Yu et al. BMC Genomics (2023) 25:127
https://doi.org/10.1186/s12864-023-09900-6

BMC Genomics

RESEARCH

Open Access

Structure-aware deep model for MHC-II peptide binding affinity prediction

Ying Yu[1†], Lipeng Zu[2†], Jiaye Jiang[1], Yafang Wu[1], Yinglin Wang[1], Midie Xu[3,4,5*] and Qing Liu[1*]

Abstract

The prediction of major histocompatibility complex (MHC)-peptide binding affinity is an important branch in immune bioinformatics, especially helpful in accelerating the design of disease vaccines and immunity therapy. Although deep learning-based solutions have yielded promising results on MHC-II molecules in recent years, these methods ignored structure knowledge from each peptide when employing the deep neural network models. Each peptide sequence has its specific combination order, so it is worth considering adding the structural information of the peptide sequence to the deep model training. In this work, we use positional encoding to represent the structural information of peptide sequences and validly combine the positional encoding with existing models by different strategies. Experiments on three datasets show that the introduction of position-coding information can further improve the performance built upon the existing model. The idea of introducing positional encoding to this field can provide important reference significance for the optimization of the deep network structure in the future.

Keywords MHC-II molecules, Affinity prediction, Positional embedding

Introduction

T-cells present on their surface a specific receptor known as the T-cell receptor (TCR) that enables the recognition of antigens when they are displayed on the surface of antigen-presenting cells (APCs) bound to major histocompatibility complex (MHC) molecules [1, 2], which play a significant role in the adaptive immune response mediated by T cells [3, 4]. Due to the time-consuming and labor-intensive process of biochemical experiments [5, 6], the machine learning-based method of predicting MHC binding peptides has attracted more and more attention and has been used to optimize the selection of a small number of promising high-affinity binding peptides, which are further verified by biochemical experiments [7–9]. In general, there are two major classes of MHC molecules: MHC Class I (MHC-I) and MHC Class II (MHC-II) with subclasses in each of these two classes. MHC-I mainly has A, B, and C subclasses, while MHC-II mainly has DP, DQ, and DR subclasses encoded in the human leukocyte antigen (HLA) gene [10] and in the histocompatibility2 (H-2) gene of mouse [11]. Furthermore, note that different from MHC-I molecules consisting of one chain [12], each MHC-II molecule has two chains, α and β [13, 14].

Most MHC-I peptide ligands have 9 residues, made from a single chain α, and can promise better predicted results for these peptides that hold this size

[†]Ying Yu and Lipeng Zu contributed equally to this work.

*Correspondence:
Midie Xu
xumd27202003@sina.com
Qing Liu
liuq@usst.edu.cn
[1] School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China
[2] Department of Computer Science, Florida State University, Tallahassee 32306, USA
[3] Department of Pathology, Fudan University, Shanghai Cancer Center, Shanghai 200032, China
[4] Department of Medical Oncology, Shanghai Medical College, Fudan University, Shanghai 200032, China
[5] Institute of Pathology, Fudan University, Shanghai 200032, China

[15]. However, the MHC-II molecules' peptide binding groove is open at both ends, allowing the peptide to extend beyond the binding groove (9-22 residues in length), although there is only a core of nine residues in the MHC-II binding groove [16, 17]. Therefore, peptide binding predictions for MHC-II molecules are of great challenge compared to those for MHC-I molecules. Therefore, how to design an effective algorithm framework to predict affinity peptides for MHC-II molecules has become a hot but difficult topic [18, 19]. To date, a variety of methods have been developed to predict the binding capacity of peptides to MHC-II [20–22], among which prediction of MHC-II peptide binding based on panspecificity is the most common and efficient computational solution [23, 24].

Early panspecific-based methods can be divided into various techniques, such as support vector machine (SVM) [25], motif matrix (MM) [26], artificial neural network (ANN) [27], and kernel-based methods [28]. Furthermore, NetMHCIIpan-4.0, an ANN-based method, exploited customized machine learning strategies to integrate different types of training data, resulting in happy performance and outperforming their competitors in that year [29]. Recently, pan-specific based methods began to be oriented toward deep learning (DL) [30]: PUFFIN used a deep residual network-based computational approach that quantifies uncertainty in peptide-MHC affinity prediction; The attention mechanism used by MHCAttnNet provided a heatmap over the peptide sequences [31]; And DeepSeaPanII was an end-to-end neural network model without the need for pre- or post-processing on input samples [32]. It is worth noting that the above DL-based methods only encode the text information of the peptide sequences and input them into the corresponding model, and do not take into account the structural information of the peptide sequence or its implicit presence in the network structure design [33].

Although some traditional methods utilize the structural information of the peptide sequences [34, 35], it is difficult to combine these deep learning algorithms properly. In recent research in the field of natural language processing (NLP) [36], positional encoding (PE) is used to encode the relative position of words in a sentence, allowing deep models to retain position information among words [37, 38]. Supplementing the structure information can effectively help the deep network to achieve better performance, especially those networks that are not sensitive to position information, such as the Transformer [39]. Apart from this, [40] showed that the structural information was crucial for modular reinforcement learning, substantially outperforming prior state-of-the-art methods on multi-task learning [41, 42]. Therefore, in the MHC-II affinity peptide prediction task, the introduction of position encoding can be expected to further improve the performance of the training deep-learning model, especially when the internal position of each peptide sequence is completely determined [43].

In this paper, the proposed algorithm is based on DeepMHCII, the current state-of-the-art DL-based algorithm [44][1], to validate our proposed strategy. To study the effectiveness of introducing positional encoding, this paper discusses the placement of positional encoding in different positions and the use of different encoding schemes. To intuitively compare the performance of different positional encoding-adding strategies, the same datasets and evaluation as the DeepMHCII algorithm are used, such as 5-fold cross-validation, independent testing set verification, and binding core prediction. Experimental results show that the introduction of positional encoding information can further improve the performance of the DeepMHCII model. We believe that introducing position encoding into this task can provide an important reference for future model optimization.

Methods

Preliminary

Consider two primary sequences: a peptide sequence denoted by P and an MHC-II molecule sequence denoted by Q. Both sequences consist of the 20 standard amino acids. Our objective is to establish a regression model that predicts the binding affinity $\hat{z} \in [0, 1]$ when given a specific pair of P and Q. The binding affinity between P and Q is primarily influenced by two factors:

- The peptide's binding core: The principal segment actively participates in interactions with the MHC-II molecule.
- The MHC-II molecule's binding groove: A region marked by its nine specialized pockets, which is paramount for the peptide's accommodation.

In addition, it should be noted that peptide flanking residues (PFRs) play a role. Although these residues reside outside the core binding groove, research has shown their significant impact. PFRs not only influence the binding affinity of the peptide, but also play a role in enhancing both peptide processing and T cell activation [45]. To facilitate our study and in alignment with prior research settings, we focus on a 34-residue pseudo sequence derived from Q. This representation of MHC-II molecules is an amalgamation of two parts: 15 amino acid residues from the α chain and 19 from the β chain of MHC-II. The extraction of these residues is based on

[1] https://github.com/yourh/DeepMHCII

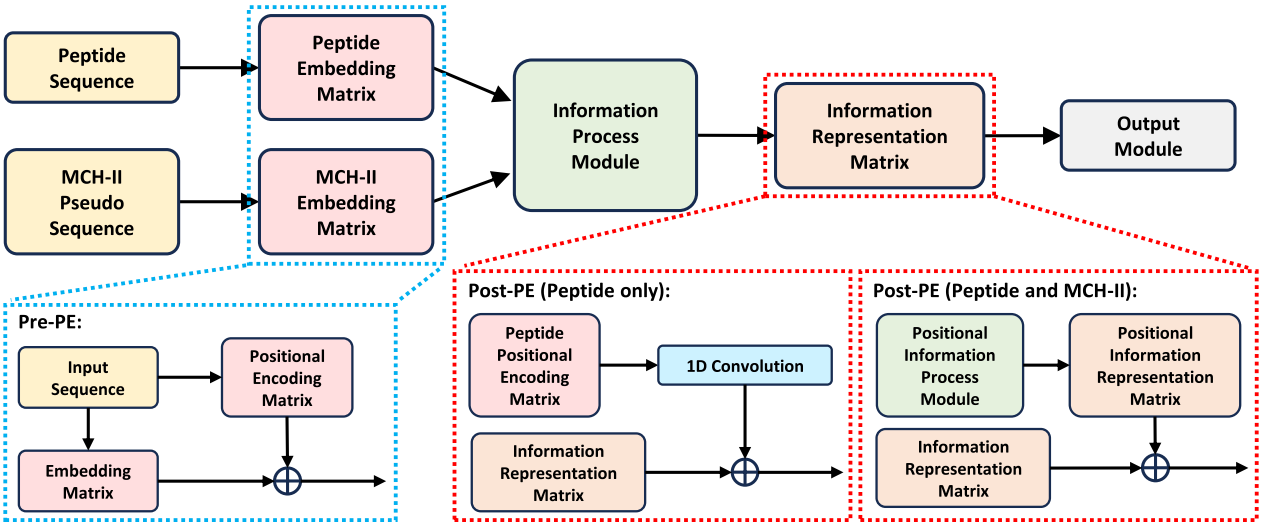

Fig. 1 The architecture of our proposed method. The blue dashed box denotes the Pre-PE condition and the red dashed box denotes the Post-PE condition

their presence in the MHC-II peptide complexes found in the Protein Data Bank (PDB) [46].

Overview of DL-based MHC-II binding prediction

The burgeoning field of immunological research has turned to Deep Learning (DL) methodologies, especially panspecific methods, to offer granular insights [47]. These methods can be distilled into a structured deep framework as shown in Fig. 1. Based on this, existing frameworks employ an embedding layer dedicated to encoding peptide sequences into an embedding matrix, $\mathbf{X} \in \mathbb{R}^{L_{peptide} \times d}$. Currently, another separate embedding layer focuses on translating the pseudo-sequences of MHC-II into a unified embedding matrix, $\mathbf{Y} \in \mathbb{R}^{L_{pseudo} \times d}$. After initial encoding, both embedding matrices undergo a transformation mediated by the **information process module**. In doing this, sophisticated structures are tailored to extract the underlying **Information Representation Matrix**, revealing the dynamic interplay between peptides and MHC-II molecules. Drawing the process to a close, the framework takes advantage of an output layer. Its primary objective is dual: to calculate the binding affinity, denoted by $\hat{z}$, and to discern the predictive scores associated with potential binding cores of nine lengths.

The purpose of this paper is to delve deeper into the architectural intricacies of this framework. Emphasis will be laid on the integration of structural data, with a dedicated segment elucidating the strategic application of positional encoding to bolster the prediction accuracy. When considering positional encoding (PE) strategies for peptide sequences in the context of machine learning, especially for tasks like predicting the binding affinity between peptides and MHC-II molecules, various strategies can be formulated by combining different aspects of positional encoding. These strategies could include variations in the position of the addition of encoding, the method of encoding, and the use of positional peptide encoding alone. Let us explore what each of these aspects means and how they can be combined.

How to implement the positional encoding

In this section, we explore the computation of positional encoding for the peptide sequence and the MHC-II pseudo-sequence. Let us consider a peptide sequence, P, and an MHC-II pseudo-sequence, Q. These sequences are mapped to their corresponding positional encodings, $\mathbf{X_{PE}} \in \mathbb{R}^{L \times d}$ and $\mathbf{Y_{PE}} \in \mathbb{R}^{34 \times d}$, representing the positional encoding matrices for P and Q, respectively. The MHC-II pseudo-sequence Q is a 34-length sequence extracted from the entire MHC-II molecule sequence. The positional encoding can be defined as:

$$\mathbf{X_{PE}} = (x_1, x_2, \ldots, x_L)^\top \in \mathbb{R}^{L \times d}, \quad \mathbf{Y_{PE}} = (y_1, y_2, \ldots, y_{34})^\top \in \mathbb{R}^{34 \times d}, \tag{1}$$

where $x_i \in \mathbb{R}^d$ denotes the positional encoding vector for the i-th residue of the peptide sequence, and $y_j \in \mathbb{R}^d$ represents the positional encoding vector for the j-th residue

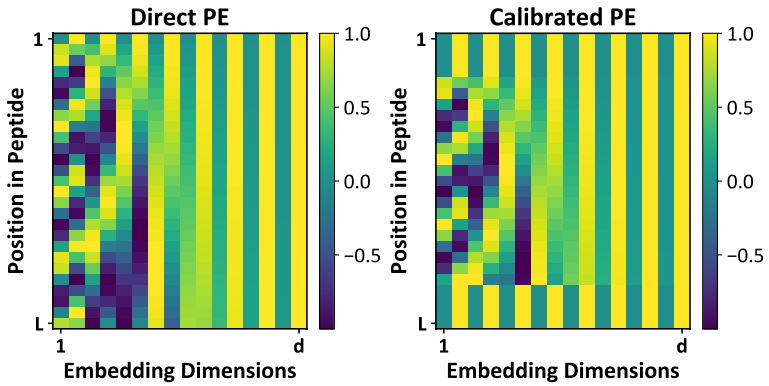

Fig. 2 Illustration of two different positional encodings: Direct PE and Calibrated PE

of the MHC-II pseudo-sequence. Here, L signifies the length of the input peptide sequence.

Due to the peptide-binding groove alignment on MHC-II molecules being open at both ends [48], which allows the peptide to extend beyond the binding groove (9-22 residues in length), practical sequences of equal length are achieved by zero padding at both ends of the peptide sequence. Thus, amino acids at these zero positions will produce invalid encoding which has been considered by the traditional peptide embedding process. Therefore, this article proposed a new positional encoding strategy for the encoding of the peptide sequence, Calibrated PE, based on Direct PE, as shown in Fig. 2. The positional encoding of a sequence P when using calibrated PE can be rewritten as follows:

$$\mathbf{X_{C-PE}} = (x_0, ..., x_0, x_1, ...x_{L'}, x_0, ..., x_0)^T \in \mathbb{R}^{L\times d}, \quad (2)$$

where L' denotes the truth length of the peptide sequence. For computing positional encoding, a fixed sinusoidal relative PE method that is derived from sinusoidal functions and fixed during the model training [49]. For example, consider the dimension $2k, (x_2, x_4, ..., x_{2k}, ...)$ and the dimension $2k+1, (x_1, x_3, ..., x_{2k+1}, ...)$ of a peptide-encoded PE, respectively:

$$\begin{aligned} \mathrm{PE}(pos, 2k) &= sin\left(\frac{pos}{1000^{2k/d}}\right) \\ \mathrm{PE}(pos, 2k+1) &= cos\left(\frac{pos}{1000^{2k+1/d}}\right) \end{aligned} \quad (3)$$

Where to add the positional encoding

The next thing to consider is where to add the positional encoding information. Positional encoding information can be inserted into the pre (pre-PE) or post (post-PE) information process module according to Fig. 1, respectively. Here, the information process module is the binding interaction convolutional layer (BICL) proposed by DeepMHCII. Briefly, BICL generates different kernels for each MHC-II molecule and performs a convolution operation on the peptide embedding matrix. On the other hand, we also consider whether the positional encoding is applied exclusively to the peptide sequence, which may highlight the importance of the peptide's position in binding without affecting the encoding of the MHC-II molecule, or to both the peptide and MHC-II sequences, allowing the model to learn the relative positions of both in the binding interaction.

As for Pre-PE, just simply add the position-encoding information matrix to the amino acid embedding matrix to make the trained model contain position-encoding information, shown in the blue dashed box of Fig. 1. Let $\mathbf{X_{Emb}} \in \mathbb{R}^{L\times d}$ and $\mathbf{Y_{Emb}} \in \mathbb{R}^{34\times d}$, representing the embedding matrices for P and Q, respectively. Thus, the combined embedding matrices which will undergo a transformation mediated by the Information Process Module can be given as follows:

$$\mathbf{X} = \mathbf{X_{Emb}} + \mathbf{X_{PE}}(or\ \mathbf{X_{C-PE}}) \in \mathbb{R}^{L\times d}, \quad \mathbf{Y} = \mathbf{Y_{Emb}} + \mathbf{Y_{PE}} \in \mathbb{R}^{34\times d}. \quad (4)$$

When it comes to Post-PE (the red dashed box of Fig. 1), things are going to get a little more complicated. As can be seen from DeepMHCII framework, the peptide sequence information representation matrix, denoted as the output of BICL can be written as:

$$C_{Emb} = f\left(f\left(W^k\mathbf{Y}\right)\mathbf{X} + b^k\right), \quad (5)$$

Where W^k is the weight matrix, b^k is the bias, f is the activation function, and k denotes different kernel sizes. Considering the effect of both the binding core and PFRs, BICL used four different kernel sizes (s^k): 9, 11, 13, and 15. For each kernel size, there is a different number of kernels (h^k). In the peptide-only strategy, a 1D convolution layer is used to learn the positional information representation matrix and add this to the output of BICL. To satisfy the

size with the output of BICL, the kernels (W^k_{PE-O}) of convolution hold the size of $h^k \times d \times s^k$ and the output of this layer can be described as follows:

$$C_{PE} = f\left(W^k_{PE-O} \cdot \mathbf{X}_{PE} + b^k_{PE-O}\right), \quad (6)$$

where b^k_{PE-O} is the bias, f is the activation function, and k denotes different kernel sizes. In both considering the peptide and the MHC-II strategy, the same structure of BICL was employed and the interaction between $\mathbf{X}_{PE}$ and $\mathbf{Y}_{PE}$ can be given as follows:

$$C_{PE} = f\left(f\left(W^k_{PE-T}\mathbf{Y}_{PE}\right) \cdot \mathbf{X}_{PE} + b^k_{PE-T}\right), \quad (7)$$

where W^k_{PE-T} with the size of $h^k \times s^k \times 34$ to generate the kernels and b^k_{PE-T} is the bias. Same as the Pre-PE condition, the combined information representation matrix can be given as follows:

$$C = C_{Emb} + C_{PE}, \quad (8)$$

Datasets

Three available benchmark datasets are used to train and evaluate our proposed method:

- BC2015: a binding core benchmark, which was used to evaluate the performance of NetMHCIIpan3.2[2] in identifying the binding core of an MHC-II peptide complex. BC2015 consists of 51 complexes from PDB.
- BD2016: It contains 134,281 data points on MHC-peptide binding affinities for 80 different MHC-II molecules, including 36 HLA-DR, 27 HLA-DQ, 9 HLA-DP and 8 H-2 molecules. BD2016 already provides a 5-fold cross-validation (5-fold CV) split that groups peptides with common motifs into the same fold.
- ID2017: an independent test dataset in DeepMHCII, ID2017, by removing data points that overlapped with BD2016 and retained MHC-II molecules with more than 50 peptides for robust performance evaluation. There are 10 HLA-DB molecules with 857 peptides in practice.

The following experiments will be conducted to validate the performance of our method: (i) the performance comparison among different PE-adding strategies on ID2017; (ii) the performance of different PE-adding strategies by 5-fold CV over BD2016; (iii) visualization of the binding motifs of MHC-II molecules obtained by each model as sequence logos; (iv) predict the binding core over BC2015.

We have set the minimization of the mean square error as our primary goal in training. To achieve this, we have implemented an ensemble learning strategy, wherein we trained T distinct models, each initialized with unique random weights. The final prediction is derived by calculating the mean of the predictions of all T models.

The area under the receiver operating characteristic curve (AUC) for each MHC-II molecule and the average AUC were reported. The Pearson correlation coefficient (PCC) was calculated to examine the linear relationship between the predicted binding affinity. Spearman rank correlation coefficient (SRCC) measured the monotonic relationship based on ranks. Furthermore, mean square error (MSE) was used to provide a measure of the average prediction error of the proposed strategy.

Results

Experimental settings

In this paper, the following hyperparameter values can be found, which are the same as DeepMHCII: $d = 16$. The number of kernels h^k with kernel sizes s^k of 9, 11, 13, and 15 was 256, 128, 64, and 64, respectively. f was ReLU. While training, the batch size was 128, the number of epochs was 20 and the optimizer was Adaelta [50] with a learning rate of 0.9 and weight decay of 1e-4. T (number of trained models) was 20. Apart from that, this paper discussed 8 types of combined PE-adding strategies: 2 PE-adding positions (where), 2 PE encoding methods (how), and consider whether to use peptide-only PE. Each strategy would have its advantages and could be tested empirically to see which yields the associated accurate predictions.

Comparison of different pe-adding strategies on ID2017

Table 1 offers a detailed analysis of the performance (AUC) on the ID2017 independent test set when different positional encoding (PE) strategies are applied to the baseline DeepMHCII model. The tables elucidate that Calibrated PE, particularly the Pre-PE(T) strategy, outperforms other configurations with an impressive AUC of 0.777, marking an enhancement over the baseline performance of DeepMHCII, which has an average AUC of 0.770. Examining individual allele performance further, the Calibrated PE method demonstrates a clear advantage. For instance, allele DRB1*0301 shows a significant increase from an AUC of 0.629 with DeepMHCII to 0.676 with Calibrated Pre-PE(T). Similarly, allele DRB1*0701's AUC escalates to 0.845 with Calibrated Post-PE(T), compared to 0.814 with the baseline. However, not all alleles react equally to the addition of PEs. Allele DRB1*0401, for example, maintains a higher AUC with the baseline

[2] http://www.cbs.dtu.dk/suppl/immunology/NetMHCIIpan3.2

Table 1 Performance (AUC) of using different positional encoding conditions and DeepMHCII on ID2017. 'O' denotes using peptide-PE only, and 'T' denotes using peptide-PE and MHC-II-PE together

Allele	DeepMHCII	Calibrated PE				Direct PE			
		Pre (O)	Pre (T)	Post (O)	Post (T)	Pre (O)	Pre (T)	Post (O)	Post (T)
DRB1*0101	0.882	0.868	0.873	**0.885**	0.871	0.883	0.883	0.875	0.877
DRB1*0301	0.629	0.602	**0.676**	0.622	0.632	0.628	0.620	0.629	0.610
DRB1*0401	**0.863**	0.813	0.790	0.854	0.814	0.807	0.778	0.810	0.816
DRB1*0701	0.814	0.792	**0.845**	0.802	0.823	0.812	0.821	0.825	0.821
DRB1*0901	0.889	0.875	0.844	**0.893**	0.859	0.856	0.844	0.841	0.842
DRB1*1101	0.657	0.642	0.649	**0.658**	0.648	0.652	0.641	0.628	0.633
DRB1*1202	0.788	0.758	**0.811**	0.763	0.725	0.788	0.733	0.716	0.735
DRB1*1301	0.615	0.566	**0.736**	0.636	0.660	0.517	0.651	0.643	0.572
DRB1*1501	0.799	0.798	0.821	0.807	**0.823**	0.810	0.798	0.817	0.806
DRB1*1502	**0.764**	0.752	0.728	0.734	0.700	0.726	0.720	0.679	0.687
Average	0.770	0.747	**0.777**	0.765	0.755	0.748	0.749	0.746	0.740

Table 2 Performance of using different positional encoding conditions and DeepMHCII on BD2016

Allele	DeepMHCII	Calibrated PE				Direct PE			
		Pre (O)	Pre (T)	Post (O)	Post (T)	Pre (O)	Pre (T)	Post (O)	Post (T)
AUC	0.856	0.844	0.820	0.855	0.856	0.843	0.820	**0.857**↑	0.855
PCC	0.691	0.679	0.634	0.689	0.693↑	0.675	0.633	**0.694**↑	0.692↑
SRCC	0.682	0.672	0.630	0.681	**0.687**↑	0.666	0.626	0.685↑	0.686↑
MSE	0.0308	0.0301↓	0.0337	0.0312	**0.0299**↓	0.0315	0.0341	**0.0299**↓	0.0301↓

DeepMHCII at 0.863, compared to 0.790 with Calibrated Pre-PE(T). Across the alleles, Calibrated PE's average AUC exhibits a notable increment compared to Direct PE's average AUC. More importantly, the same conclusion can be drawn from the results in Tables A1 and A2. These observations corroborate the inference that calibrated PE, especially the Pre-PE(T) strategy, is generally more beneficial than direct PE in improving the prediction accuracy of the DeepMHCII model on the ID2017 dataset.

Comparison of different PE-adding strategies on BD2016

Table 2 delineates the performance metrics of various positional encoding (PE) strategies in comparison to the baseline DeepMHCII on the BD2016 dataset. A careful examination of the figures reveals that Post-PE configurations generally surpass Pre-PE strategies across all measured criteria. This finding contrasts with the observations from the ID2017 dataset, where the Calibrated Pre-PE(T) strategy was notably effective. In particular, the Direct Post-PE(O) approach outperforms other strategies with the highest AUC of 0.857, which is marginally better than the baseline DeepMHCII's AUC of 0.856. Similarly, this strategy achieved the top Pearson Correlation Coefficient (PCC) of 0.694, slightly improving upon the baseline's 0.691. This trend continues with Spearman's Rank Correlation Coefficient (SRCC), where the Direct Post-PE(O) achieves an SRCC of 0.685, approaching the highest value in the table, 0.687, as seen with Calibrated Post-PE(T). Furthermore, the Mean Squared Error (MSE) metric also supports the superiority of Post-PE methods. Both Calibrated Post-PE(T) and Direct Post-PE(O) strategies share the lowest MSE of 0.0299, indicating a statistically significant reduction from the DeepMHCII baseline of 0.0308. These results highlight that, for the BD2016 dataset, the application of Post-PE, especially the Direct Post-PE(O) strategy, is particularly beneficial, surpassing the pre-encoding strategies and improving upon the baseline DeepMHCII model across multiple performance metrics. Detailed results for BD2016 are shown in Tables A3, A4 and A5. Besides, we further showed the PCC difference in Calibrated Post-PE (T) and Direct Post-PE (O) surpasses that in DeepMHCII (Fig. 3).

Binding core prediction and sequence logos

Furthermore, we have graphically represented the binding motifs of MHC-II molecules derived from each

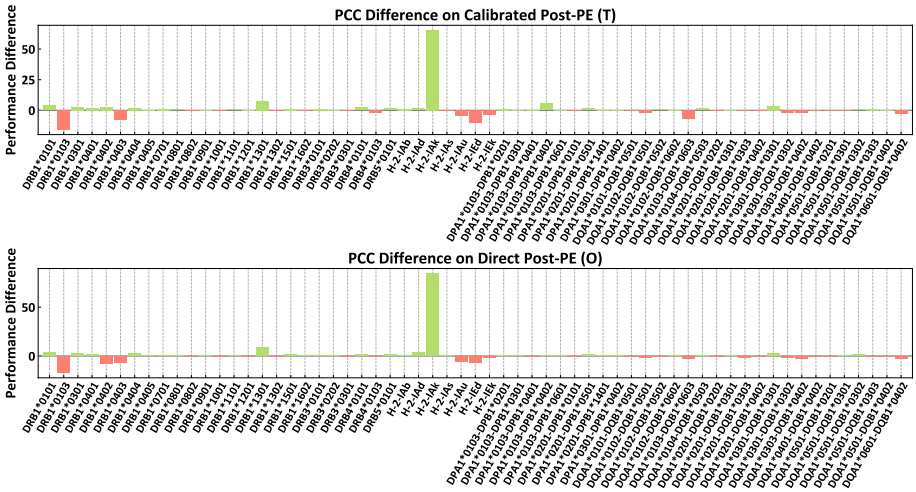

Fig. 3 PCC difference in Calibrated Post-PE (T) and Direct Post-PE (O) surpasses that in DeepMHCII

PE-adding strategy in the form of sequence logos, which are accessible at the WebLogo portal[3] [51, 52]. For illustrative purposes, we selected three MHC-II molecules-DRB10401, DRB10901, and DRB1*1202-from the ID2017 set and subjected them to random testing. Figure 4 displays the sequence logos corresponding to these MHC-II molecules as influenced by different encoding strategies. On the sequence logos, the x-axis encompasses positions 1 through 9 (referred to as pockets), where the overall height at each position is indicative of the relative informational significance attributed to that specific site within the motif. Concurrently, the stature of individual letters within each position correlates to the prevalence of the respective amino acid at that site. Typically, pockets 1, 4, 6, and 9 constitute the four main anchor positions, considered critical for the binding affinity of peptides [53]. The findings suggest that all strategies except for Direct Post-PE(O) exhibited substantial promise and conferred valuable insights pertinent to peptide binding. Upon evaluating the prediction accuracy of the binding core over the BC2015 dataset, the performance of Calibrated PE was observed to eclipse that of Direct PE, as evidenced by the data articulated in Tables A6 and A7.

Discussion

In this study, we introduced a novel approach to integrate structural information to enhance the predictions of binding affinity of MHC-II molecule using the DeepMHCII framework. The results from our proposed methodology indicate a noticeable improvement in DeepMHCII's performance, affirming the utility of positional encoding (PE) strategies in this domain. The comprehensive analysis provided herein lays a foundation for empirical understanding and sets a precedent for future investigative pursuits. Intriguingly, our findings underscore the variability in performance outcomes contingent upon the dataset and evaluative measures employed, thereby suggesting a necessity for meticulously tailored positional encoding techniques to accommodate different research requirements.

While our results are promising, they also reveal the heterogeneity of performance across different datasets and evaluation metrics, suggesting that a one-size-fits-all approach to positional encoding may not be viable. Instead, there is a compelling need for the careful design of positional information that aligns with specific tasks and datasets. Moreover, the landscape of positional encoding methodologies extends beyond the scope of our current work, with avenues such as learnable PEs presenting opportunities for further exploration. These adaptive encoding methods could potentially reveal more nuanced structural relationships within the binding affinities of MHC-II molecules.

Potential applications of this research are vast, including the development of more accurate predictive models for vaccine design, where understanding peptide-MHC-II interactions is crucial. Such models could substantially expedite the identification of potent epitopes, thereby bolstering the development of peptide-based vaccines and therapeutics. However, the limitations must be acknowledged. One such constraint is the reliance on available structural data, which may not fully capture the dynamic nature of peptide-MHC interactions. Future work could aim

[3] https://weblogo.berkeley.edu/

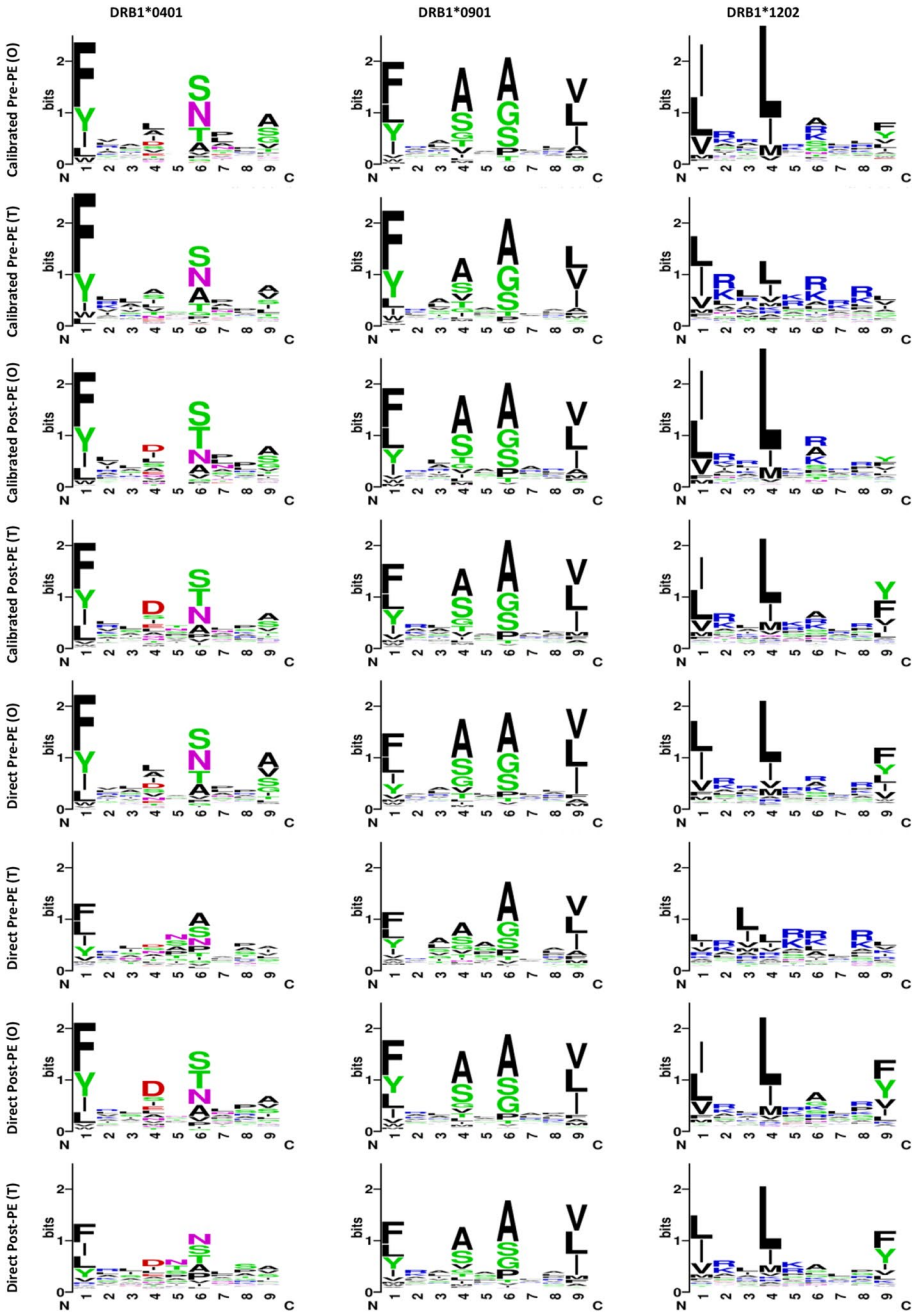

Fig. 4 Sequence logos by using different positional encoding

to incorporate three-dimensional spatial structures, potentially offering a more holistic view of the binding process. This could be achieved through the integration of molecular dynamics simulations or the application of advanced imaging techniques, further enhancing the predictive capabilities of deep learning models in this field. Ultimately, our work serves as a stepping stone toward the realization of deep learning methodologies that not only utilize positional information, but also encapsulate the rich structural intricacies inherent to biological processes. This could pave the way for a new era of bioinformatic tools capable of tackling complex biological predictions with greater accuracy and efficiency.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-023-09900-6.

Additional file 1.

Acknowledgements
We thanked Prof. Jia Yang for helping revise the manuscript.

Code availability
Code will be available on request.

Authors' contributions
Ying Yu: Writing the original draft. Lipeng Zu: Contribution methodology. Jiaye Jiang: Formal analysis. Yafang Wu: Visualization. N.S. Yinglin Wang: Data curation. Midie Xu: Review writing, editing. Qing Liu: Conceptualization, Supervision, Resources.

Funding
This work is supported in part by grants from National Natural Science Foundation of China (No. 32370092), from the National Key Research and Development Program (No. 2023YFF1103600) and supported by the Shanghai Science and Technology Development Fund (No. 22DZ2202500).

Availability of data and materials
All data generated or analyzed during this study are included in this published article; our manuscript does not contain any data, as no data set was generated or analyzed during the study.

Declarations

Ethics approval and consent to participate
This study did not include the use of animals, humans or otherwise, and therefore did not require ethical approval.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 30 August 2023 Accepted: 12 December 2023
Published: 30 January 2024

References

1. Wieczorek M, Abualrous ET, Sticht J, Álvaro Benito M, Stolzenberg S, Noé F, et al. Major Histocompatibility Complex (MHC) Class I and MHC Class II Proteins: Conformational Plasticity in Antigen Presentation. Front Immunol. 2017;8:1–16. https://doi.org/10.3389/fimmu.2017.00292.
2. Birnbaum ME, Mendoza JL, Sethi DK, Dong S, Glanville J, Dobbins J, et al. Deconstructing the Peptide-MHC Specificity of T Cell Recognition. Cell. 2014;157(5):1073–87. https://doi.org/10.1016/j.cell.2014.03.047.
3. Kelley J, Walter L, Trowsdale J. Comparative Genomics of Major Histocompatibility Complexes. Immunogenetics. 2005;56(10):683–95. https://doi.org/10.1007/s00251.004.0717.7.
4. Hue S, Ahern P, Buonocore S, Kullberg MC, Cua DJ, McKenzie BS, et al. Interleukin-23 Drives Innate and T Cell-Mediated Intestinal Inflammation. J Exp Med. 2006;203(11):2473–83. https://doi.org/10.1084/jem.20061099.
5. Jiang W, Boder ET. High-throughput Engineering and Analysis of Peptide Binding to Class II MHC. Proc Natl Acad Sci. 2010;107(30):13258–63. https://doi.org/10.1073/pnas.1006344107.
6. Tripathi NM, Bandyopadhyay A. High Throughput Virtual Screening (HTVS) of Peptide Library: Technological Advancement in Ligand Discovery. Eur J Med Chem. 2022;243:114766. https://doi.org/10.1016/j.ejmech.2022.114766.
7. Bravi B, Tubiana J, Cocco S, Monasson R, Mora T, Walczak AM. RBM-MHC: A Semi-Supervised Machine-Learning Method for Sample-Specific Prediction of Antigen Presentation by HLA-I Alleles. Cell Syst. 2021;12(2):195–202. https://doi.org/10.1016/j.cels.2020.11.005.
8. 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.
9. Fonseca AF, Antunes DA. CrossDome: An Interactive R Package to Predict Cross-Reactivity Risk Using Immunopeptidomics Databases. Front Immunol. 2023;14:1–15. https://doi.org/10.3389/fimmu.2023.1142573.
10. 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.
11. Moore MJ, Zhong M, Hansen J, Gartner H, Grant C, Huang M, et al. Humanization of T Cell-Mediated Immunity in Mice. Sci Immunol. 2021;6(66):eabj4026. https://doi.org/10.1126/sciimmunol.abj4026.
12. Lantz O, Bendelac A. An Invariant T Cell Receptor Alpha Chain Is Used by A Unique Subset of Major Histocompatibility Complex Class I-Specific CD4+ and CD4-8- T Cells in Mice and Humans. J Exp Med. 1994;180(3):1097–106. https://doi.org/10.1084/jem.180.3.1097.
13. Zhang L, Udaka K, Mamitsuka H, Zhu S. Toward More Accurate Pan-Specific MHC-Peptide Binding Prediction: A Review of Current Methods and Tools. Brief Bioinform. 2011;13(3):350–64. https://doi.org/10.1093/bib.bbr060.
14. Hu X, Zhou W, Udaka K, Mamitsuka H, Zhu S. MetaMHC: A Meta Approach to Predict Peptides Binding to MHC Molecules. Nucleic Acids Res. 2010;38(suppl_2):474–479. https://doi.org/10.1093/nar.gkq407.
15. Nielsen M, Lund O, Buus S, Lundegaard C. MHC Class II Epitope Predictive Algorithms. Immunology. 2010;130(3):319–28. https://doi.org/10.1111/j.1365-2567.2010.03268.x.
16. Stern LJ, Wiley DC. Antigenic Peptide Binding by Class I and Class II Histocompatibility Proteins. Structure. 1994;2(4):245–51. https://doi.org/10.1016/S0969.2126(00)00026.5.
17. Trowitzsch S, Tampé R. Multifunctional Chaperone and Quality Control Complexes in Adaptive Immunity. Annu Rev Biophys. 2020;49(1):135–61. https://doi.org/10.1146/annurev-biophys-121219-081643.
18. Barra C, Alvarez B, Paul S, Sette A, Peters B, Andreatta M, et al. Footprints of Antigen Processing Boost MHC Class II Natural Ligand Predictions. Genome Med. 2018;10(1):84. https://doi.org/10.1186/s13073-018-0594-6.
19. Frankiw L, Baltimore D, Li G. Alternative mRNA Splicing in Cancer Immunotherapy. Nat Rev Immunol. 2019;19(11):675–87. https://doi.org/10.1038/s41577-019-0195-7.
20. Reche PA, Glutting JP, Zhang H, Reinherz EL. Enhancement to the RANKPEP Resource for the Prediction of Peptide Binding to MHC Molecules using Profiles. Immunogenetics. 2004;56(6):405–19. https://doi.org/10.1007/s00251-004-0709-7.
21. Oyarzún P, Ellis JJ, Bodén M, Kobe B. PREDIVAC: CD4+ T-cell Epitope Prediction for Vaccine Design that Covers 95% of HLA Class II DR Protein Diversity. BMC Bioinformatics. 2013;14(1):52. https://doi.org/10.1186/1471-2105-14-52.
22. Zhang Q, Wang P, Kim Y, Haste-Andersen P, Beaver J, Bourne PE, et al. Immune Epitope Database Analysis Resource (IEDB-AR). Nucleic Acids Res. 2008;36(suppl_2):513–518. https://doi.org/10.1093/nar/gkn254.
23. Singh SP, Mishra BN. Major Histocompatibility Complex Linked Databases and Prediction Tools for Designing Vaccines. Hum Immunol. 2016;77(3):295–306. https://doi.org/10.1016/j.humimm.2015.11.012.
24. 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(1):238. https://doi.org/10.1186/1471-2105-8-238.
25. Bhasin M, Raghava GPS. SVM Based Method for Predicting HLA-DRB1*0401 Binding Peptides in An Antigen Sequence. Bioinformatics. 2004;20(3):421–3. https://doi.org/10.1093/bioinformatics/btg424.
26. He Y, Xiang Z, Mobley HLT. Vaxign: The First Web-Based Vaccine Design Program for Reverse Vaccinology and Applications for Vaccine Development. J Biomed Biotechnol. 2010;2010:297505. https://doi.org/10.1155/2010/297505.
27. Nielsen M, Lundegaard C, Blicher T, Peters B, Sette A, Justesen S, et al. Quantitative Predictions of Peptide Binding to Any HLA-DR Molecule of

Known Sequence: NetMHCIIpan. PLoS Comput Biol. 2008 07;4(7):1–10. https://doi.org/10.1371/journal.pcbi.1000107.
28. Guo L, Luo C, Zhu S. MHC2SKpan: A Novel Kernel Based Approach for Pan-Specific MHC Class II Peptide Binding Prediction. BMC Genomics. 2013;14(5):11. https://doi.org/10.1186/1471-2164-14-S5-S11.
29. 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):449–454. https://doi.org/10.1093/nar/gkaa379.
30. Cheng J, Bendjama K, Rittner K, Malone B. BERTMHC: Improved MHC-Peptide Class II Interaction Prediction with Transformer and Multiple Instance Learning. Bioinformatics. 2021;37(22):4172–9. https://doi.org/10.1093/bioinformatics/btab422.
31. 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(Supplement_1):399–406. https://doi.org/10.1093/bioinformatics/btaa479.
32. Liu Z, Jin J, Cui Y, Xiong Z, Nasiri A, Zhao Y, et al. DeepSeqPanII: An Interpretable Recurrent Neural Network Model With Attention Mechanism for Peptide-HLA Class II Binding Prediction. IEEE/ACM Trans Comput Biol Bioinforma. 2022;19(4):2188–96. https://doi.org/10.1109/TCBB.2021.3074927.
33. Zeng H, Gifford DK. Quantification of Uncertainty in Peptide-MHC Binding Prediction Improves High-Affinity Peptide Selection for Therapeutic Design. Cell Syst. 2019;9(2):159–66. https://doi.org/10.1016/j.cels.2019.05.004.
34. Dimitrov I, Garnev P, Flower DR, Doytchinova I. EpiTOP–A Proteochemometric Tool for MHC Class II Binding Prediction. Bioinformatics. 2010;26(16):2066–8. https://doi.org/10.1093/bioinformatics/btq324.
35. Zhang L, Chen Y, Wong HS, Zhou S, Mamitsuka H, Zhu S. TEPITOPEpan: Extending TEPITOPE for Peptide Binding Prediction Covering over 700 HLA-DR Molecules. PLoS ONE. 2012;7(2):1–10. https://doi.org/10.1371/journal.pone.0030483.
36. Jacob D, Ming-Wei C, Kenton L, Kristina NT. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.
37. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is All You Need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS). Red Hook; 2017. p. 6000–6010. https://doi.org/10.5555/3295222.3295349.
38. Wang H, Yin H, Zhang M, Li P. Equivariant and Stable Positional Encoding for More Powerful Graph Neural Networks. arXiv preprint arXiv:2203.00199.
39. Liu Z, Ning J, Cao Y, Wei Y, Zhang Z, Lin S, et al. Video Swin Transformer. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Los Alamitos: IEEE Computer Society; 2022. p. 3192–3201. https://doi.org/10.1109/CVPR52688.2022.00320.
40. Hong S, Yoon D, Kim KE. Structure-Aware Transformer Policy for Inhomogeneous Multi-Task Reinforcement Learning. In: International Conference on Learning Representations (ICLR). Virtual; 2022. p. 1–13.
41. Huang W, Mordatch I, Pathak D. One Policy to Control Them All: Shared Modular Policies for Agent-Agnostic Control. In: Proceedings of the 37th International Conference on Machine Learning (ICML), vol. 119. PMLR; 2020. p. 4455–4464.
42. Kurin V, Igl M, Rocktäschel T, Böhmer W, Whiteson S. My Body Is A Cage: The Role of Morphology in Graph-Based Incompatible Control. In: International Conference on Learning Representations (ICLR). Vienna; 2021. p. 1–14.
43. Bjellqvist B, Hughes GJ, Pasquali C, Paquet N, Ravier F, Sanchez JC, et al. The Focusing Positions of Polypeptides in Immobilized pH Gradients Can be Predicted from Their Amino Acid Sequences. Electrophoresis. 1993;14(1):1023–31. https://doi.org/10.1002/elps.11501401163.
44. 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(Supplement_1):220–228. https://doi.org/10.1093/bioinformatics/btac225.
45. Godkin AJ, Smith KJ, Willis A, Tejada-Simon MV, Zhang J, Elliott T, et al. Naturally Processed HLA Class II Peptides Reveal Highly Conserved Immunogenic Flanking Region Sequence Preferences That Reflect Antigen Processing Rather Than Peptide-MHC Interactions1. J Immunol. 2001;166(11):6720–7. https://doi.org/10.4049/jimmunol.166.11.6720.
46. Velankar S, Burley SK, Kurisu G, Hoch JC, Markley JL. The Protein Data Bank Archive. Struct Proteomics High-Throughput Methods. 2021:3–21. https://doi.org/10.1007/978-1-0716-1406-8_1.
47. Mayne J, Ning Z, Zhang X, Starr AE, Chen R, Deeke S, et al. Bottom-up proteomics (2013–2015): keeping up in the era of systems biology. Anal Chem. 2016;88(1):95–121.
48. 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(12):22. https://doi.org/10.1186/1471-2105-9-S12-S22.
49. Wei J, Ren X, Li X, Huang W, Liao Y, Wang Y, et al. NEZHA: Neural Contextualized Representation for Chinese Language Understanding. arXiv preprint arXiv:1909.00204.
50. Zeiler MD. ADADELTA: An Adaptive Learning Rate Method. arXiv preprint arXiv:1212.5701.
51. 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.
52. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: A Sequence Logo Generator. Genome Res. 2004;14(6):1188–90. https://doi.org/10.1101/gr.849004.
53. Rammensee HG, Bachmann J, Emmerich NPN, Bachor OA, Stevanović S. SYFPEITHI: Database for MHC Ligands and Peptide Motifs. Immunogenetics. 1999;50(3):213–9. https://doi.org/10.1007/s002510050595.

Publisher's Note