and limited biological interpretability.

Results

In this study, we present DSCA-HLAII, a novel predictive framework for peptide-HLA-II interactions and presentation based on a dual-stream cross-attention architecture. The framework proposes a dual-stream cross-attention (DSCA) mechanism to integrate pre-trained semantic embedding ESMC with sequence-level ONE-HOT features. The DSCA mechanism effectively models the interaction dynamics between peptides and HLA-II molecules, enabling the precise identification of key binding sites. Experimental results demonstrate that DSCA-HLAII consistently surpasses existing state-of-the-art approaches, demonstrating high accuracy and robustness in predicting peptide-HLA-II interactions and presentation. We further demonstrate the capability of DSCA-HLAII for predicting peptide binding cores and assessing antibody

(No. L232067) to K.Y., the National Natural Science Foundation of China (No. 62271049 to B.L., 62473049 to K.Y., U22A2039 to B.L.), the Zhongguancun Academy (Project No. 20240101) to B.L., and HSE basic research program to A.K.S.. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

immunogenicity, which is expected to advance artificial intelligence-based peptide drug discovery.

Author summary

This paper proposes a novel predictive framework for peptide-HLA-II interactions and presentation based on a dual-stream cross-attention architecture (DSCA-HLAII). DSCA-HLAII is a unified predictive framework that integrates sequence-based ONE-HOT features with pre-trained semantic embeddings (ESMC) to construct a comprehensive hybrid representation of peptides and full-length HLA-II sequences. By introducing a Dual-Stream Cross-Attention (DSCA) module, the model enables fine-grained characterization of peptide–HLA-II interactions and assigns differential scores across sequence positions, thereby improving the identification of critical binding sites and enhancing generalization. DSCA-HLAII simultaneously predicts peptide presentation probability and binding core location, and further supports systematic assessment of antibody immunogenicity risk. Extensive experiments on multiple warm-start test datasets and cold-start test datasets demonstrate that DSCA-HLAII surpasses existing state-of-the-art methods in both accuracy and robustness. Additionally, a publicly accessible web server (http://bliulab.net/DSCA-HLAII) has been established to facilitate practical application.

1. Introduction

Major histocompatibility complex (MHC) molecules hold a central position in adaptive immunity. In humans, the MHC molecules are known as human leukocyte antigens (HLAs). HLAs are divided into two primary classes: class I (HLA-I) and class II (HLA-II). Typically, HLA-II molecules are divided by three highly polymorphic loci: HLA-DR, HLA-DP, and HLA-DQ [1]. These isotypes form heterodimers composed of an α chain and a β chain protein [2]. Exogenous antigenic peptides are loaded onto HLA-II molecules, and the resulting complexes are transported to the cell surface for presentation. The binding affinity between a peptide and an HLA-II molecule is primarily determined by interactions within the peptide's binding core [3].These complexes are recognized by $CD4^+$ helper T cells, thereby inducing both humoral and cellular immune responses [4,5]. Therefore, the specificity and strength of peptide–HLA-II interactions directly influence which peptides can be effectively presented and recognized. Due to the highly complex and polymorphic nature of HLA-II molecules, accurately predicting peptide presentation and assessing their immunogenic potential remain critical challenges in immunology research and clinical immunotherapy.

In previous studies, researchers employed *in vitro* experimental techniques such as mass spectrometry and immunopeptidomics to investigate the binding and presentation mechanisms of peptides by HLA-II molecules [6–8]. These efforts

have provided abundant reliable data and prior knowledge for subsequent research. However, *in vitro* experiments are costly and time-consuming, and thus cannot achieve comprehensive coverage of such interactions [9]. To overcome these limitations, a series of pan-specific computational tools have been developed and applied to the prediction of peptide-HLA-II binding and presentation, becoming essential methods in this field [10–12]. These pan-specific tools construct computational models to predict peptide–HLA-II interactions for both known alleles and previously unseen alleles [13]. The strategy effectively alleviates the constraints caused by limited data for certain alleles and offers broader applicability. Racle et al. proposed a probabilistic method, MixMHC2pred, which employs the Expectation–Maximization (EM) algorithm to learn probabilistic graphical models of multiple motifs [14]. Nilsson et al. introduced NetMHCIIpan, a multilayer perceptron (MLP)-based model for predicting peptide–HLA-II binding and presentation [15]. Thrift et al., leveraging prior knowledge from AlphaFold2-multimer, developed a graph convolutional network model Graph-pMHC that explicitly captures peptide–HLA-II interactions [16,17]. Wang et al. constructed TripHLApan, a deep neural network integrating gated recurrent units (GRUs) and attention mechanisms to predict peptide–HLA-II binding [18]. Chang et al. designed CapHLA, a hybrid network to jointly predict peptide–HLA-II binding affinity and presentation probability [19].

Despite the remarkable progress made in peptide-HLA-II interactions and presentation studies in recent years, there remain several disadvantages [20–22]. Firstly, previous studies typically represented HLA-II molecules by extracting specific residues from resolved allele structures as pseudo sequences [23,24]. Peptide sequences and pseudo sequences of HLA-II molecules were then represented by hand-crafted features, such as amino acid composition and physicochemical properties [3,23,25–29]. However, such representations fail to capture the contextual dependencies and global semantic information across residues, which are essential for accurate identifying peptide-HLA-II binding [30–34]. Secondly, the key sites directly influence the interactions between peptides and HLA-II molecules. These key sites are typically located in specific regions, such as the peptide binding cores and certain pockets of HLA-II molecules [35]. Accurately identifying key sites such as the binding core not only reflects whether the model has adequately captured the underlying mechanism of peptide–HLA-II presentation, but also serves as an indicator of its interpretability and biological reliability. Moreover, accurate identification of the binding core facilitates downstream tasks such as $CD4^+$T cell recognition. Nevertheless, existing approaches often treat all residues uniformly and fail to explicitly capture the key sites which are essential for accurately characterizing the interaction between peptides and HLA-II molecules [36]. Thirdly, despite their satisfactory performance with known alleles, their models fail to generalize effectively, exhibiting poor performance for identifying unknown alleles.

To address the challenges, we propose DSCA-HLAII, an end-to-end deep learning framework based on a Dual-Stream Cross-Attention (DSCA) mechanism, designed to provide a robust and high-performance solution for predicting peptide–HLA-II presentation probabilities. The contributions of the proposed method are as follows:

(1) DSCA-HLAII integrates sequence-based ONE-HOT features [37,38] with pre-trained semantic embeddings ESMC, which is derived from large-scale unsupervised data [39]. By fusing multi-modal features into a hybrid representation of both peptide sequences and complete sequences of HLA-II chains, DSCA-HLAII achieves a more comprehensive representational space than that provided by hand-crafted features.

(2) DSCA-HLAII introduces the DSCA module to facilitate more precise modeling of peptide–HLA-II interactions and allow differential weighting across sequence sites, thereby enhancing the identification of key sites. By partially mimicking the biological interaction process, the DSCA module enhances the model's generalization ability.

(3) The ONE-HOT features capture the residue-level information, whereas the ESMC features encode the semantic and structural characteristics of the sequences. The DSCA mechanism captures the interaction information between peptides and HLA-II molecules. By leveraging fused features and integrating the DSCA mechanism, we further enhance the model's predictive capability, particularly improving its generalization performance on unknown alleles.

(4) DSCA-HLAII is capable of jointly predicting peptide presentation probability and binding core position, while also providing a systematic assessment of antibody immunogenicity risk. Comparative experiments across multiple independent test datasets demonstrate that DSCA-HLAII outperforms current state-of-the-art methods in both predictive accuracy and generalization performance, highlighting its potential value for peptide–HLA-II research and immunological applications. Finally, we developed a user-friendly web server, which has been made publicly accessible at http://bliulab.net/DSCA-HLAII.

2. Results

2.1 Comparison with the existing baseline methods on the warm-start test dataset

In this study, we employed a 5-fold cross-validation strategy to compare the performance of DSCA-HLAII with several baseline methods on the warm-start test dataset, including NetMHCIIpan-4.3 [15], CapHLA [19], TripHLApan [18], Graph-pMHC [16] and MixMHCIIPred [14].

In the warm-start test dataset, all HLA-II alleles appeared in the benchmark dataset, aiming to evaluate the predictive performance on previously seen alleles. As shown in Table 1 and Fig 1, DSCA-HLAII demonstrates superior performance than the five baseline methods in terms of AUROC, AUPR, PCC, Precision, Recall, MCC, ACC, and F1. To evaluate whether the observed performance differences were statistically significant, we conducted a bootstrap analysis on the warm-start test dataset. Compared to the strongest baseline, TripHLApan, DSCA-HLAII achieved an AUROC improvement of 0.0038, with a 95% bootstrap confidence interval of 0.0033–0.0043, indicating that the improvement is statistically significant ($p < 0.001$). On the warm-start test dataset, DSCA-HLAII demonstrates a genuine and statistically significant improvement over the strongest baseline, TripHLApan, confirming the effectiveness of the model in enhancing predictive performance.

The results show that (1) The proposed method achieves a strong linear correlation between the predicted probabilities and the ground truth labels, as evidenced by a high PCC value, confirming its reliability in generating probability estimates. (2) Unlike MixMHCIIPred and NetMHCIIpan, which extract features from HLA-II pseudo-sequences, or Graph-pMHC, which uses limited inputs, our method leverages ESMC to build intrinsic biological representations from complete chain sequences. By operating on the entire chain sequence, ESMC directly learns to decipher the intrinsic information that encode both structure and function. This provides a more comprehensive representation than pseudo sequence which offers only an incomplete sequence information. (3) Compared with CapHLA and TripHLApan adopt relatively simple network architectures and lack explicit modeling of the antigen presentation process, the proposed method utilized DSCA mechanism enables the capture of contextual information and the identification of key sites, thereby achieving explicit modeling of the interaction process between peptides and HLA-II molecules. Therefore, DSCA-HLAII achieves superior accuracy and robustness against existing state-of-the-art methods in peptide–HLA-II presentation prediction.

To further assess the predictive performance of our method across different HLA-II alleles, we conducted a per-allele comparison between DSCA-HLAII and TripHLApan [18] on the warm-start test dataset. In terms of best AUROC and

Table 1. Performance comparison on the warm-start test dataset in terms of AUROC, AUPR and PCC.

	AUROC	AUPR	PCC
Graph-pMHC	0.891	0.747	0.291
MixMHCIIPred	0.845	0.706	0.628
NetMHCIIpan	0.889	0.811	0.751
CapHLA	0.970	0.926	0.795
TripHLApan	0.984	0.955	0.889
DSCA-HLAII	**0.988**	**0.963**	**0.909**

https://doi.org/10.1371/journal.pcbi.1013836.t001

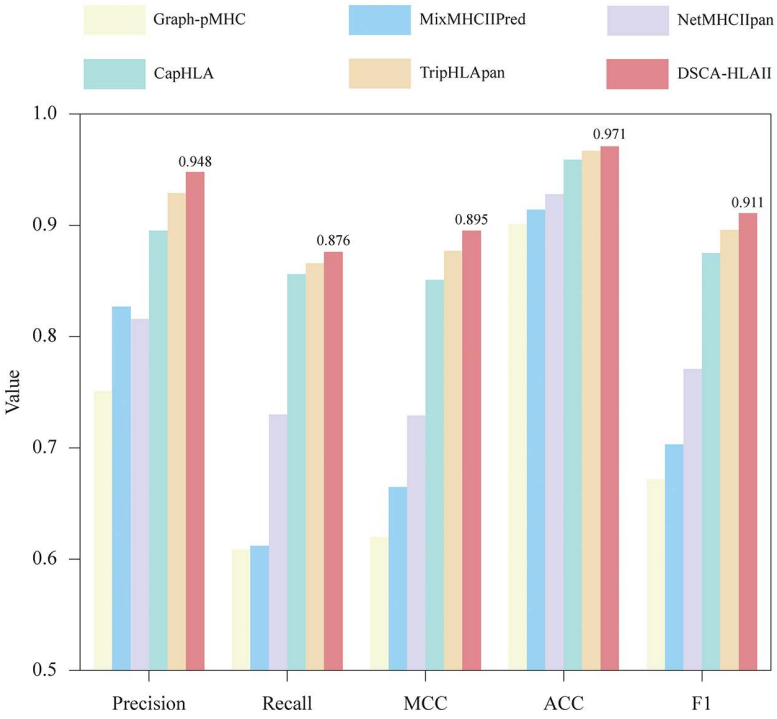

Fig 1. Performance comparison on the warm-start test dataset in terms of Precision, Recall, MCC, ACC and F1.

https://doi.org/10.1371/journal.pcbi.1013836.g001

AUPR, TripHLApan takes the second place in Table 1 by integrating triple coding matrix and transfer learning strategy. As illustrated by the allele distribution in Fig 2, DSCA-HLAII demonstrates superior performance over TripHLApan on 63 alleles. For example, for the allele DRA*01:01-DRB1*08:02, DSCA-HLAII achieves AUROC and AUPR scores of 0.926 and 0.796, whereas TripHLApan attains only 0.898 and 0.758. This performance gap can be partially attributed to the fact that TripHLApan relies solely on AAindex, Blosum62, and integer encoding as input features, while the number of samples available for this allele in the benchmark dataset is limited to 4,861. Consequently, the model struggles to capture effective peptide–HLA-II interaction patterns from such sparse data. In contrast, DSCA-HLAII leverages a fused representation of ONE-HOT and ESMC, enabling it to extract deeper biological information from large-scale sequence data and thereby improving predictive accuracy even for alleles with limited sample sizes. Therefore, DSCA-HLAII achieves consistently strong predictive performance across a wide range of alleles, rather than being effective only for a subset, underscoring its robust generalization capability.

2.2 Comparison with existing baseline methods on the cold-start test dataset

To evaluate the generalization capability of the model on unseen alleles, we compared the performance of DSCA-HLAII with several baseline methods on the cold-start test dataset. The methods included NetMHCIIpan-4.3 [15], CapHLA [19], TripHLApan [18], Graph-pMHC [16] and MixMHCIIPred [14]. In the cold-start test dataset, all HLA-II alleles are entirely absent from the benchmark dataset, specifically designed to assess the model's generalization performance on novel alleles. The experimental results are presented in Table 2 and Fig 3. Similarly, we performed a bootstrap analysis on the cold-start test dataset, which confirmed that DSCA-HLAII achieved a significant improvement over TripHLApan, with an AUROC increase of 0.0138 and a 95% bootstrap confidence interval of 0.0070–0.0202 ($p < 0.001$). On the cold-start test dataset, DSCA-HLAII similarly demonstrates a genuine and statistically significant improvement in performance.

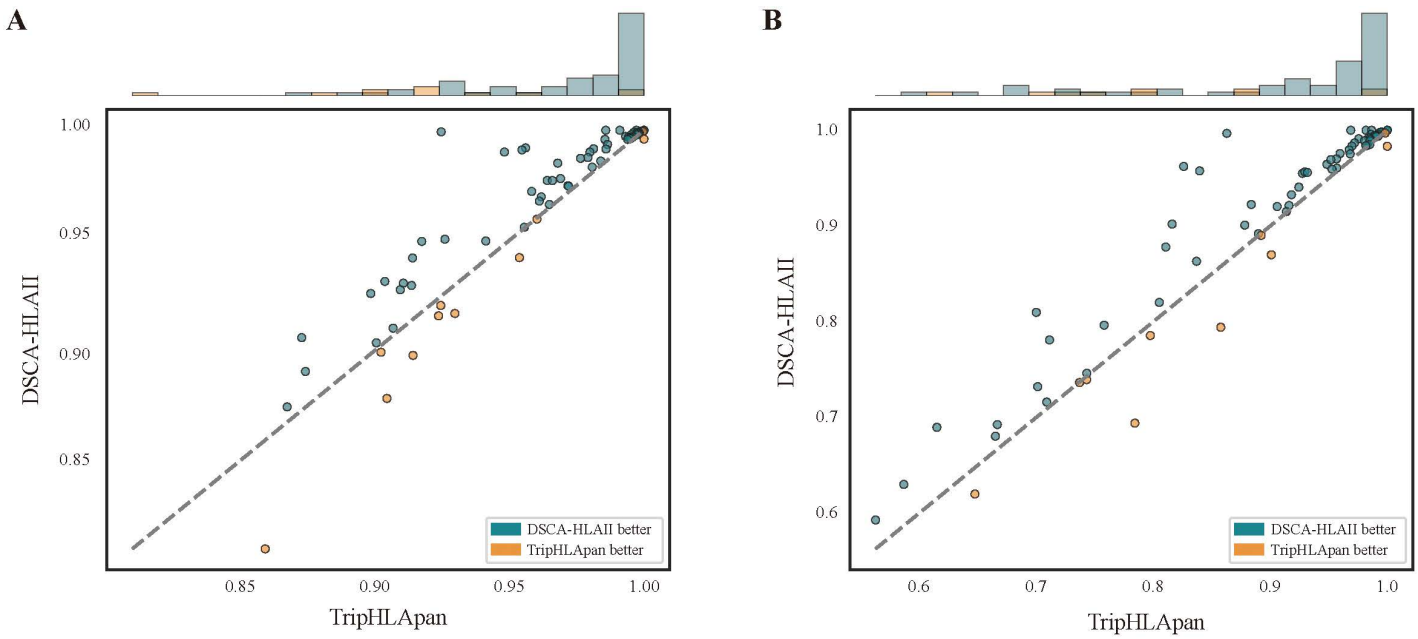

Fig 2. Performance comparison between DSCA-HLAII and TripHLApan across different alleles on the warm-start test dataset. A: The results based on AUROC. B: The results based on AUPR.

https://doi.org/10.1371/journal.pcbi.1013836.g002

Table 2. Performance comparison on the cold-start test dataset in terms of AUROC, AUPR and PCC.

	AUROC	AUPR	PCC
Graph-pMHC	0.773 (-0.118)	0.474 (-0.273)	0.236 (-0.055)
MixMHCIIPred	0.806 (-0.039)	0.586 (-0.120)	0.517 (-0.111)
NetMHCIIpan	0.882 (-0.007)	0.654 (-0.157)	0.545 (-0.206)
CapHLA	0.890 (-0.08)	0.762 (-0.164)	0.644 (-0.151)
TripHLApan	0.949 (-0.035)	0.863 (-0.092)	0.769 (-0.12)
DSCA-HLAII	**0.963 (-0.025)**	**0.894 (-0.069)**	**0.803 (-0.106)**

https://doi.org/10.1371/journal.pcbi.1013836.t002

The results indicate that: (1) On the cold-start test dataset, DSCA-HLAII outperforms all other methods across multiple metrics, including Precision, AUROC, and AUPR. Even when the dataset contains entirely novel alleles, our method maintains high predictive performance. (2) We further quantified the performance differences of each method between the cold-start and warm-start test datasets with the detailed results (AUROC, AUPR and PCC) shown in parentheses in Table 2. DSCA-HLAII proved to be the most robust method, exhibiting only a minimal performance degradation in the cold-start setting, in contrast to the significant declines seen in other methods. These findings indicate that DSCA-HLAII, based on the DSCA mechanism, maintains robust predictive performance when encountering unseen alleles.

We further compared the performance of DSCA-HLAII and TripHLApan on individual alleles within the cold-start test dataset. According to the allele distribution shown in Fig 4, DSCA-HLAII outperforms TripHLApan on 34 alleles. For example, for DRA*01:01-DRB1*14:02, DSCA-HLAII achieves AUROC and AUPR values of 0.910 and 0.799, respectively, whereas TripHLApan attains only 0.782 and 0.615. This indicates that TripHLApan experiences a substantial performance decline when predicting entirely novel alleles. In contrast, DSCA-HLAII maintains superior predictive capability even on previously unseen alleles.

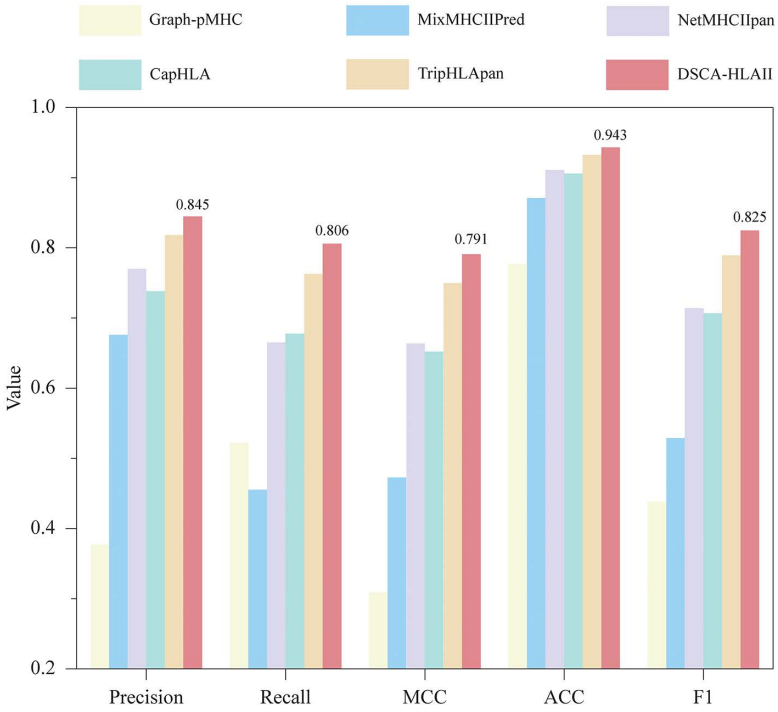

Fig 3. Performance comparison on the cold-start test dataset in terms of Precision, Recall, MCC, ACC and F1.

https://doi.org/10.1371/journal.pcbi.1013836.g003

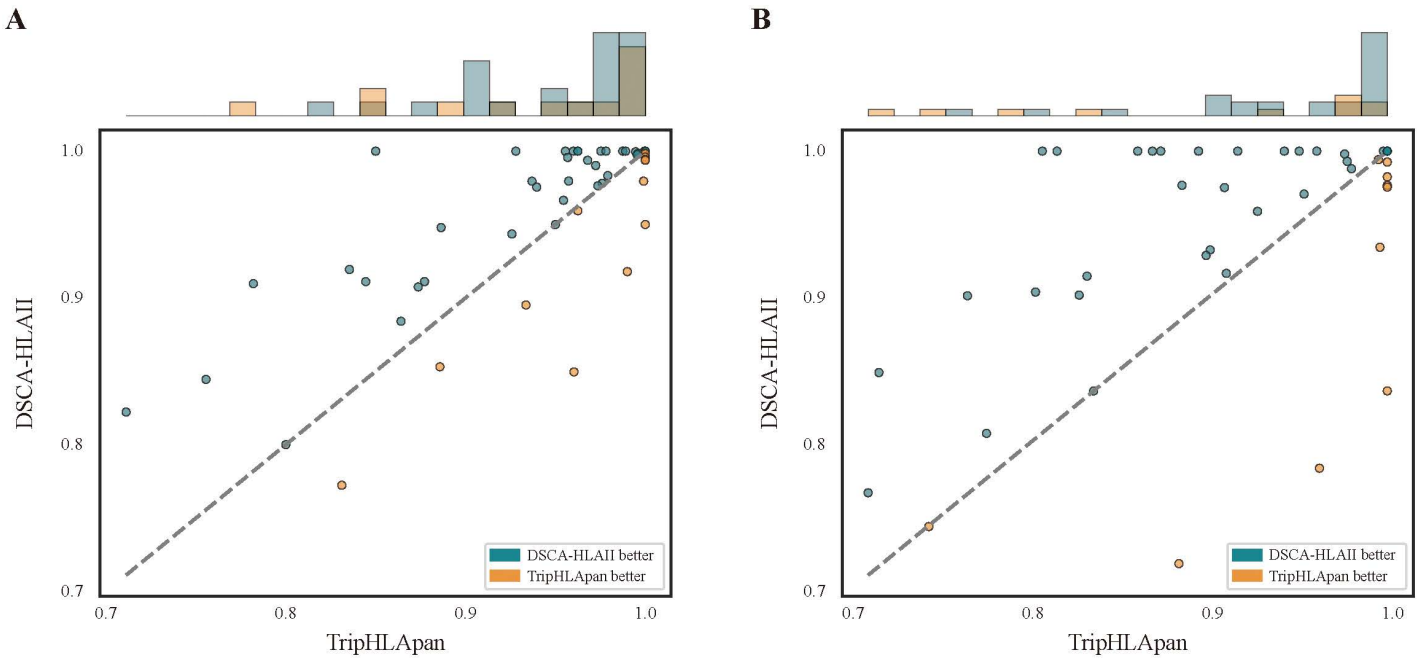

Fig 4. Performance comparison between DSCA-HLAII and TripHLApan across different alleles on the cold-start test dataset. A: The results based on AUROC. (B)The results based on AUPR.

https://doi.org/10.1371/journal.pcbi.1013836.g004

In summary, these results demonstrate that the incorporation of ESMC features enables DSCA-HLAII to achieve a more comprehensive and fine-grained representation of sequence semantics, structural characteristics, and functional attributes, even in the context of previously unobserved alleles. Furthermore, the DSCA mechanism facilitates the efficient extraction of presentation-relevant information through explicit modeling and precise identification of key residues. Therefore, the test results demonstrate that DSCA-HLAII achieves superior accuracy and robustness across all metrics compared to the baseline methods.

2.3 Ablation analysis of DSCA-HLAII

In this section, we conducted ablation experiments on the warm-start test dataset to evaluate the contributions of the fused ESMC–ONE-HOT representations and the DSCA module. The results are summarized in Table 3. Compared with Model 1, it is evident that the model can learn comprehensive semantic, structural, and functional information from ESMC, thereby enhancing predictive accuracy. Compared with Model 2, the fused representation of the complete sequences of HLA-II chains using ESMC and ONE-HOT outperforms the ONE-HOT representation of pseudo sequences. These results indicate that representing HLA-II molecules with complete sequences of chains and employing the fused representation further improves model performance. The fused representation incorporates external knowledge, such as semantic and structural information, augmenting the representation space of peptides and HLA-II molecules and enhancing the predictive capability of DSCA-HLAII. In Model 3 and Model 4, we employed other pretrained protein language models, ProtBert [40] and ProtT5 [41], respectively, to compare the performance differences between different embeddings. Compared with Model 3 and Model 4, the model using ESMC demonstrates superior performance on the warm-start test dataset. ProtBert and ProtT5 primarily learn the statistical patterns of amino acid sequences through self-supervised sequence prediction. In contrast, ESMC integrates both sequence and structural information in its training objectives. As a result, ESMC captures spatial structures, functional sites, and other biologically relevant features more effectively, which are highly relevant to the peptide–HLA-II binding mechanism. Therefore, we attribute the advantages of ESMC over other models primarily to this capability.

Compared with Model 5, the DSCA mechanism demonstrates superior performance over the conventional cross-attention mechanism, confirming its advantage and applicability in modeling peptide-HLA-II interactions and improving both representational power and predictive performance. We attribute this improvement to the dual-stream and cross-stream attention design, as well as the Query–Key–Value construction based on global and local features, all of which are absent in the conventional cross-attention mechanism.

Table 3. The ablation analysis in the DSCA on the warm-start test dataset.

Method	$Repr^a$	$Repr^b$	$Repr^c$	$Repr^d$	$Repr^e$	CA^a	CA^b	AUROC	AUPR
baseline	√					√		**0.988**	**0.963**
Model1		√				√		0.981	0.945
Model2			√			√		0.980	0.955
Model3				√		√		0.981	0.947
Model4					√	√		0.982	0.951
Model5	√						√	0.788	0.454

Note: $Repr^a$ denotes the representation of HLA-II molecules using the fused information of ESMC and ONE-HOT from the complete sequences of chains. $Repr^b$ denotes the representation using only ONE-HOT from the complete sequences of chains. $Repr^c$ denotes the representation using ONE-HOT from pseudo sequences. $Repr^d$ denotes the representation of HLA-II molecules using the fused information of ProtBert and ONE-HOT from the complete sequences of chains. $Repr^e$ denotes the representation of HLA-II molecules using the fused information of ProtT5 and ONE-HOT from the complete sequences of chains. CA^a indicates that the attention module employs the DSCA mechanism, whereas CA^b indicates that the attention module employs the conventional cross-attention mechanism.

https://doi.org/10.1371/journal.pcbi.1013836.t003

In addition, the longitudinal comparison of the experimental results shows that both ESMC and the DSCA mechanism contribute positively to model performance. By incorporating the rich representational information provided by ESMC, the performance in terms of AUROC and AUPR improves from 0.981 and 0.945 in Model 1 to 0.988 and 0.963, respectively. Through the effective modeling enabled by the DSCA mechanism, the AUROC and AUPR similarly increase from 0.788 and 0.454 in Model 1 to 0.988 and 0.963. These results indicate that DSCA-HLAII, supported by a well-designed modeling strategy and architecture, achieves superior performance on the peptide–HLA-II presentation task by further integrating the rich semantic and structural information encoded by ESMC.

Overall, the experimental results demonstrate that the combination of fused representations and the DSCA module synergistically promotes accurate modeling of peptide–HLA-II interactions and presentation. Specifically, for HLA-II molecules, the fused representation of complete sequences of HLA-II chains introduces a more comprehensive knowledge context compared with pseudo sequences. The DSCA mechanism effectively captures key sites, partially mitigating potential noise introduced by full-length sequences. The integration of these two components achieves complementarity, ultimately yielding optimal modeling and predictive performance.

2.4 Binding core prediction

In the interaction between peptides and MHC-II molecules, the peptide's core region typically consists of a contiguous stretch of nine amino acid residues, referred to as the binding core [42]. This region is accommodated within the binding groove of the MHC-II molecule and interacts with HLA-II via key anchor residues, thereby influencing peptide binding and presentation. As illustrated in Fig 9F, DSCA-HLAII is designed to effectively capture the positions of binding cores based on the fused peptide and HLA-II features. The position with the highest score is regarded as the predicted start position of the binding core by DSCA-HLAII.

The binding cores predicted by DSCA-HLAII provide insights into the potential biological mechanisms underlying the peptide-MHC-II presentation pathway. To validate the effectiveness of binding core prediction, we used the allele DRA*01:01-DRB1*01:01 as an example and visualized peptides corresponding to binding cores at different start positions as sequence logos [43] to examine residue conservation. Specifically, we first employed DSCA-HLAII to predict the presentation probabilities of 100,000 randomly selected peptide segments from the UniProt database. The top 1% of peptides with the highest predicted presentation probabilities were then selected, and sequence logos were generated using Seq2Logo with default settings [44].The results are shown in Fig 5.

As shown in Fig 5, the binding core motifs remain highly conserved regardless of the start position of the binding core. Previous studies have indicated that pocket 1 represents the largest and most critical side-chain binding pocket in HLA-DR1 [45]. Our results demonstrate that the P1 position exhibits the highest relative information content, which further supports the validity of DSCA-HLAII in binding core prediction. Moreover, the P1 position is predominantly enriched with large hydrophobic or aromatic residues such as Phe (F), Tyr (Y), and Ile (I), consistent with previous findings on the pocket binding preferences of HLA-DR molecules [46]. In addition to P1, the P6 and P9 positions were identified by DSCA-HLAII as key anchor residues essential for peptide–HLA-II binding [47]. Therefore, DSCA-HLAII is capable of accurately identifying informative binding cores across different start positions of peptide binding cores.

2.5 Interpretability in DSCA-HLAII: effective feature learning and identifying key binding sites

We explore the interpretable insights of DSCA-HLAII from two perspectives: the explainability analysis of input feature importance and the visualization of attention weights from the DSCA mechanism. This interpretability allows for a more detailed, residue-level characterization of the interaction mechanisms between peptides and HLA-II molecules.

2.5.1 Analysis of the importance of multi-view features. We applied the SHapley Additive exPlanations (SHAP) method [48] to evaluate the importance of different features on both the peptide and HLA-II molecule sides. In the context of peptide–HLA-II interaction prediction, Fig 6A and Fig 6B respectively present the top 20 features contributing most

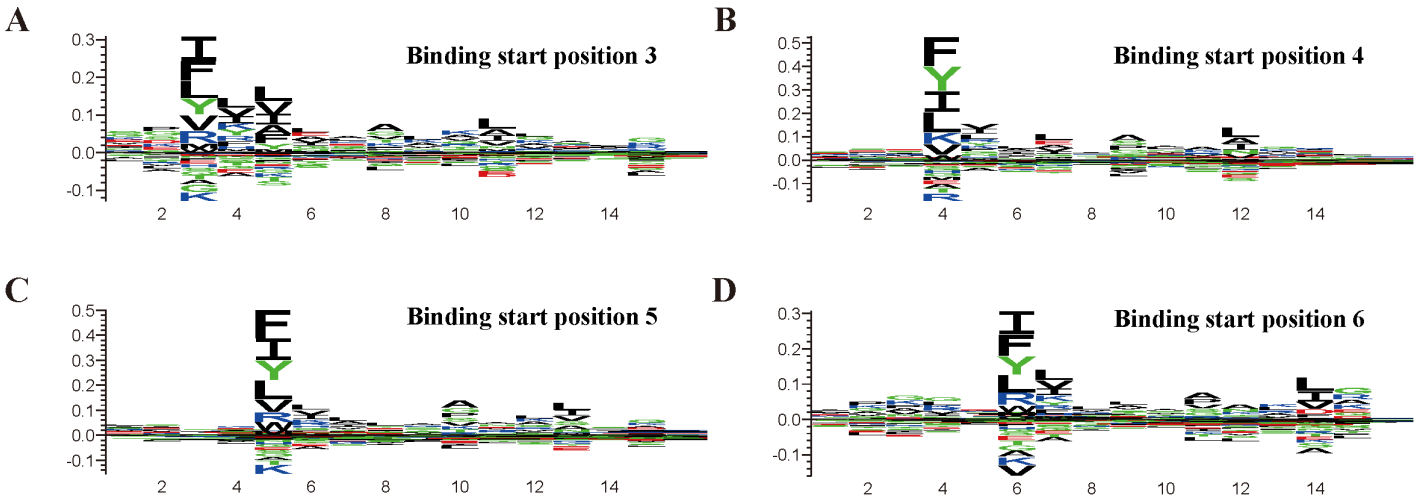

Fig 5. Sequence logos of binding cores at different start positions by DSCA-HLAII. The x-axis denotes residue positions within the peptide. At each position, the total height represents the relative information content, reflecting the degree of conservation, while the height of each letter corresponds to the contribution of the respective amino acid at that position.

https://doi.org/10.1371/journal.pcbi.1013836.g005

substantially to the DSCA-HLAII model from the peptide side and the HLA-II molecule side. The results indicate that the ESMC features dominate on both sides, highlighting their critical role in the model's predictive process. This suggests that incorporating pre-trained embeddings facilitates the capture of key semantic information influencing presentation, thereby enhancing prediction accuracy and generalization capability. Notably, on the peptide side, features corresponding to residues 5–8 constitute a large portion of the top 20 features, whereas on the HLA-II molecule side, features near residues 23 and 170 occupy a significant proportion. These findings partially reveal the positional distribution of key sites in peptides and HLA-II molecules that are relevant for antigen presentation.

2.5.2 Residue-level analysis of dual-stream cross-attention. The DSCA module can compute the attention weight for each site within a peptide sequence, reflecting the relative contribution of individual residues to peptide presentation. We selected peptide-HLA-II pairs from the BC2015 dataset [3] and computed the attention weights for the peptide sequences using DSCA-HLAII, visualizing the distribution of these weights via heatmaps. As shown in Fig 7A, high-attention sites are predominantly concentrated within the binding core region, indicating that the DSCA module effectively highlights key residues relevant to binding and provides meaningful interpretability for subsequent peptide–HLA-II presentation predictions. Furthermore, we compared the binding core positions predicted by DSCA-HLAII with the experimentally validated cores. The results demonstrate that DSCA-HLAII can accurately identify the true binding core regions. Owing to its integration of contextual information from both the peptide sequence and the HLA-II molecular sequence, a small number of predicted cores exhibit a positional shift of 1–2 residues. From the visualized dataset, we further selected representative complexes formed by HLA-DR, HLA-DP, and HLA-DQ molecules and illustrated their PDB structures in Fig 7B. Residues with higher attention weights are generally located within the binding groove of the HLA-II molecules and participate in peptide–HLA-II interactions. For instance, in the HLA-DRA*01:01–DRB1*01:01–peptide complex (PDB ID: 4OV5), the sixth residue of the peptide, ARG, forms a stable hydrogen bond with ASN82 of the HLA β-chain. Similar interaction patterns were observed in other HLA-DP and HLA-DQ complexes, suggesting that the attention weights captured by the model reasonably correspond to structurally critical interacting residues. These findings further demonstrate that the DSCA mechanism can identify the key sites of peptides within HLA-II molecules. In addition, it highlights residues that play crucial roles in molecular interactions, providing structural-level interpretability for the model's predictions.

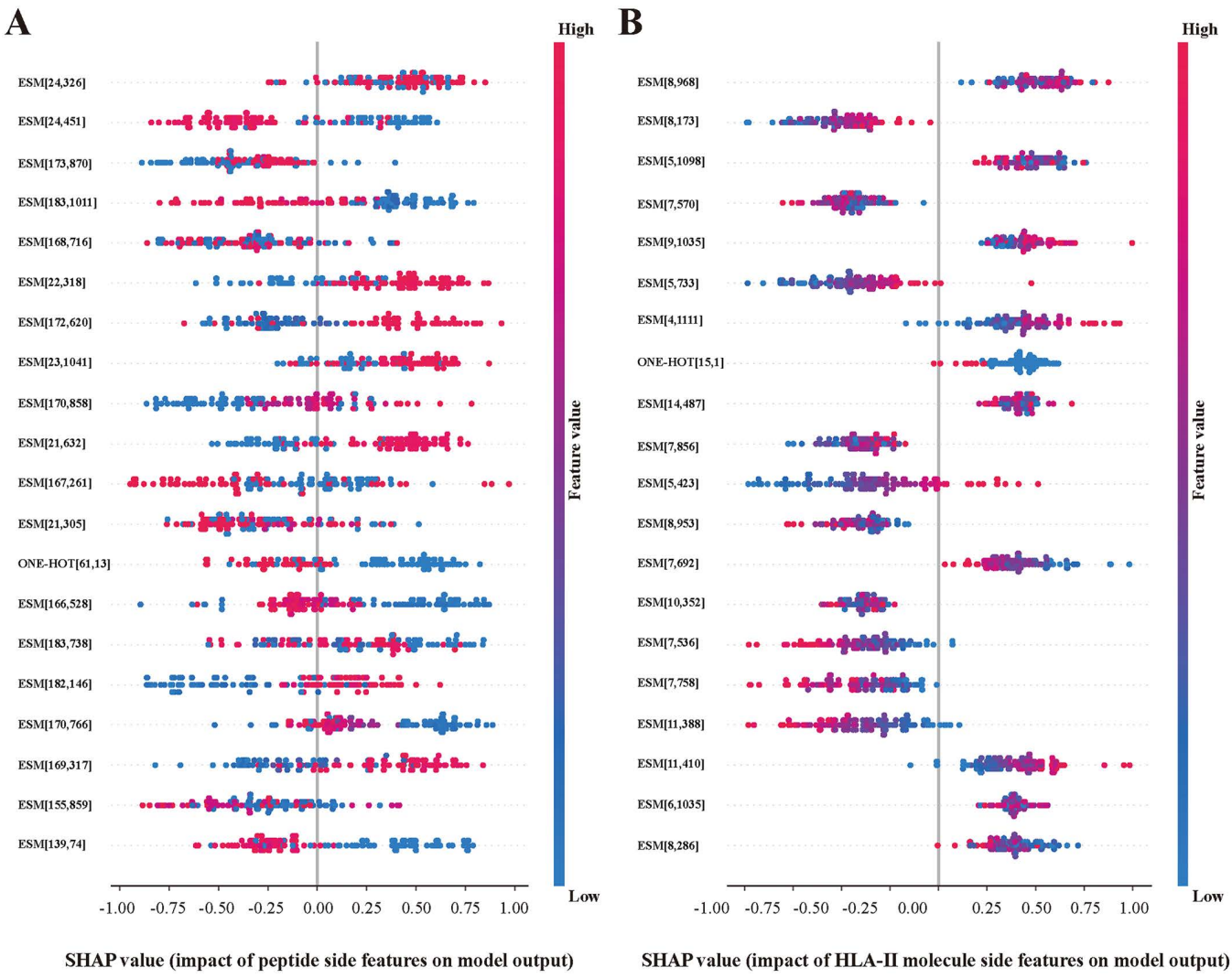

Fig 6. Interpretable Analysis of DSCA-HLAII based on SHAP. (A) The top 20 most influential features of peptide side that impact the model's predictions, where ESM denotes the ESMC features, ONE-HOT denotes the ONE-HOT features, [*a*, *b*] represents position *a* of the amino acid and position *b* of the feature. (B) The top 20 most influential features of HLA-II molecule side that impact the model's predictions.

https://doi.org/10.1371/journal.pcbi.1013836.g006

2.6 Antibody immunogenicity risk assessment

The immunogenicity of an antibody is one of the key indicators for evaluating its potential as a safe and effective therapeutic candidate, as it determines whether the antibody may elicit adverse immune responses in vivo [49]. Previous studies have shown that immunogenicity is closely associated with the recognition of potential T-cell epitopes within the antibody molecule, a process that primarily depends on the binding and presentation of antibody-derived peptide fragments by HLA-II molecules [50]. Therefore, elucidating peptide–HLA-II interactions is of critical importance for predicting antibody immunogenicity risk [51]. In this study, DSCA-HLAII not only accurately predicts the binding affinity between peptides and HLA-II molecules but also facilitates the assessment of antibody immunogenicity risk, thereby providing valuable guidance for the safety evaluation and rational design of antibody therapeutics.

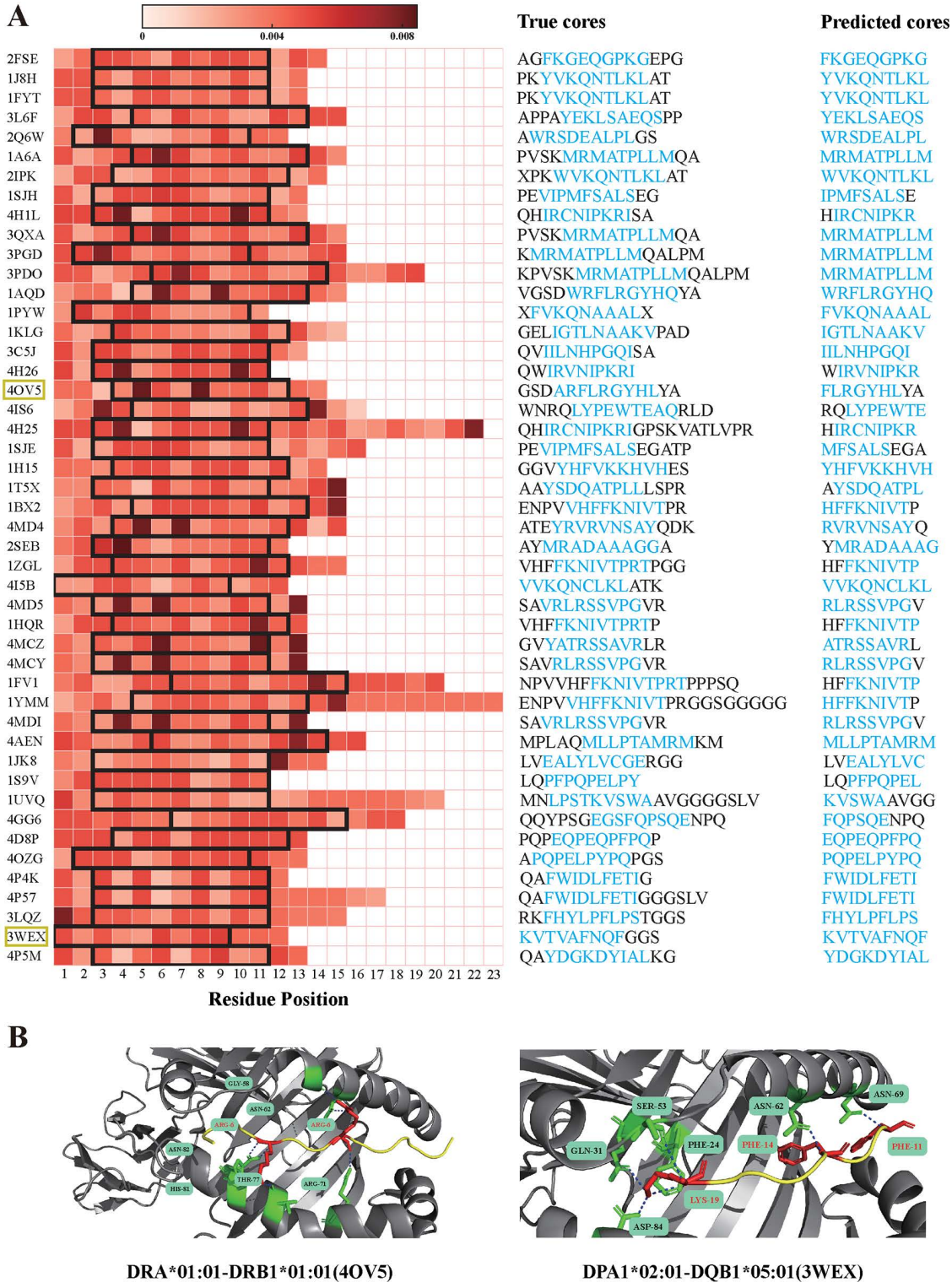

Fig 7. Visualization of DSCA-HLAII average attention weight. (A) Heatmap visualizing the positional attention of peptide sequences within peptide–HLA-II complexes. Each row corresponds to a peptide sequence in a given complex, and each column represents a residue position. The color intensity indicates the magnitude of the attention weight, with black boxes highlighting the binding core. The True cores column on the right annotates each complex with its corresponding peptide sequence and the experimentally validated binding core positions. The Predicted cores column indicates the binding

core positions predicted by DSCA-HLAII. (B) Structural visualization of peptide–HLA-II complexes. Peptide sequences are shown in yellow, residues with high attention weights are highlighted in red, polar interactions are indicated in blue, and HLA-II residues interacting with high-attention peptide residues are marked in green.

https://doi.org/10.1371/journal.pcbi.1013836.g007

Firstly, we extracted all potential peptide fragments from antibody sequences using sliding windows of lengths 12–19 and predicted their presentation probabilities and binding cores for eight specific HLA-II alleles. Subsequently, peptides whose binding cores appeared in more than 22 subjects were filtered out using the OASis [52] to reduce interference from self-derived peptides. Peptides with presentation probabilities below a defined CUTOFF threshold were then further excluded. Finally, the total number of unique binding cores for each allele was counted and used as an indicator of antibody immunogenicity risk. The eight selected HLA-II alleles in this study correspond to common HLA-DR subtypes, including DRB1*01:01, DRB1*03:01, DRB1*04:01, DRB1*07:01, DRB1*08:01, DRB1*11:01, DRB1*13:01, and DRB1*15:01 [16].

As shown in Fig 8, we compared the performance of DSCA-HLAII with Sapiens [52], Graph-pMHC [16], and NetMHCIIpan [15] on the clinical antibody immunogenicity dataset [16]. The results demonstrate that DSCA-HLAII outperforms the other methods in predicting antibody immunogenicity. This indicates that, in addition to its superior performance in peptide–HLA-II presentation prediction, DSCA-HLAII also possesses strong predictive capability for assessing the immunogenicity risk of antibodies.

3. Discussion

This study introduces a novel multi-task predictive framework, DSCA-HLAII, based on a DSCA architecture, which provides a unified approach for modeling peptide–HLA class II interactions, identifying binding cores, and predicting peptide presentation. By integrating ONE-HOT encoded sequence features with pre-trained semantic embeddings (ESMC), the framework constructs a comprehensive hybrid representation of peptide and full-length HLA-II sequences, enabling a thorough encoding of the structural semantics of peptide–HLA-II complexes. Notably, the proposed DSCA module allows for fine-grained, position-aware interaction modeling between peptides and HLA sequences, significantly enhancing the

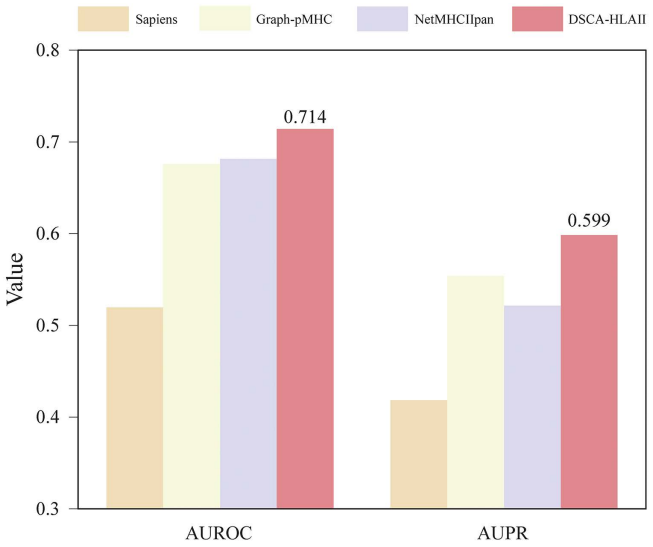

Fig 8. Comparison of the performance of DSCA-HLAII and other methods on the clinical antibody immunogenicity dataset.

https://doi.org/10.1371/journal.pcbi.1013836.g008

interpretability of critical binding site identification and laying a structural foundation for the model's generalization capability in cold-start scenarios.

Experimental validation demonstrates that DSCA-HLAII outperforms existing state-of-the-art methods across multiple warm-start and cold-start datasets, highlighting its dual strengths in accuracy and robustness. The model simultaneously predicts peptide presentation probabilities and binding core locations while supporting systematic assessment of antibody immunogenicity risk, thereby offering a novel computational tool for vaccine design and immunotherapeutic research. To promote practical application, we have further developed a publicly accessible web server (http://bliulab.net/DSCA-HLAII), lowering the barrier to adoption for researchers in the field.

In summary, DSCA-HLAII provides a more accurate, interpretable, and practical solution for predicting peptide–HLA class II interactions by integrating multi-source sequence features with attention-driven interaction modeling. While the model has demonstrated superior performance across multiple test sets, future work may further explore its predictive capacity for rare alleles, incorporate structural information to enrich representations, and deepen its application in peptide vaccine design and personalized immunotherapy. The framework proposed in this study also offers a transferable modeling strategy for other protein–ligand interaction prediction tasks.

4. Conclusion

In this study, we propose DSCA-HLAII, a novel predictive model for peptide–HLA-II interactions and presentation. The model systematically integrates pre-trained semantic and structural ESMC information and ONE-HOT to construct multi-view representations. It then leverages a specially designed DSCA to enable comprehensive learning and effective capture of interaction patterns between peptides and HLA-II molecules. Extensive experimental evaluations demonstrate that DSCA-HLAII consistently achieves state-of-the-art predictive performance and robust generalization capability, while simultaneously offering valuable biological interpretability. Moreover, the model can be extended to binding core prediction and antibody immunogenicity assessment, where it likewise exhibits superior performance. Finally, we have developed an online server, which is publicly accessible at http://bliulab.net/DSCA-HLAII.

Although DSCA-HLAII demonstrates substantial predictive performance, several limitations remain. Our approach only considers single-allelic (SA) samples while neglecting multi-allelic (MA) samples that contain multiple alleles. As a result, DSCA-HLAII may miss learning allele-specific information present exclusively in MA samples. Additionally, because the model incorporates contextual information, while it enhances the modeling of peptides and HLA-II molecules, it may also introduce slight deviations in accurately identifying the precise positions of binding cores. In future work, we aim to integrate MA and SA data more effectively to further enhance the model's generalization capability. Some recent advanced studies have introduced graph-based learning strategies to recognize and model protein complexes, such as FMvPCI [53] and LCAAG [54]. We plan to achieve a comprehensive improvement in both the modeling of peptide–HLA-II interactions and the accurate identification of key sites through more advanced graph learning mechanisms.

5. Methods

5.1 Benchmark and independent test datasets

For the benchmark dataset, the data were systematically curated from multiple authoritative sources, including the IEDB [55], UniProt [56], and IPD-IMGT/HLA [57] databases. For positive peptide–HLA-II pairs, we selected data annotated with the mass spectrometry eluted ligand (MS EL) label from IEDB, as this label reflects the processes of antigen processing and presentation and is generated under more standardized experimental conditions. To obtain complete sequences of HLA-II chains corresponding to the positive pairs, the complete α chain and β chain sequences of each allele were retrieved from the IPD-IMGT database. Based on the records in the UniProt database, the α chain and β chain sequences of different alleles were then segmented at specific positions and concatenated to construct the complete sequences of HLA-II chains for each allele.

For negative peptide–HLA-II pairs, peptide sequences with lengths ranging from 8 to 32 amino acids were randomly extracted from the UniProt database to generate a peptide pool of two million sequences. The complete sequences of HLA-II chains were constructed using the same strategy as for the positive pairs. To enhance the model's generalization capability, we constructed a negative-to-positive sample ratio of 5:1 for each allele. When the available negative samples for a particular allele did not suffice to meet this ratio, additional negative instances were produced by randomly sampling peptides from the peptide pool and pairing them with that allele. To ensure that the peptide length distributions of the positive and negative samples remained consistent, we enforced that the peptide lengths of the newly generated negative samples matched those of the positive samples for each allele.

To rigorously evaluate model performance, all peptide sequences that appeared in the independent test dataset were removed from the benchmark dataset, rather than only excluding peptide–allele pairs with exact matches. This deduplication procedure effectively prevents potential information leakage caused by repeated peptide sequences. Finally, the benchmark dataset comprised 76 HLA-II alleles and a total of 942,602 peptide–HLA-II pairs, including 154,324 positives and 788,278 negatives.

For the independent test dataset, we constructed warm-start and cold-start test datasets based on allele coverage [18]. Alleles in the warm-start test dataset had appeared in the benchmark dataset, whereas alleles in the cold-start test dataset were entirely absent from the benchmark dataset. The warm-start test dataset ultimately included 76 HLA-II alleles and comprised a total of 192,846 peptide–HLA-II pairs, of which 32,141 were positive samples and 160,705 were negative samples. The cold-start test dataset contained 55 HLA-II alleles and a total of 5,262 peptide–HLA-II pairs, including 877 positive and 4,385 negative samples.

5.2 The architecture of the DSCA-HLAII

The framework of DSCA-HLAII is illustrated in Fig 9. It consists of four main stages: Residue-level Embedding, Representation Extraction, Cross-Attention, and Presentation Prediction. This framework is designed to extract both global and local features enriched with semantic information from peptide and HLA-II sequences. By incorporating the DSCA mechanism, the model explicitly captures the dependencies between the two sequences. This design enables the prediction of presentation probabilities and supports downstream tasks such as binding core identification.

In the Residue-level Embedding stage, peptides and HLA-II molecules are represented from two perspectives: sequence-based features and pre-trained embeddings. Specifically, sequence features are encoded using ONE-HOT representations to capture residue-level information. ESMC embeddings encode semantic, structural, and functional information of the sequences. During the Representation Extraction stage, sequence information is extracted from both global and local perspectives. Global features are captured through several layers of transformer encoders and local features are obtained via one-dimensional convolution operations that downsample the peptides and HLA-II sequences. In the Cross-Attention stage, the DSCA mechanism is employed. This mechanism not only captures key sites information within peptide and HLA-II sequences but also explicitly models the dependencies between the peptide and HLA-II sequences. In the Presentation Prediction stage, the fused peptide–HLA-II representations are used to predict presentation probabilities via a binary cross-entropy (BCE) loss function. Moreover, DSCA-HLAII can be extended to downstream tasks, including binding core prediction and antibody immunogenicity assessment, demonstrating the potential utility of our approach in immunological research.

5.3 Residue-level embedding module

The Residue-level Embedding module directly determines whether the model can effectively utilize the relevant inputs of peptides and HLA-II molecules, thereby influencing the predictive capability and accuracy of downstream tasks. In this study, DSCA-HLAII extracts both the sequence-based ONE-HOT encoding and the pretrained semantic ESMC representation of peptides and HLA-II molecules. The ONE-HOT encoding reflects residue-level information and preserves

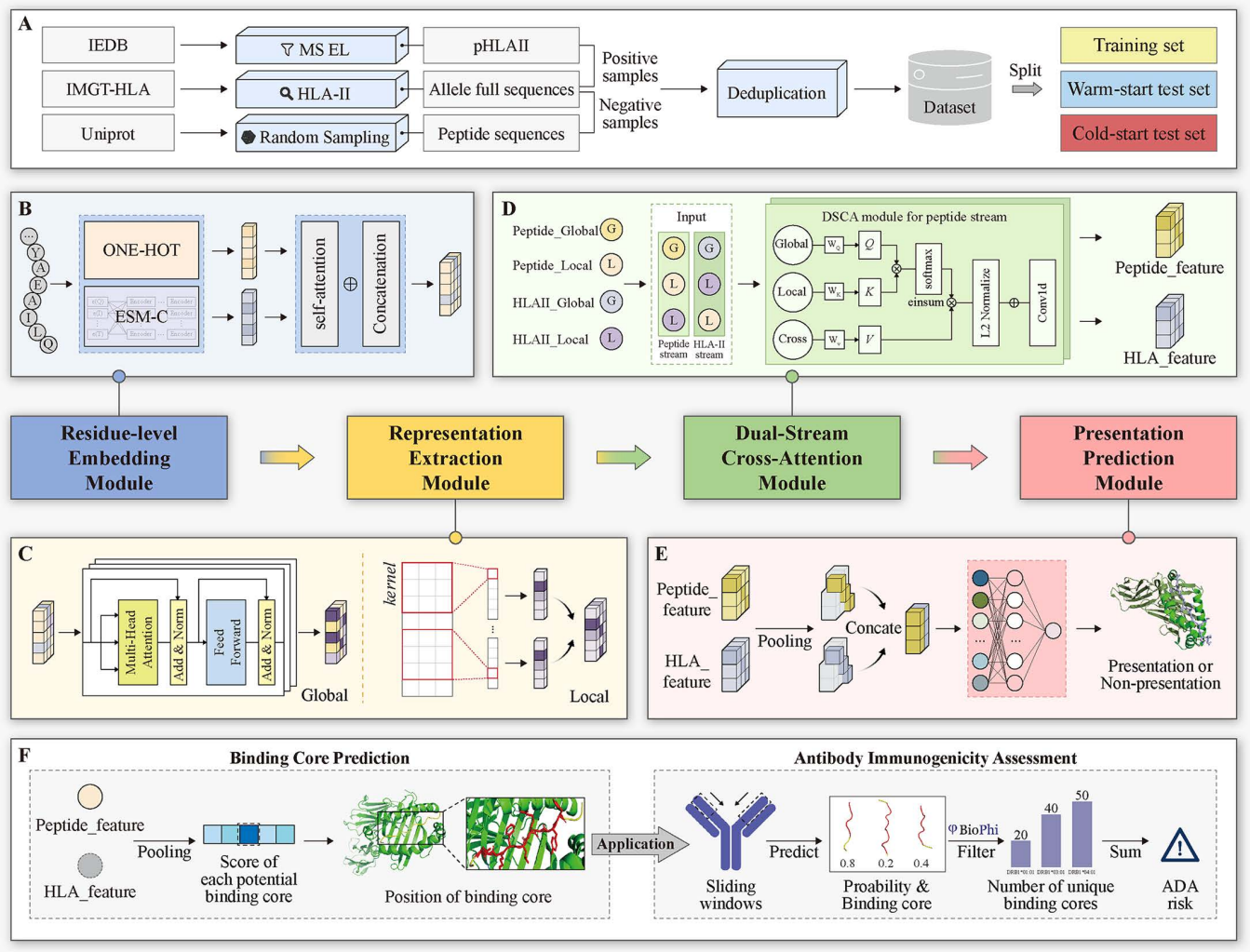

Fig 9. Overview of the DSCA-HLAII framework. A: Data preparation workflow. B: The Residue-level Embedding module. This module extracts ONE-HOT and ESMC representations while incorporating context-enhanced embeddings, providing a comprehensive representation of both peptides and HLA-II molecules. C: The Representation Extraction module. This module captures multi-level dependencies in sequences from global and local perspectives. D: The Cross Attention module. This module employs the DSCA mechanism to capture the information interaction between peptides and HLA-II molecules. The peptide_feature and HLA_feature are constructed through the peptide stream and HLA stream, respectively. E: The Presentation Prediction module. This module outputs the predicted presentation probability based on the integrated interaction features of peptides and HLA-II molecules. F: Downstream Tasks. DSCA-HLAII is used for predicting peptide binding cores and assessing antibody immunogenicity.

https://doi.org/10.1371/journal.pcbi.1013836.g009

fundamental sequence features [58,59]. The ESMC representation captures deeper semantic information, revealing the contextual relationships across amino acids and encoding potential structural and functional characteristics [39]. The sequences of peptides and HLA-II molecules can be represented as $S_{pep} = \{R_1R_2R_3 \ldots R_{L_{pep}}\}$ and $S_{hla} = \{R_1R_2R_3 \ldots R_{L_{hla}}\}$, where R_i denotes one of the standard amino acids or the non-standard amino acid 'X', L_{pep} represents the length of the peptide, and L_{hla} refers to the length of the concatenated α chains and β chains of the HLA-II molecule.

ESMC, a widely used protein language model from the ESM3 family, is purpose-built for advancing protein representation learning [39]. By automatically extracting deep biological features, it avoids the need for the supervisory sequences and feature engineering that traditional methods rely on. ESMC achieves significant gains in predictive performance

alongside enhanced computational efficiency. In this study, we employed the ESMC-600M model to extract features from peptide sequences and the full-length chain sequences of HLA-II molecules. After processing by the ESMC model, peptide sequences S_{pep} and HLA-II sequences S_{hla} are represented as vectors of dimensions $(L_{pep} * 1152)$ and $(L_{hla} * 1152)$, respectively, as defined in the following equation:

$$E_{pep} = ESM(S_{pep}) \tag{1}$$

$$E_{hla} = \left[ESM\left(S_{hla_{alpha}}\right); ESM\left(S_{hla_{beta}}\right)\right] \tag{2}$$

where S_{hla_alpha} and S_{hla_beta} denote the α chain and β chain of the HLA-II molecule, respectively. Due to the heterodimeric nature of HLA-II molecules, the α chains and β chains of S_{hla} are processed independently. Specifically, each chain is separately fed into the ESMC model to generate its representation, and the two representations are then concatenated to construct the comprehensive representation of S_{hla} using Eq.(2).

The ONE-HOT feature encodes the sequence into a high-dimensional, sparse binary vector representation. Since the lengths of peptide sequences are inherently heterogeneous, we standardize the sequences to a fixed length of 32 through padding or truncation. Consequently, the peptide sequence S_{pep} and the HLA-II molecular sequence S_{hla} are represented as $(32 * 21)$-D and $(200 * 21)$-D vectors. In addition, we design a self-attention layer to extract context-enhanced representations from the ONE-HOT, enabling the capture of context-dependent relationships across residues within both the peptide and the complete sequences of HLA-II chains. By aligning and concatenating the two types of representations in a latent space, we obtain the fused features of the peptide and HLA-II molecule, denoted as E_{multi}:

$$E_{multi} = [E_{esm}; E_{one\text{–}hot}; SelfAttn(E_{one\text{–}hot})] \tag{3}$$

where E_{esm} and $E_{one\text{–}hot}$ denote the features of ESMC and ONE-HOT, respectively. $SelfAttn()$ denotes the self-attention layer.

Therefore, the fused feature representation comprises both ONE-HOT and ESMC features. The ONE-HOT representation captures the fundamental sequence features, while the ESMC representation encodes deeper semantic information, reflecting contextual dependencies and latent relationships within the sequence. The fused representation provides the model with comprehensive and informative input features, enriching the representational space of peptides and HLA-II molecules.

5.4 Representation extraction module

The model is designed with a feature extraction framework that integrates both global and local encoding strategies to fully capture the multi-level dependencies in features. The synergy between these two strategies allows the model to balance long-range structural information with local functional site information, thereby enhancing both prediction accuracy and generalization capability.

To capture global contextual dependencies, we designed a Transformer-based encoding module, consisting of a stack of three Transformer encoder layers. Formally, each Transformer encoder layer updates the input $\mathbf{X}^{(l)} \in R^{B\times L\times D}$ by applying multi-head self-attention (MHSA) followed by a position-wise feed-forward network (FFN), with residual connections and layer normalization:

$$\mathbf{Z}^{(l)} = \text{LayerNorm}\left(\mathbf{X}^{(l)} + \text{MHSA}\left(\mathbf{X}^{(l)}\right)\right) \tag{4}$$

$$\mathbf{X}^{(l+1)} = \text{LayerNorm}\left(\mathbf{Z}^{(l)} + \text{FFN}\left(\mathbf{Z}^{(l)}\right)\right) \tag{5}$$

where $\mathbf{X}^{(l)}$ denotes the input to the l-th Transformer encoder layer, and $\mathbf{Z}^{(l)}$ represents the intermediate representation at the same layer. Through the global encoding process, the model is able to capture dependencies within the multi-source fused features. Compared with standard self-attention, the MHSA mechanism allows multiple queries to focus on different relational patterns within the sequence, thereby capturing comprehensive information.

To capture local short-range dependencies, we apply one-dimensional convolution operations to downsample the features. The design of the convolutional kernels incorporates biological prior knowledge of both peptides and HLA-II molecules. For peptide sequences, considering the length of the binding core, the kernel size is set to 9 to effectively capture local information for each potential binding core. For HLA-II sequences, taking into account the positional distribution of key binding pockets, the kernel size is set to 5 to facilitate the capture of contextual relationships among residues within local pockets. Finally, the resulting local short-range dependency features are denoted as H_{local}. This process can be formally represented as:

$$H_{local} = Conv1D\left(E_{multi}\right) \tag{6}$$

Therefore, the global encoding is designed to capture contextual associations across the entire sequence, enabling the characterization of both overall structural features and long-range interactions. In contrast, the local encoding targets fine-grained features by incorporating biological prior knowledge into the design of convolutional kernels, which enables the effective capture of local dependencies around the binding core and key pocket regions.

5.5 Dual-stream cross-attention module

In this section, we propose the DSCA model to precisely capture the interaction between the peptide and HLA-II sequences. In conventional cross-attention mechanisms, the Query is typically derived from the peptide sequence, whereas the Key and Value originate from the HLA-II sequence [60]. This design ensures a unidirectional information flow, where features are queried and fused exclusively from the peptide to the HLA-II molecule. However, peptides and HLA-II molecules influence each other during the actual interaction process. The conventional cross-attention mechanism based on the unidirectional interaction may be insufficient to fully characterize their interaction process. Moreover, the attention computation for each position of the peptide and HLA-II molecule is homogeneous, which may hinder the identification of key sites. To address the above limitations, we propose the DSCA module, which operates by processing the global and local features of the peptide and the HLA-II molecule through Peptide stream and HLA-II stream. Through the Peptide stream and HLA-II stream dual-stream design, the interaction process between the peptide and the HLA-II molecule can be more comprehensively modeled. Using the peptide stream as an illustrative example, the operational pipeline of the DSCA module can be described as follows.

Initially, in the Peptide stream, the global feature representation of the peptide is mapped via a linear transformation into the Query space, while the local feature representation of the peptide is mapped via a linear transformation into the Key space. The local feature representation of the HLA-II serves as the Cross information and is mapped via a linear transformation into the Value space. This process can be formally represented as:

$$Q_{cross} = X_{pep\_global}W_Q, K_{cross} = X_{pep\_local}W_K, V_{cross} = Y_{HLA\_local}W_V \tag{7}$$

where X_{pep_global} denotes the global feature of the peptide, X_{local} denotes the local feature of the peptide, and Y_{HLA_local} denotes the local feature from the HLA-II.

Subsequently, the attention scores, computed from the Query and Key, are then leveraged to aggregate the Value via the Einstein summation convention [61], producing preliminary cross-fused information. The resulting representation is subsequently normalized to ensure a stable feature distribution. Notably, the dimensions of the cross-fused information

are aligned with the sequence positions, such that the values of the fused representation partially reflect the distribution of key sites.

$$Attn = \text{Softmax}\left(\frac{Q_{cross}K_{cross}{}^{T}}{\sqrt{d_{pep\_global}}}\right) \tag{8}$$

$$O = Attn \cdot V_{cross} \tag{9}$$

$$\widetilde{O} = \|O\|_2 \tag{10}$$

where d_{pep_global} denotes the dimensionality of the global feature in the peptide stream.

Finally, to capture potential key regions, a one-dimensional convolution, using the same parameters as in the encoding stage, is applied to the fused representation. The result is then combined with the input global features via a residual connection to obtain the final peptide_feature representation H_{pep}. On the peptide side, the dimensions of the final feature representation are aligned with the potential start positions of the binding core, such that the values of the final representation further reflect the distribution of key sites.

$$H_{pep} = Conv1D\left(\widetilde{O}\right) + X_{pep\_global} \tag{11}$$

The Query, derived from the global feature, represents the overall semantics of the entire sequence. The Key, derived from the local feature, captures fine-grained information for each segment. The global feature queries the local feature, directing the attention to key local positions that are most relevant for interaction and reducing the impact of irrelevant positions. This mechanism is utilized to capture contextual dependencies. The Value, derived from the local feature of the other sequence, facilitates cross-stream information transfer. Specifically, the Query–Key attention first identifies key sites within the current sequence, and then the Value extracts the corresponding key sites from the other sequence. This process is consistent with the biological principles of peptide–HLA-II interactions and improves the model's generalization performance.

Similarly, in the HLA-II stream, the DSCA module follows the same workflow as described for the peptide stream, which generates the final HLA_feature H_{HLA}. In summary, the DSCA module, through its dual-stream and cross-stream attention design, enables precise modeling of the bidirectional interactions between peptides and HLA-II molecules. The Query–Key–Value construction based on global and local features not only enhances the identification of key sites but also improves the generalization performance of the model.

5.6 Presentation prediction

The Presentation Prediction module predicts the presentation probability by integrating the interaction features of peptides and HLA-II molecules. Firstly, the peptide feature H_{pep} and HLA-II feature H_{HLA} are separately subjected to global max pooling and concatenation to obtain a joint interaction representation of the peptide and HLA-II, denoted as $G_{interaction}$. Next, a MLP module followed by a linear projection and a sigmoid activation function is used to generate the presentation probability $\hat{y}$, represented as:

$$\hat{y} = \sigma\left(W \cdot \text{MLP}\left(G_{interaction}\right) + b\right) \tag{12}$$

where W and b represent the weight and bias of the final linear projection layer, and σ denotes the sigmoid activation function.

Given that peptide–HLA-II presentation prediction constitutes an imbalanced binary classification problem, the model is trained using the BCELoss function [62], which effectively quantifies the discrepancy between predicted probabilities and true labels, thereby improving the model's discriminative performance between positive and negative classes. It is formally defined as follows:

$$\updownarrow_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log\left(\hat{y}_i\right) + \left(1-y_i\right)\log\left(1-\hat{y}_i\right)\right] \tag{13}$$

where N denotes the number of samples, $\{y_i\}_{i=1}^{N} \in \{0,1\}$ represents the ground-truth label of the i-th sample, and $\{\hat{y}_i\}_{i=1}^{N} \in [0,1]$ denotes the predicted probability of the i-th sample being positive.

5.7 Implementation and training

We implemented DSCA-HLAII using the PyTorch framework (https://pytorch.org/). The model was trained with the Adadelta optimization algorithm [63] for parameter optimization. The learning rate for model training was set to 0.01. To enhance model generalization, we incorporated several key techniques: (1) Dropout regularization to prevent overfitting [64], (2) Early stopping based on validation performance to optimize training duration [65]. The patience parameter for early stopping was set to 7, (3) A weight decay strategy was employed to regularize the model parameters [66], with the weight decay coefficient set to 1×10^{-4}.

To ensure reproducibility and effective performance, these hyperparameters were selected and fine-tuned using 5-fold cross-validation on the benchmark set. Specifically, we explored a range of values across the folds and chose the combination that yielded the best average validation performance.

5.8 Performance evaluation

The presentation probability prediction task is formulated as a binary classification problem. Therefore, we evaluate the model's performance using seven metrics: Precision, Recall, MCC, ACC, F1, AUROC, AUPR, and PCC. AUROC measures [67–71] the overall discriminative ability of a model in distinguishing positive and negative samples [72,73]. AUPR assesses model performance by integrating precision and recall trade-offs [74].

Assume the test set contains N samples, each consisting of a peptide–HLA-II pair. The ground-truth label is denoted as $\{y_i\}_{i=1}^{N} \in \{0,1\}$, and the model's predicted score is represented as $\{\{\hat{s}_i\}_{i=1}^{N} \in [0,1]$. The predicted label $\hat{y}_i \in \{0,1\}$ is then obtained by applying a threshold τ. The definitions of each metric are as follows [3,75]:

$$\begin{cases} \text{Precision} = \frac{TP}{TP+FP} \\ \textit{Recall} = \frac{TP}{TP+FN} \\ \text{MCC} = \frac{TP\cdot TN - FP\cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \\ \text{F1} = \frac{2\cdot\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}} \\ \text{ACC} = \frac{TP+TN}{TP+TN+FP+FN} \\ \text{PCC} = \frac{\sum_{i=1}^{N}\left(y_i-\bar{y}\right)\left(\hat{s}_i-\bar{\hat{s}}\right)}{\sqrt{\sum_{i=1}^{N}\left(y_i-\bar{y}\right)^2}\sqrt{\sum_{i=1}^{N}\left(\hat{s}_i-\bar{\hat{s}}\right)^2}} \end{cases} \tag{14}$$

where TP is the true positive, TN is the true negative, FN is the false negative, FP is the false positive, Precision is the proportion of predicted positives that are actually positive, Recall is the proportion of actual positives that are correctly predicted.

Author contributions

Conceptualization: Ke Yan, Hongjun Yu, Shutao Chen, Youyu Wang.

Funding acquisition: Ke Yan, Alexey K. Shaytan, Bin Liu.

Methodology: Ke Yan, Hongjun Yu, Shutao Chen, Alexey K. Shaytan.

Project administration: Bin Liu, Youyu Wang.

Software: Hongjun Yu, Shutao Chen.

Writing – original draft: Ke Yan, Hongjun Yu, Shutao Chen, Bin Liu.

Writing – review & editing: Ke Yan, Hongjun Yu, Alexey K. Shaytan, Bin Liu, Youyu Wang.

References

1. Kumánovics A, Takada T, Lindahl KF. Genomic organization of the mammalian MHC. Annu Rev Immunol. 2003;21:629–57. https://doi.org/10.1146/annurev.immunol.21.090501.080116 PMID: 12500978
2. Stražar M, Park J, Abelin JG, Taylor HB, Pedersen TK, Plichta DR, et al. HLA-II immunopeptidome profiling and deep learning reveal features of antigenicity to inform antigen discovery. Immunity. 2023;56(7):1681-1698.e13. https://doi.org/10.1016/j.immuni.2023.05.009 PMID: 37301199
3. You R, Qu W, Mamitsuka H, Zhu S. DeepMHCII: a novel binding core-aware deep interaction model for accurate MHC-II peptide binding affinity prediction. Bioinformatics. 2022;38(Suppl 1):i220–8. https://doi.org/10.1093/bioinformatics/btac225 PMID: 35758790
4. Venkatesh G, Grover A, Srinivasaraghavan G, Rao S. MHCAttnNet: predicting MHC-peptide bindings for MHC alleles classes I and II using an attention-based deep neural model. Bioinformatics. 2020;36(Suppl_1):i399–406. https://doi.org/10.1093/bioinformatics/btaa479 PMID: 32657386
5. Neefjes J, Jongsma MLM, Paul P, Bakke O. Towards a systems understanding of MHC class I and MHC class II antigen presentation. Nat Rev Immunol. 2011;11(12):823–36. https://doi.org/10.1038/nri3084 PMID: 22076556
6. Abelin JG, Keskin DB, Sarkizova S, Hartigan CR, Zhang W, Sidney J, et al. Mass Spectrometry Profiling of HLA-Associated Peptidomes in Mono-allelic Cells Enables More Accurate Epitope Prediction. Immunity. 2017;46(2):315–26. https://doi.org/10.1016/j.immuni.2017.02.007 PMID: 28228285
7. Purcell AW, Ramarathinam SH, Ternette N. Mass spectrometry-based identification of MHC-bound peptides for immunopeptidomics. Nat Protoc. 2019;14(6):1687–707. https://doi.org/10.1038/s41596-019-0133-y PMID: 31092913
8. Wei L, Xing P, Zeng J, Chen J, Su R, Guo F. Improved prediction of protein-protein interactions using novel negative samples, features, and an ensemble classifier. Artif Intell Med. 2017;83:67–74. https://doi.org/10.1016/j.artmed.2017.03.001 PMID: 28320624
9. Lin HH, Zhang GL, Tongchusak S, Reinherz EL, Brusic V. Evaluation of MHC-II peptide binding prediction servers: applications for vaccine research. BMC Bioinformatics. 2008;9 Suppl 12(Suppl 12):S22. https://doi.org/10.1186/1471-2105-9-S12-S22 PMID: 19091022
10. Yang Y, Wei Z, Cia G, Song X, Pucci F, Rooman M, et al. MHCII-peptide presentation: an assessment of the state-of-the-art prediction methods. Front Immunol. 2024;15:1293706. https://doi.org/10.3389/fimmu.2024.1293706 PMID: 38646540
11. Yan K, Lv H, Shao J, Chen S, Liu B. TPpred-SC: multi-functional therapeutic peptide prediction based on multi-label supervised contrastive learning. Sci China Inf Sci. 2024;67(11). https://doi.org/10.1007/s11432-024-4147-8
12. Mohapatra M, Sahu C, Mohapatra S. Trends of Artificial Intelligence (AI) use in drug targets, discovery and development: current status and future perspectives. Curr Drug Targets. 2025;26(4):221–42. https://doi.org/10.2174/0113894501322734241008163304 PMID: 39473198
13. Jensen KK, Andreatta M, Marcatili P, Buus S, Greenbaum JA, Yan Z, et al. Improved methods for predicting peptide binding affinity to MHC class II molecules. Immunology. 2018;154(3):394–406. https://doi.org/10.1111/imm.12889 PMID: 29315598
14. Racle J, Michaux J, Rockinger GA, Arnaud M, Bobisse S, Chong C, et al. Robust prediction of HLA class II epitopes by deep motif deconvolution of immunopeptidomes. Nat Biotechnol. 2019;37(11):1283–6. https://doi.org/10.1038/s41587-019-0289-6 PMID: 31611696
15. Nilsson JB, Kaabinejadian S, Yari H, Kester MGD, van Balen P, Hildebrand WH, et al. Accurate prediction of HLA class II antigen presentation across all loci using tailored data acquisition and refined machine learning. Sci Adv. 2023;9(47):eadj6367. https://doi.org/10.1126/sciadv.adj6367 PMID: 38000035
16. Thrift WJ, Perera J, Cohen S, Lounsbury NW, Gurung HR, Rose CM, et al. Graph-pMHC: graph neural network approach to MHC class II peptide presentation and antibody immunogenicity. Brief Bioinform. 2024;25(3):bbae123. https://doi.org/10.1093/bib/bbae123 PMID: 38555476
17. Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, Green T. Protein complex prediction with AlphaFold-Multimer. bioRxiv. 2021. https://doi.org/10.1101/2021.10.04.463034
18. Wang M, Lei C, Wang J, Li Y, Li M. TripHLApan: predicting HLA molecules binding peptides based on triple coding matrix and transfer learning. Brief Bioinform. 2024;25(3):bbae154. https://doi.org/10.1093/bib/bbae154 PMID: 38600667
19. Chang Y, Wu L. CapHLA: a comprehensive tool to predict peptide presentation and binding to HLA class I and class II. Brief Bioinform. 2024;26(1):bbae595. https://doi.org/10.1093/bib/bbae595 PMID: 39688477
20. 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. 2020;48(W1):W449–54. https://doi.org/10.1093/nar/gkaa379 PMID: 32406916

21. Slathia PS, Sharma P. In Silico Designing of Vaccines: Methods, Tools, and Their Limitations. Computer-Aided Drug Design. Springer Singapore. 2020. p. 245–77. https://doi.org/10.1007/978-981-15-6815-2_11
22. Chen Y, Wang Z, Wang J, Chu Y, Zhang Q, Li ZA, et al. Self-supervised learning in drug discovery. Sci China Inf Sci. 2025;68(7). https://doi.org/10.1007/s11432-024-4453-4
23. Karosiene E, Rasmussen M, Blicher T, Lund O, Buus S, Nielsen M. NetMHCIIpan-3.0, a common pan-specific MHC class II prediction method including all three human MHC class II isotypes, HLA-DR, HLA-DP and HLA-DQ. Immunogenetics. 2013;65(10):711–24.
24. Wang Y, Zhai Y, Ding Y, Zou Q. SBSM-Pro: support bio-sequence machine for proteins. Sci China Inf Sci. 2024;67(11). https://doi.org/10.1007/s11432-024-4171-9
25. Wang X, Wu T, Jiang Y, Chen T, Pan D, Jin Z, et al. RPEMHC: improved prediction of MHC-peptide binding affinity by a deep learning approach based on residue-residue pair encoding. Bioinformatics. 2024;40(1):btad785. https://doi.org/10.1093/bioinformatics/btad785 PMID: 38175759
26. Jiang Y, Wang R, Feng J, Jin J, Liang S, Li Z, et al. Explainable Deep Hypergraph Learning Modeling the Peptide Secondary Structure Prediction. Adv Sci (Weinh). 2023;10(11):e2206151. https://doi.org/10.1002/advs.202206151 PMID: 36794291
27. Meng Q, Guo F, Tang J. Improved structure-related prediction for insufficient homologous proteins using MSA enhancement and pre-trained language model. Brief Bioinform. 2023;24(4):bbad217. https://doi.org/10.1093/bib/bbad217 PMID: 37321965
28. Lai L, Liu Y, Song B, Li K, Zeng X. Deep Generative Models for Therapeutic Peptide Discovery: A Comprehensive Review. ACM Comput Surv. 2025;57(6):1–29. https://doi.org/10.1145/3714455
29. Zulfiqar H, Guo Z, Ahmad RM, Ahmed Z, Cai P, Chen X, et al. Deep-STP: a deep learning-based approach to predict snake toxin proteins by using word embeddings. Front Med (Lausanne). 2024;10:1291352. https://doi.org/10.3389/fmed.2023.1291352 PMID: 38298505
30. Zou X, Ren L, Cai P, Zhang Y, Ding H, Deng K, et al. Accurately identifying hemagglutinin using sequence information and machine learning methods. Front Med (Lausanne). 2023;10:1281880. https://doi.org/10.3389/fmed.2023.1281880 PMID: 38020152
31. Jha K, Saha S, Singh H. Prediction of protein-protein interaction using graph neural networks. Sci Rep. 2022;12(1):8360. https://doi.org/10.1038/s41598-022-12201-9 PMID: 35589837
32. Zhang W, Wei H, Zhang W, Wu H, Liu B. Multiple types of disease-associated RNAs identification for disease prognosis and therapy using heterogeneous graph learning. Sci China Inf Sci. 2024;67(8). https://doi.org/10.1007/s11432-024-4100-7
33. Yan K, Chen S, Liu B, Wu H. Accurate prediction of toxicity peptide and its function using multi-view tensor learning and latent semantic learning framework. Bioinformatics. 2025;41(9):btaf489. https://doi.org/10.1093/bioinformatics/btaf489 PMID: 40905623
34. Wang R, Jiang Y, Jin J, Yin C, Yu H, Wang F, et al. DeepBIO: an automated and interpretable deep-learning platform for high-throughput biological sequence prediction, functional annotation and visualization analysis. Nucleic Acids Res. 2023;51(7):3017–29. https://doi.org/10.1093/nar/gkad055 PMID: 36796796
35. Wang P, Sidney J, Dow C, Mothé B, Sette A, Peters B. A systematic assessment of MHC class II peptide binding predictions and evaluation of a consensus approach. PLoS Comput Biol. 2008;4(4):e1000048. https://doi.org/10.1371/journal.pcbi.1000048 PMID: 18389056
36. Chen S, Yan K, Liu B. PDB-BRE: A ligand-protein interaction binding residue extractor based on Protein Data Bank. Proteins. 2024;92(1):145–53. https://doi.org/10.1002/prot.26596 PMID: 37750380
37. Zou Q, Xing P, Wei L, Liu B. Gene2vec: gene subsequence embedding for prediction of mammalian N6-methyladenosine sites from mRNA. RNA. 2019;25(2):205–18. https://doi.org/10.1261/rna.069112.118 PMID: 30425123
38. Lv Z, Ding H, Wang L, Zou Q. A Convolutional Neural Network Using Dinucleotide One-hot Encoder for identifying DNA N6-Methyladenine Sites in the Rice Genome. Neurocomputing. 2021;422:214–21. https://doi.org/10.1016/j.neucom.2020.09.056
39. Team E. ESM Cambrian: Revealing the mysteries of proteins with unsupervised learning. Evolutionary Scale. 2024.
40. Brandes N, Ofer D, Peleg Y, Rappoport N, Linial M. ProteinBERT: a universal deep-learning model of protein sequence and function. Bioinformatics. 2022;38(8):2102–10. https://doi.org/10.1093/bioinformatics/btac020 PMID: 35020807
41. Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, et al. ProtTrans: toward understanding the language of life through self-supervised learning. IEEE Trans Pattern Anal Mach Intell. 2022;44(10):7112–27. https://doi.org/10.1109/TPAMI.2021.3095381 PMID: 34232869
42. Janeway CA, Travers P, Walport M, Shlomchik MJ. Immunobiology: the immune system in health and disease. New York, USA: Garland Publishing. 2001.
43. Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. 1990;18(20):6097–100. https://doi.org/10.1093/nar/18.20.6097 PMID: 2172928
44. Nielsen M, Lundegaard C, Lund O. Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. BMC Bioinformatics. 2007;8:238. https://doi.org/10.1186/1471-2105-8-238 PMID: 17608956
45. Murthy VL, Stern LJ. The class II MHC protein HLA-DR1 in complex with an endogenous peptide: implications for the structural basis of the specificity of peptide binding. Structure. 1997;5(10):1385–96. https://doi.org/10.1016/s0969-2126(97)00288-8 PMID: 9351812
46. Rappazzo CG, Huisman BD, Birnbaum ME. Repertoire-scale determination of class II MHC peptide binding via yeast display improves antigen prediction. Nat Commun. 2020;11(1):4414. https://doi.org/10.1038/s41467-020-18204-2 PMID: 32887877
47. Rammensee H, Bachmann J, Emmerich NP, Bachor OA, Stevanović S. SYFPEITHI: database for MHC ligands and peptide motifs. Immunogenetics. 1999;50(3–4):213–9. https://doi.org/10.1007/s002510050595 PMID: 10602881

48. Ribeiro MT, Singh S, Guestrin C, editors. " Why should i trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016: 1135-1144.
49. Schellekens H. Immunogenicity of therapeutic proteins: clinical implications and future prospects. Clin Ther. 2002;24(11):1720–40; discussion 1719. https://doi.org/10.1016/s0149-2918(02)80075-3 PMID: 12501870
50. Jawa V, Terry F, Gokemeijer J, Mitra-Kaushik S, Roberts BJ, Tourdot S, et al. T-Cell Dependent Immunogenicity of Protein Therapeutics Pre-clinical Assessment and Mitigation-Updated Consensus and Review 2020. Front Immunol. 2020;11:1301. https://doi.org/10.3389/fimmu.2020.01301 PMID: 32695107
51. Qi R, Liu S, Hui X, Shaytan AK, Liu B. AI in drug development: advances in response, combination therapy, repositioning, and molecular design. Sci China Inf Sci. 2025;68(7). https://doi.org/10.1007/s11432-024-4461-0
52. Prihoda D, Maamary J, Waight A, Juan V, Fayadat-Dilman L, Svozil D, et al. BioPhi: A platform for antibody design, humanization, and humanness evaluation based on natural antibody repertoires and deep learning. MAbs. 2022;14(1):2020203. https://doi.org/10.1080/19420862.2021.2020203 PMID: 35133949
53. Yang Y, Hu L, Li G, Li D, Hu P, Luo X. FMvPCI: a multiview fusion neural network for identifying protein complex via fuzzy clustering. IEEE Trans Syst Man Cybern, Syst. 2025;55(9):6189–202. https://doi.org/10.1109/tsmc.2025.3578348
54. Yang Y, Hu L, Li G, Li D, Hu P, Luo X. Link-based attributed graph clustering via approximate generative Bayesian learning. IEEE Trans Syst Man Cybern, Syst. 2025;55(8):5730–43. https://doi.org/10.1109/tsmc.2025.3572738
55. Vita R, Blazeska N, Marrama D, IEDB Curation Team Members, Duesing S, Bennett J, et al. The Immune Epitope Database (IEDB): 2024 update. Nucleic Acids Res. 2025;53(D1):D436–43. https://doi.org/10.1093/nar/gkae1092 PMID: 39558162
56. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019;47(D1):D506–15. https://doi.org/10.1093/nar/gky1049 PMID: 30395287
57. Barker DJ, Maccari G, Georgiou X, Cooper MA, Flicek P, Robinson J, et al. The IPD-IMGT/HLA Database. Nucleic Acids Res. 2023;51(D1):D1053–60. https://doi.org/10.1093/nar/gkac1011 PMID: 36350643
58. ElAbd H, Bromberg Y, Hoarfrost A, Lenz T, Franke A, Wendorff M. Amino acid encoding for deep learning applications. BMC Bioinformatics. 2020;21(1):235. https://doi.org/10.1186/s12859-020-03546-x PMID: 32517697
59. Li H, Pang Y, Liu B. BioSeq-BLM: a platform for analyzing DNA, RNA, and protein sequences based on biological language models. Nucleic Acids Research. 2021;49(22):e129.
60. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN. Attention is all you need. Advances in neural information processing systems. 2017;30.
61. Misner CW, Thorne KS, Wheeler JA, Gravitation W. Freeman and company. San Francisco. 1973;891.
62. Goodfellow I, Bengio Y, Courville A, Bengio Y. Deep learning. Cambridge: MIT press; 2016.
63. Zeiler MD. Adadelta: an adaptive learning rate method. arXiv preprint arXiv:12125701. 2012.
64. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research. 2014;15(1):1929–58.
65. Prechelt L. Early stopping-but when? In: Neural Networks: Tricks of the trade. Berlin, Heidelberg: Springer; 2002. pp. 55–69.
66. Krogh A, Hertz J. A simple weight decay can improve generalization. Advances in Neural Information Processing Systems. 1991;4.
67. Huang Z, Xiao Z, Ao C, Guan L, Yu L. Computational approaches for predicting drug-disease associations: a comprehensive review. Front Comput Sci. 2024;19(5). https://doi.org/10.1007/s11704-024-40072-y
68. Huang Z, Guo X, Qin J, Gao L, Ju F, Zhao C, et al. Accurate RNA velocity estimation based on multibatch network reveals complex lineage in batch scRNA-seq data. BMC Biol. 2024;22(1):290. https://doi.org/10.1186/s12915-024-02085-8 PMID: 39696422
69. Guo X, Huang Z, Ju F, Zhao C, Yu L. Highly Accurate Estimation of Cell Type Abundance in Bulk Tissues Based on Single-Cell Reference and Domain Adaptive Matching. Adv Sci (Weinh). 2024;11(7):e2306329. https://doi.org/10.1002/advs.202306329 PMID: 38072669
70. Shao J, Chen J, Liu B. ProFun-SOM: protein function prediction for specific ontology based on multiple sequence alignment reconstruction. IEEE Trans Neural Netw Learn Syst. 2025;36(5):8060–71. https://doi.org/10.1109/TNNLS.2024.3419250 PMID: 38980781
71. Yan K, Lv H, Guo Y, Peng W, Liu B. sAMPpred-GAT: prediction of antimicrobial peptide by graph attention network and predicted peptide structure. Bioinformatics. 2023;39(1):btac715. https://doi.org/10.1093/bioinformatics/btac715 PMID: 36342186
72. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition. 1997;30(7):1145–59. https://doi.org/10.1016/s0031-3203(96)00142-2
73. Xie X, Wu C, Dao F, Deng K, Yan D, Huang J, et al. scRiskCell: A single-cell framework for quantifying islet risk cells and their adaptive dynamics in type 2 diabetes. Imeta. 2025;4(4):e70060. https://doi.org/10.1002/imt2.70060 PMID: 40860447
74. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015;10(3):e0118432. https://doi.org/10.1371/journal.pone.0118432 PMID: 25738806
75. Chen S, Yan K, Li X, Liu B. Protein Language Pragmatic Analysis and Progressive Transfer Learning for Profiling Peptide-Protein Interactions. IEEE Trans Neural Netw Learn Syst. 2025;36(8):15385–99. https://doi.org/10.1109/TNNLS.2025.3540291 PMID: 40100664