BMC Bioinformatics

RESEARCH

Open Access

Multimodal learning on heterogeneous subgraphs and LLMs representation for MHC-peptide binding affinity prediction

Ruimeng Li[1], Ying Wang[1], Haozhou Li[1], Biyi Zhou[1] and Qinke Peng[1*]

https://github.com/Limomo33/CMSH.git.

*Correspondence:
Qinke Peng
qkpeng@mail.xjtu.edu.cn
[1]Faculty of Electronic and Information Engineering, The Systems Engineering Institute, Xi'an Jiaotong University, Xi'an 710049, China

Abstract

Accurate prediction of MHC-peptide binding affinity remains a challenge for immunotherapeutic development. Existing methods struggle to jointly model functional semantics of polymorphic residues, evolutionary conservation constraints, and structural dynamic. We propose the Contrast learning-based Multi-feature Heterogeneous Subgraph model (CMHS) with sequence and structural representation. For sequence representation, we introduce LoRA fine-tuning to obtain the MHC-exclusive sequence representation from ESM2, then jointly BLOSUM50 to capture long-range functional dependencies and evolutionarily conserved residues. For structural representation, we use the biophysics-guided heterogeneous graph network. Constructing an MHC-peptide graph with a novel trainable Gaussian noise layer guided by crystallographic B-factors to dynamically simulate electron density uncertainty, coupled with a three-stage message-passing framework with subgraph aggregation, subgraph extraction and heterogeneous. Finally, to align sequence and graph representation spaces, we use contrastive learning to obtain a more comprehensive representation and to enhance the ability of model prediction. Evaluations on 16 HLA allele benchmarks show average SRCC improvements of 8.7%, with improvements of average AUC of 7.6%. This work establishes a new paradigm for predicting hypervariable immune interactions. The corresponding code can be founded in github.

Keywords Heterogeneous graph, Cross-attention, Edge-induced subgraph extract, GCN, Contrastive learning

Introduction

The Major Histocompatibility Complex (MHC) mediates adaptive immunity by presenting peptide fragments on cell surfaces for T-cell recognition [1]. However, its extreme genetic polymorphism encompassing over 20,000 documented alleles poses a significant challenge: experimental determination of binding affinities is infeasible at this scale due to combinatorial explosion [2–5].

Computational methods for predicting MHC-peptide binding affinities are generally divided into two categories: allele-specific and pan-allele methods [2]. While

allele-specific models train predictors for each allele, they fail to generalize across the polymorphic allele [4–6]. Pan-allele methods (e.g., the NetMHCpan series) leverage shared binding patterns. For instance, NetMHCpan1.0 to NetMHCpan3.0 are sequence-based tools using artificial neural networks [7, 8]. To enhance performance, NetMHCpan4.0 integrated mass spectrometry (MS) data to capture in vivo peptide repertoires [9, 10]. Other models, such as MHCflurry, extract features via fully connected layers [11]. However, MS data suffers from systematic deviations due to limitations in proteolytic efficiency, ionization, and detection sensitivity, and only a small subset of peptides have experimentally validated binding specificities. Moreover, MS data lacks structural information and evolutionarily conserved signals [12]. Thus, building a universal, robust, and interpretable MHC-peptide binding model solely on MS data remains challenging. Many models employ biophysical and model-based representations for sequence data. However, current representation schemes face limitations. Matrix-based encodings, such as BLOSUM50 in ACME capture residue substitution constraints but are processed by CNNs without explicitly modeling allele-specific functional semantics or spatial conformations [13]. Similarly, MHCSeqNet uses one-hot encoding combined with skip-gram methods (designed for natural language processing) to learn amino acid correlations and incorporates one-hot allele representations [14]. These approaches fail to capture the spatial topology of the binding interface.

Pseudo-sequence methods represent MHC-peptide binding patterns by selecting residues experimentally verified to bind peptides. However, compressing 3D binding sites into 1D residue indexes disrupts spatial relationships [15–17]. While methods like RPEMHC [18] and ConvNeXt-MHC [19] introduce residue-pair encodings to capture spatial interactions, they are confined to local contacts and neglect long-range and indirect contributions [20, 21]. Moreover, by focusing on single amino acids, they ignore the broader context within the MHC. Crucially, all static representations fail to model peptide-induced conformational changes, introducing systematic bias in flexible regions.

Recently, Large Language Models (LLMs) based on the Transformer framework have been applied to learn richer semantic representations, capturing functional and structural features [22, 23]. For example, pMHChat [12, 24] integrates ESM-MSA-1b and ESM-2 embeddings and uses BiLSTM to capture long-range dependencies in MHC sequences, addressing the limitations of traditional LSTMs in modeling remote associations of polymorphic residues. It also employs hypergraph neural networks for anchor prediction. Despite these advances, LLMs exhibit limitations: they are insufficiently sensitive to single-point variations (common in MHC alleles), and direct fusion of sequence embeddings with structural features creates representation space discordance, propagating learning bias.

To address these challenges, we propose CMHS (Contrastive Multi-feature Heterogeneous Subgraph), a deep learning framework for predicting MHC-peptide binding affinity. CMHS introduces three key innovations: To overcome the limitations of matrix-based representations, we jointly fine-tune ESM-2 on MHC sequences with BLOSUM50 constraints, synergizing conserved residue signals and long-range dependencies. To address the lack of structural data and spatial topology loss, we employ ESM-3 to predict MHC structures, modeling inter-domain couplings. We then construct MHC-peptide heterogeneous graphs incorporating a B-factor-guided trainable Gaussian noise layer to dynamically simulate electron density uncertainty. To resolve representation space

discordance, we use contrastive learning to align sequence and graph representations, eliminating modality bias.

Methods

Preliminaries

Denote $M = \{m^{(1)}, m^{(2)}, ..., m^{(n)}\}$ as MHC samples, $P = \{p^{(1)}, p^{(2)}, ..., p^{(q)}\}$ as peptides samples. Given $\left\{\left(m^{(i)}, p^{(i)}\right), a^{(i)}\right\}$ where $\left(m^{(i)}, p^{(i)}\right)$ is a MHC-peptide pair and $a^{(i)}$ is the corresponding afinity value. The main goal is to design a system that takes the MHC-peptide pair as the input to predict the affinity values as shown in the Fig. 1.

The embedding of MHC and peptides sequences

BLOSUM50 embedding

Firstly, we use BLOSUM50 to represent MHC and peptides. The BLOSUM matrices is one of the most useful representation methods calculated in large protein dataset and is able to evaluate the probability of replacement of each AA to deduce the physical properties [7, 8]. Considering the length of the peptide sequences ranges from 9 to 12, we use an encoded method to unify the input length [13].

LLM-based

BLOSUM50 is a static matrix and cannot capture context information. For instance, the function of an amino acid may vary at different positions, but BLOSUM50 only considers pairing substitution and does not take the surrounding environment into account. Second, it may not be able to handle polymorphic regions in the MHC that are highly variable but structurally critical. In addition, BLOSUM50 is based on distant homologous sequences and may not be applicable to rapidly evolving gene families like MHC. Large language models are widely used in protein representation due to their ability of context-aware dynamic representation. MHC molecules are highly polymorphic, and there are key functional sites (such as peptide-binding groove pockets) in the sequence. General pre-trained models may not be able to focus on these key regions or capture the functional implications of subtle differences among alleles. We use a large model representation combined with fine-tuning to characterize MHC sequences. ESM2, a transformer-based protein language model at scales from 8 million parameters up to 15

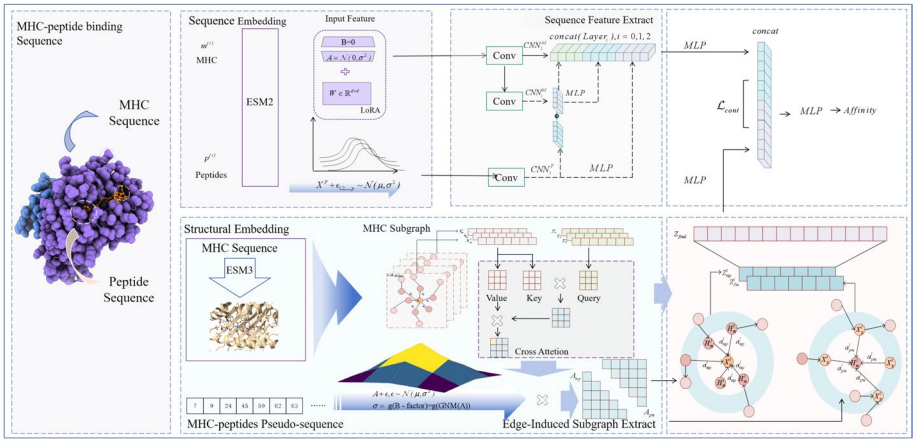

Fig. 1 Overall framework of CMHS model

billion parameters, is one of the widely used protein Masked Language Model (MLM) [24]. To reduce the fine-tuning parameters, we use the LoRA for ESM2 fine-tuning [25].

The BLOSUM50 embeddings of the MHC sequence is denoted as $X^{MB} = \{x_1^{MB}, x_2^{MB}, ..., x_n^{MB}\}$, $x_i^{MB} \in \mathbb{R}^{l \times d}$ represent the $i-th$ MHC sequence with length of *l*. The BLOSUM50 embeddings of the peptide sequence is denoted as $X^{PB} = \{x_1^{PB}, x_2^{PB}, ..., x_q^{PB}\}$, $x_i^{PB} \in \mathbb{R}^{p \times d}$ represent the $i-th$ peptides sequence with length of *p*.

The ESM2 of the MHC sequence is denoted as $X^{ME} = \{x_1^{ME}, x_2^{ME}, ..., x_n^{ME}\}$, $x_i^{ME} \in \mathbb{R}^{l \times e}$ represent the $i-th$ MHC sequence with length of *l*, $e = 320$. Let ϕ be the frozen pretrained ESM2 model with fine-tuning parameters θ. For output MHC sequence x_i^{ME}:

$$x_i^{ME} = \phi(p^{(i)}; \theta) \tag{1}$$

where $x_i^{ME} \in \mathbb{R}^{l \times e}$ is the output embedding. For any weight matrix $W \in \mathbb{R}^{d_a \times d_b}$ in ESM2 in Q/K/V projections, LoRA decomposes updates into low-rank matrices:

$$W_{new} = W + \Delta W = W + BA^T \tag{2}$$

where $\mathrm{A} \in \mathbb{R}^{d_b \times r}$ and $\mathrm{B} \in \mathbb{R}^{d_a \times r}$ is trainable parameters, $r \ll \min(d_a, d_b)$ is the LoRA rank which is hyperparameter. For input *M* to a self-attention layer:

$$\begin{aligned} Q &= M(W_Q + B_Q A_Q^T) \\ K &= M(W_K + B_K A_K^T) \\ V &= M(W_V + B_V A_V^T) \end{aligned} \tag{3}$$

Only $\mathrm{B}_Q, \mathrm{A}_Q, \mathrm{B}_K, \mathrm{A}_K, \mathrm{B}_V, \mathrm{A}_V$ are updated during fine-tuning.

The final output of embedding of MHC sequence can be denoted as $X^M \in \mathbb{R}^{m \times d+e}$:

$$X^M = concat(X^{ME}, X^{MB}) \tag{4}$$

The ESM2 embeddings of the peptide sequence is denoted as $X^{PE} = \{x_1^{PE}, x_2^{PE}, ..., x_q^{PE}\}$, $x_i^{PE} \in \mathbb{R}^{p \times e}$ represent the $i-th$ peptides sequence with length of *p*. Due to the short length of peptides, they are directly characterized using ESM2. Let ϕ be the frozen pretrained ESM2 model with parameters θ. For output peptides sequence x_i^{PE}:

$$x_i^{PE} = \phi(p^{(i)}) \tag{5}$$

where $x_i^{PE} \in \mathbb{R}^{p \times e}$ is the output embedding with frozen ESM2.

Same, the final output of embedding of MHC sequence can be denoted as $X^P \in \mathbb{R}^{p \times d+e}$:

$$X^P = concat(X^{PE}, X^{PB}) \tag{6}$$

Multi-level sequence feature pyramid-module

To simulate the recognition mechanism: MHC-peptide binding relies on both global structural complementarity and local matching of key anchor residues, we use a layer-by-layer fusion approach to extract the sequence features introduced in [13].

To simulate natural amino acid mutations in peptide-MHC interactions and enhances generalization to unseen alleles through controlled perturbation, we introduce the trainable knowledge-based learnable Gaussian noise module. This module enhances model robustness by injecting learnable Gaussian noise into the embedding space. Unlike traditional fixed noise, both the mean (μ) and standard deviation (σ) are dynamically generated by neural networks conditioned on input features. Specifically, For input of peptides embedding feature $X^P \in \mathbb{R}^{p\times d+e}$:

$$\tilde{X^P} = X^P + \epsilon \odot I, \epsilon_{1,2,\ldots,p} \sim \mathcal{N}(\mu, \sigma^2) \tag{7}$$

where, ϵ is the inject gaussian noise, I is the dimensionally adapted unit tensor, $\sigma = \exp(\log \sigma)$ and $\mu$is the trainable standard deviation parameter.

Then we used two layers of CNN to extract MHC sequence feature and a single CNN layer to extract the peptides sequence feature based on the different length. We have established feature link modules at each layer to obtain features of different dimensions. Denoted the *i*-th CNN layer for MHC as $CNN_i^M, i = 1, 2$, the CNN^P for peptide, the module can be defined as:

$$\begin{aligned} Layer_0 &= concat(X^M, X^P) \\ Layer_i &= concat(CNN_i^M, CNN^P) \end{aligned} \tag{8}$$

Final output of Multi-Level Sequence Feature Pyramid-module is:

$$S_{final} = concat(Layer_i), i = 0, 1, 2 \tag{9}$$

The embedding of MHC and peptides structure

The real structure data for MHC and peptides given by experiment is limitation compared with sequences data. Due to the efficiency of structure prediction model in protein domain, we use ESM3 model to predict the structure data for each MHC atom. To transform atomic-level information into residue-level representations that are more suitable for graph neural network processing, we calculate the 3-dimensional geometric center for each amino acid residue, based on the coordinates of its relevant atoms. This centroid coordinate represents the position of the residue.

The binding pratten we used the pseudo-sequence proposed by [7], which contain the position of binding sites obtained from biological experiments. We defined the MHC-peptides binding systems as heterogeneous graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A}, \mathcal{R}, \tau, \phi)$, thereinto, $\mathcal{V} = \mathcal{V}_M \cup \mathcal{V}_P$, $\mathcal{V}_M$ is the nodes in MHC, $\mathcal{V}_P$ is the nodes in peptide; $\mathcal{A} = MHC, Peptide$ is the type of nodes; $\mathcal{R} = MHC - MHC, MHC - Peptide, Peptide - MHC$ shows the type of edges, also, we have:

$$\tau(v) = \begin{cases} MHC & \forall v \in \mathcal{V}_M \\ Peptide & \forall v \in \mathcal{V}_P \end{cases} \tag{10}$$

$$\phi(e) = \begin{cases} MHC - MHC & if(\nu_i, \nu_j) \in \mathcal{E}_{mm} \\ MHC - Peptide & if(\nu_i, \nu_j) \in \mathcal{E}_{mp} \\ Peptide - MHC & if(\nu_i, \nu_j) \in \mathcal{E}_{pm} \end{cases} \tag{11}$$

where $\mathcal{E}_{mm} \subseteq \mathcal{V}_m \times \mathcal{V}_m$ represent the interior edges in MHC. $\mathcal{E}_{mp} \subseteq \mathcal{V}_m \times \mathcal{V}_p$ represent the outside edges from MHC to peptide, $\mathcal{E}_{pm} \subseteq \mathcal{V}_p \times \mathcal{V}_m$ represent the outside edges from peptide to MHC.

The definition of $\mathcal{E}_{mm}$ based on the result from ESM3. Specifically, for each C_α atom, calculate the Euclidean distance between it and other C_α atoms, and then generate an adjacency matrix based on the threshold we defined as 8 Å (if the distance is less than the threshold, the corresponding position is 1; otherwise, it is 0). Note that the adjacency matrix should be symmetrical and have a diagonal of 0. The prediction structure data often introduce bias and noise, limiting effectiveness [26].

Assumed that the deformation field is constant locally while docking, the density around the molecule moves in a similar way. The deformation in the continuous molecular docking can be described as the displacement of the Gaussian center [27]. Here, we introduce trainable 2-dimensional Gaussian noise in $\mathcal{E}_{mp}$ to $\mathcal{E}_{pm}$ simulate the flexibility of side chain, and enhance the robustness. Specifically:

$$\tilde{A} = A + \epsilon, \epsilon \sim \mathcal{N}(\mu, \sigma^2) \tag{12}$$

$\epsilon$is implemented through trainable parameters with μ and σ. The μ parameter represents the distance distribution between the C_α atom and the actual anchor positioning, calculate by:

$$\mu = f(A) \tag{13}$$

σ parameter fuses the crystallographic B factor to simulate the uncertainty of the electron density cloud, calculate by:

$$\sigma = g(B - factor) = g(GNM(A)) \tag{14}$$

g is the transform function, introduced by the relationship between factor-B and the standard deviation of atomic shift in crystallography as inverse. Due to the nonlinear threshold response of the flexible change of residues, which shows a tendency towards saturation at positions with large distances, we select the sigmoid form to model the relationship between sigma and the B-factor.

$$g(B) = \sigma_{\min} + (\sigma_{\max} - \sigma_{\min}) \times \frac{1}{1 + e^{-k(B - B_m)}} \tag{15}$$

here, $\sigma_{\min}$, $\sigma_{\max}$, B_m and k is the hyperparameters. *GNM* denote the Gaussian Network Model (GNM), introduced in [28] is able to predict the B-factor:

$$\Gamma_{ij} = \begin{cases} -1 & if\ i \neq j\ and\ A_{ij} > 0 \\ 0 & if\ i \neq j\ and\ A_{ij} < 0 \\ -\sum_{k \neq i} \Gamma_{ik} & if\ i = j \end{cases} \tag{16}$$

$$B_i = \frac{k_B T}{\gamma} [\Gamma^{-1}]_{ii} \tag{17}$$

where k_B is the Boltzmann constant, T denote the temperature constant and γ is the force constant.

Subgraph extraction-based heterogeneous graphs

MHC structures and their interactions with peptides is the hierarchical organization. To decouple intrinsic MHC structural representation learning from peptide-induced

binding interactions, while enabling effective information exchange between the two components, we formulate the binding system as a subgraph extraction-based heterogeneous graph.

Specifically, the MHC molecule is first represented as an independent subgraph, where intra-MHC residue interactions are modeled using a graph attention network (GAT) to capture its structural topology and long-range dependencies. Subsequently, MHC-peptide interactions are incorporated by introducing a cross-attention mechanism between the MHC subgraph and the peptide residues. The resulting cross-attention scores are used to construct an interaction-aware adjacency matrix, which explicitly encodes peptide-MHC coupling. Based on this adjacency structure, a graph convolutional network (GCN) is applied to jointly propagate information across the peptide and MHC subgraph, enabling effective integration of structural context and binding-induced perturbations. The final graph representation is then used to predict peptide-MHC binding affinity.

Prior attention-guided MHC subgraph module

When MHC interacts with small peptides, it is often the interaction of microenvironments with adjacent spatial structures. To obtain the characteristics of adjacent nodes within the spatial structure of MHC, we first use the Graph Attention Network (GAT) to calculate the MHC subgraph under the heterogeneous graph framework.

For this part, we directly use the BLOSUM50 encoding result as the node representation to reduce the network depth of the graph part. For each MHC node $v_m \in \mathcal{V}_m$, calculate its interaction attention:

$$H_m = GAT(X^{MB}, A_{mm}) \tag{18}$$

$$A_{mm}^{ij} = \begin{cases} Adjacency(v_i, v_j) & \tau(v_i) = \tau(v_i) = MHC \\ 0 & otherwise \end{cases} \tag{19}$$

$H_m^{(l+1)}$ is the output of GAT layer, and $Adjacency(v_i, v_j)$ is the adjacency relationship calculated from the structure predicted by ESM3. The implementation of the GAT layer is:

$$h_i^{(l+1)} = \sum_{k=1}^{K} \sigma \left(\sum_{j \in \mathcal{N}_i} \alpha_{ij}^k W^k h_j^{(l)} \right) \tag{20}$$

$\alpha_{ij}^k$is the attention weight of the *K*-th attention head:

$$\alpha_{ij} = \frac{\exp(Leaky\,\mathrm{Re}\,LU(a([Wh_i||Wh_j])))}{\sum_{k \in \mathcal{N}_i} \exp(Leaky\,\mathrm{Re}\,LU(a([Wh_i||Wh_k])))} \tag{21}$$

Cross-attention-based edge-induced subgraph extraction module

Our heterogeneous graph is defined as a graph containing both MHC and peptide nodes. Now, we need to extract the subgraphs in the MHC-peptide and peptide-MHC relationships. The subgraphs include all nodes in peptide and the top-k nodes in MHC that are closest to the peptide node. Based on the definitions extracted from the Edge-Induced

subgraphs in the graph, we provide the definitions: Given a heterogeneous graph $G = (\mathcal{V}, \mathcal{E})$ with:

- Vertex set $\mathcal{V} = \mathcal{V}_M \cup \mathcal{V}_P$ ($\mathcal{V}_M$: MHC nodes, $\mathcal{V}_P$: peptide nodes)
- Edge set $\mathcal{E} \subseteq \mathcal{V}_M \times \mathcal{V}_P$ (MHC-peptide edges only)

Let $\text{score} : \mathcal{E} \to \mathbb{R}$ be an edge weighting function and $k \in \mathbb{N}^+$ a predefined constant. The subgraph $G_{\text{sub}} = (\mathcal{V}_{\text{sub}}, \mathcal{E}_{\text{sub}})$ is constructed as:

1. Preserve all peptide nodes: $\mathcal{V}_p \subseteq \mathcal{V}_{\text{sub}}$
2. For each $p \in \mathcal{V}_p$, select top-k incident edges:
   $\mathcal{E}_p = \{e \in \mathcal{E} \mid e = (m, p) \wedge m \in V_m \wedge \text{rank}_p(e) \leq k\}$
3. Vertex set:
   $\mathcal{V}_{\text{sub}} = \mathcal{V}_p \cup \left\{ m \in \mathcal{V}_m \mid \exists p \in \mathcal{V}_p : (m, p) \in \bigcup_{q \in \mathcal{V}_p} \mathcal{E}_q \right\}$
4. Edge set: $\mathcal{E}_{\text{sub}} = \bigcup_{p \in \mathcal{V}_p} \mathcal{E}_p$

where $\text{rank}_p(e)$ denotes the rank of edge e when sorting all edges incident to p by $\text{score}(e)$ in descending order. Define the dynamic adjacency matrix generation function:

$$E_{MHC-peptide} = \text{CrossAttn}(\text{H}_m, X_{PB}) \tag{22}$$

$$e^{ij} = \frac{\exp\left(\text{LeakyReLU}\left(a^T \left[\text{Wh}_m^{(i)} \| \text{Wh}_p^{(j)}\right]\right)\right)}{\sum_{k \in \mathcal{V}_p} \exp\left(\text{LeakyReLU}\left(a^T \left[\text{Wh}_m^{(i)} \| \text{Wh}_p^{(k)}\right]\right)\right)} \tag{23}$$

Based on the edge weights, we construct two adjacency matrices. For each peptide node $\mathcal{P} \subseteq \mathcal{V}_p$, select the top-k MHC nodes, that is, to get the induced subgraph$\mathcal{G}_{\text{sub}} = (\mathcal{V}_{\text{sub}}, \mathcal{E}_{\text{sub}})$, where $\mathcal{V}_{\text{sub}} = \mathcal{P} \cup \{m \in \mathcal{V}_m \mid \exists p \in \mathcal{P} : (m, p) \in \mathcal{E}_{\text{sel}}\}$, $\mathcal{E}_{\text{sub}} = \bigcup_{p_j \in \mathcal{P}} \{(m, p_j) \mid m \in \text{Top}_k(p_j)\}$. To be specific:

$$\text{A}^{ij} = \begin{cases} \tau(m_i, p_j) & \text{if } m_i \in \text{Top}_k(p_j) \\ 0 & \text{otherwise} \end{cases} \tag{24}$$

$$\text{Top}_k(p_j) = \underset{m \in \mathcal{N}(p_j)}{\arg \text{ top-k}} \quad \tau(m, p_j) \tag{25}$$

Adjacency matrix from peptide to MHC and MHC to peptide defined as: $\text{A}_{mp} \in \mathbb{R}^{|\mathcal{V}_m| \times |\mathcal{V}_p|}$ (MHC-peptide), $\text{A}_{pm} \in \mathbb{R}^{|\mathcal{V}_p| \times |\mathcal{V}_m|}$ (peptide-MHC). We define the link relationship as $A = k(A_{pm} + A_{mp})$, $k = 0.5$.

GCN module

We use GCN to extract the two types of relationship graph respectively. For (MHC-peptide) and (peptide-MHC) graph, we have:

$$\begin{aligned} \text{GCN}_{mp}(\text{H}_m, X_{PB}) &= \sigma\left(\hat{\mathbf{D}}_{mp}^{-1/2} \text{A}_{mp} \hat{\mathbf{D}}_{pm}^{-1/2} \text{H}_{mp} \mathbf{\Theta}_m\right) \\ \text{GCN}_{pm}(\text{H}_m, X_{PB}) &= \sigma\left(\hat{\mathbf{D}}_{pm}^{-1/2} \text{A}_{pm} \hat{\mathbf{D}}_{mp}^{-1/2} \text{H}_{pm} \mathbf{\Theta}_p\right) \end{aligned} \tag{26}$$

where $\hat{\mathbf{D}}_{mp,pm}$ is degree matrix, $\mathbf{\Theta}$ is the parameter and σ is the activation function. The specific calculation is as follows:

$$\begin{aligned}\hat{\mathbf{D}}_{mp} &= \mathrm{diag}(\sum_{j} \mathrm{A}_{mp}^{ij}) \\ \hat{\mathbf{D}}_{pm} &= \mathrm{diag}(\sum_{i} \mathrm{A}_{pm}^{ji})\end{aligned} \tag{27}$$

The two subgraphs represent the graph aggregation relationships under different paths ((MHC-peptide) and (peptide-MHC)). Further, we use a weighted approach to fuse these two parts of information to obtain the final result.

$$\begin{aligned}\mathrm{Z}_{mp} &= \mathrm{GCN}_{mp}(\mathrm{H}_m, \mathrm{H}_p) \\ \mathrm{Z}_{pm} &= \mathrm{GCN}_{pm}(\mathrm{H}_m, \mathrm{H}_p) \\ \mathrm{Z}_{\mathrm{final}} &= \alpha \cdot \mathrm{Z}_{mp} + \beta \cdot \mathrm{Z}_{pm}\end{aligned} \tag{28}$$

α, $\beta = 1 - \alpha$ are parameters

Contrastive learning-based multi-feature module

The sequence features are based on the amino acid level and the structural feature is based on the spatial structure of atoms, direct fusion risks modality misalignment. We bridge this gap through contrastive learning, enforcing consistency between sequence and structural embeddings [29]. First, we map the two representations to the contrastive learning space through a projection head:

$$\begin{aligned}\mathrm{H}_g &= \mathrm{LayerNorm}(\mathrm{ReLU}(\mathrm{Z}_{final}\mathrm{W}_g + \mathrm{b}_g)) \\ \mathrm{H}_s &= \mathrm{LayerNorm}(\mathrm{ReLU}(\mathrm{S}_{final}\mathrm{W}_s + \mathrm{b}_s))\end{aligned} \tag{29}$$

Here, W_g, b_g, W_s, and b_s is the is the trainable parameters within the linear layer.The two parts of features are ultimately fused through the adaptive weight module by the linear layer:

$$\mathrm{z}^{(i)} = Linear(\mathrm{z}_g^{(i)}, \mathrm{z}_s^{(i)}) \tag{30}$$

Loss function

The representation of the same group of MHC-peptides were taken as positive sample pairs, denote as $\mathcal{P} = \{(i,i) \mid i \in \mathcal{B}\}$, and two characterizations of different MHC-peptides were taken as negative sample pairs, denote as $\mathcal{N} = \{(i,j) \mid i,j \in \mathcal{B}, i \neq j\}$.

$$\begin{aligned}\mathcal{L}_{\mathrm{cont}} = &\frac{1}{|\mathcal{P}| + |\mathcal{N}|}[\sum_{(i,i)\in\mathcal{P}} 1 - \mathrm{S}_{ii} \\ &+ \sum_{(i,j)\in\mathcal{N}} \max(0, \mathrm{S}_{ij} - \mathrm{margin})]\end{aligned} \tag{31}$$

where S is the distance matrix, calculated by cosine similarity denoted as:

$$\mathrm{S}_{ij} = \frac{1}{\tau}\left\langle \hat{\mathrm{h}}_g^{(i)}, \hat{\mathrm{h}}_s^{(j)} \right\rangle = \frac{1}{\tau}(\hat{\mathrm{h}}_g^{(i)})^{\top}\hat{\mathrm{h}}_s^{(j)} \tag{32}$$

Our final objective function integrates two complementary loss components, including supervised regression loss and contrastive alignment loss:

$$L_{total} = \lambda L_{reg} + \beta L_{cont} \tag{33}$$

where L_{reg} is calculated by Mean squared error (MSE) between ground-truth binding affinities and predicted values:

$$L_{reg} = \frac{1}{N}\sum_{i=1}^{N}(\hat{z}_i - z_i)^2 \tag{34}$$

Result

Datasets

We used the experimentally verified dataset downloaded from Immune Epitope Database(IEDB), which is the most widely used binding affinity (BA) dataset. The dataset covered 80 MHC alleles (43 Human and 12 non-human) with peptides of length 8, 9, 10, 11 for all alleles, and affinity data is given as IC50 values [7, 8, 13]. We further expand MHC data by downloading the HLA-C MHC molecule data and other animal MHC data, including chimpanzee, mouse, and rat, from the Ontobee website, which contains most complete MHC alleles data (https://www.iedb.org/database_export_v3.php).

The peptide-MHC binding affinities are represented by the IC50 values in nM units. We also test the model on the 40 latest weekly benchmark datasets (2021.1.1–2022.11.24), to verify the generality of our model. The datasets are also from IEDB (http://tools.iedb.org/auto_bench/mhci/weekly/). In order to retain the information of the original sequence as much as possible and avoid interference of excessive redundancy, 300 units of amino acids were selected as the final sequence length of MHC I.

Training set

To evaluate the performance of our CMHS model, we used Pearson Correlation Coefficient (PCC) and Spearmans rank correlation coefficient (SRCC) between predicted and measured binding affinities.

$$\begin{aligned} \mathrm{SRCC} &= 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \\ \text{where} \quad d_i &= \mathrm{rank}(x_i) - \mathrm{rank}(y_i) \end{aligned} \tag{35}$$

$$\mathrm{PCC} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{36}$$

Specifically, we use the a threshold of 500 nM to convert IC50 values into one or zero labels. Undering this setting, we used Area Under the Receiver Operating Characteristic Curve (AUROC), Accuray (ACC) as evaluation metrics. PCC measures linear correlation between predicted and true values defined as:

$$\begin{aligned} \mathrm{AUC} &= \int_0^1 \mathrm{TPR}(f) \cdot \left| \frac{d\mathrm{FPR}(f)}{df} \right| df \\ \text{with} \quad \mathrm{TPR} &= \frac{TP}{TP+FN}, \quad \mathrm{FPR} = \frac{FP}{FP+TN} \end{aligned} \tag{37}$$

$$\mathrm{ACC} = \frac{TP+TN}{TP+TN+FP+FN} \tag{38}$$

Due to equipment limitations, we adopted performance benchmarks from literature to enable unbiased comparisons. To ensure robust evaluation, we divided the dataset into a training set (80%) and a validation set (20%). A separate test dataset, excluded from model training, was reserved solely for final performance testing. The training dataset underwent 5-fold cross-validation (5-CV) to optimize and validate the model. We use the AdamW optimizer. The epoch is selected as 200 based on the process of the loss during the training Fig. 2. We compared these results with some popular and state-of-the-art methods, including MHCflurry, NetMHCpan 4.1, ANN 4.0, ACME.

- ANN 4.0: Gapped sequence alignment using artificial neural networks for MHC-I and peptide binding affinity prediction [30].
- MHCflurry: A PSSM-based method that predicts binding affinity using sequence similarity to characterized alleles [31].
- NetMHCpan 3.0: An artificial neural network-based method that integratesmass spectrometry-eluted ligands for MHC class I molecules, supporting peptides of length 8-11 [8].
- NetMHCpan 4.1: An artificial neural network-based method that integrates quantitative binding data and mass spectrometry-eluted ligands for a wide range of MHC class I molecules, supporting peptides of length 8-11 [10].
- ACME: Attention-based Convolutional neural networks for MHC Epitope binding prediction model [13].
- RPEMHC: A deep learning-based residue-residue pair encoding method for MHC-peptides affinity prediction [18].

5-CV result on the IEDB MHC class I binding affinity dataset

We use $5-$ CV to train the model, with the loss curve is shown in the Fig. 2, on the IEDB dataset on different length of peptides, including 8-mer, 9-mer, 10-mer and 11-mer. We also compared the performance of our model and state-of-the-art methods to shows the efficiency of our model.

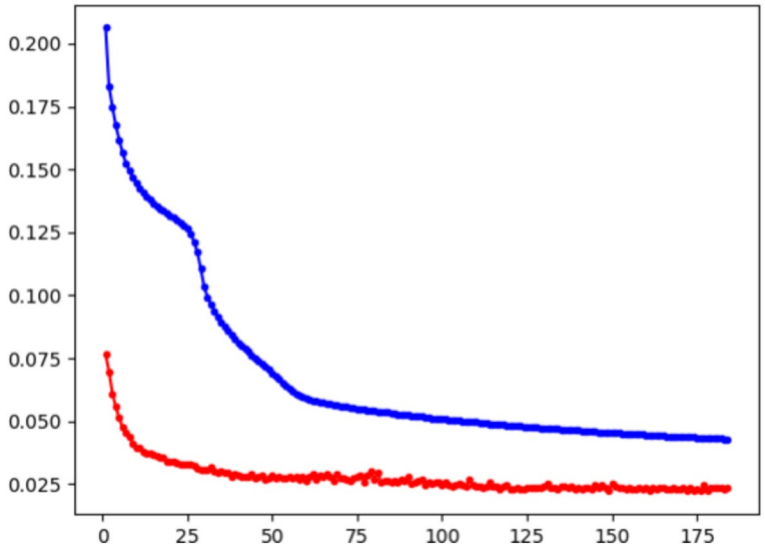

Fig. 2 5-CV Training Loss for CMHS in training process. Red line indicate the val loss, blue line is the train loss

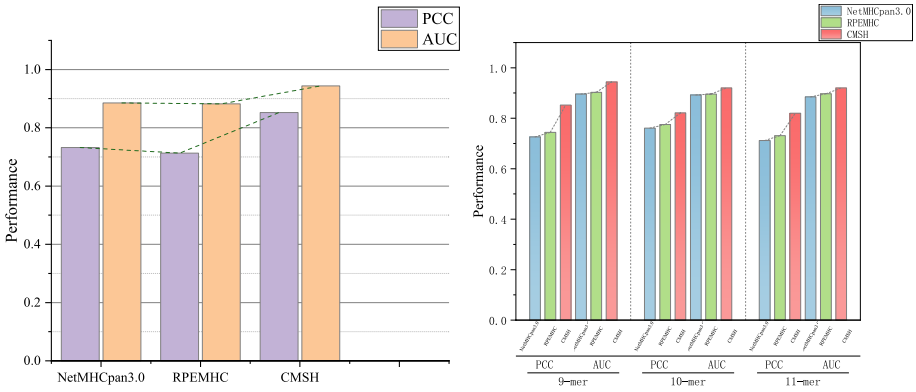

Fig. 3 5-CV Training result for CMHS. Lift figure shows 5-CV Training Performances for CMHS on IEDB comparing to SOTA Model. Right figure shows performances for CMHS and SOTA on different length of peptides

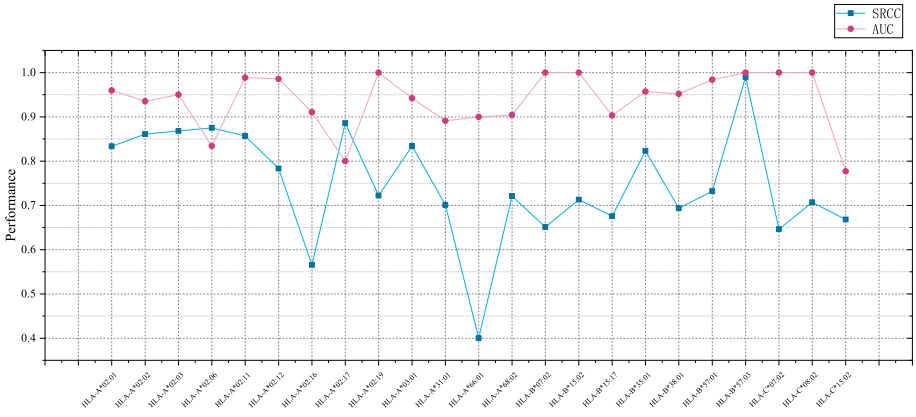

Fig. 4 5-CV Training result of performances for CMHS on different MHC alleles

The average performance comparison between SOTA and CMHS for different peptide lengths is shown in Fig. 3. For PCC, CMHS achieves 0.8526, outperforming ACME (0.81837) by 4.1% and significantly exceeding NetMHCpan3.0 (0.73233) and RPEMHC (0.7132); For AUROC, CMHS reaches 0.9436, 0.6% higher than ACME (0.9353) with enhanced robustness for all peptides; For ACC, CMHS sets a new record at 0.8911, 1.2% above ACME (0.89077).

Also, we give the result of testing on the IEDB test dataset, demonstrates that CMHS achieves favorable performance across varying peptide lengths, including 9, 10, 11-mer in Fig. 3 right. For 9-mer predictions, CMHS achives a PCC of 0.8526 and AUC of 0.9436, outperforming NetMHCpan3.0: PCC of 0.726, AUC of 0.896) and RPEMHC: PCC of 0.744, AUC of 0.903. With longer peptides, CMHS maintains robust results: 10-mer PCC reaches 0.8215, which has 5.6–6.0 percentage points higher than baselines, while 11-mer AUC sustains at 0.9201 while 0.885 for NetMHCpan3.0. These observations indicate that the CMHS exhibits improved adaptability to diverse peptide-length.

Comparsion of baselines

Additional testing across 16 alleles for 9-10 mer peptides indicates that CMHS achieves relatively favorable predictive performance in Fig. 4. As shown in Table 1, CMHS attains a PCC value of 0.753, showing improvement over existing methods and exceeding the

Table 1 Overall performance of CMHS and baseline methods on MHC-peptide binding prediction

Name	Performance	
	SRCC	**AUC**
ANN4.0	0.559	0.784
IEDB Consensus [32]	0.569	0.8
NetMHCcons [33]	0.595	0.813
NetMHCpan 3.0	0.59	0.81
NetMHCpan 4.0	0.588	0.815
PickPocket [34]	0.528	0.811
SMM [35]	0.578	0.804
SMMPMBEC [36]	0.605	0.824
ACME [13]	0.666	0.854
CMHS(Ours)	0.753	0.93

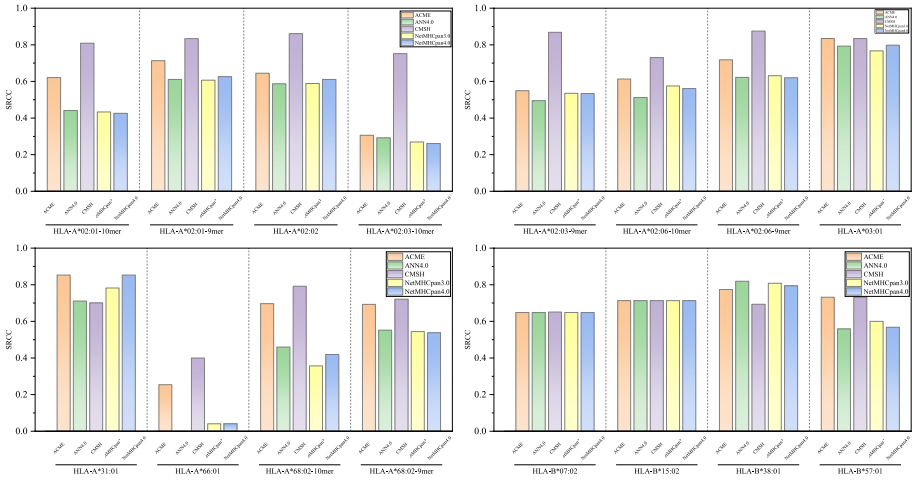

Fig. 5 Testing result for CMHS. SRCC Performances for CMHS on IEDB comparing with SOTA Model. Purple line corresponding to the CMHS

second-best performer ACME of 0.666 by 8.7%. For AUC metrics, CMHS reaches 0.93, outperforming other benchmark models such as NetMHCpan4.0 of 0.815 and SMMPMBEC of 0.824. It is noteworthy that conventional approaches like PickPocket PCC of 0.528 and ANN4.0 with PCC of 0.559 demonstrate relatively limited efficacy in multi-allele dataset.

Besides, detailed testing across 16 alleles indicates that CMHS delivers relatively consistent predictive performance for SRCC shown in Fig. 5, AUC shown in Fig. 6. Taking HLA-A*02:01 as an example, for 10-mer predictions, CMHS achieves an SRCC of 80.9%, representing an 18.8 percentage point improvement over ACME eith 62.1%. In 9-mer scenarios, its AUC 95.96% exceeds NetMHCpan4.0 58.2% by 37.76 percentage points. For highly polymorphic HLA-A*02:03, CMHS attains 75.19% SRCC on 10-mers, outperforming baseline methods by approximately 45 percentage points. Notably, on data-scarce HLA-A*66:01, CMHS achieves 40% SRCC percentage points higher than the second-best method, while conventional approaches like ANN4.0 yield negative correlation -28.7%. Although performance converges on prevalent alleles like HLA-B*07:02, CMHS maintains comparatively favorable stability across most complex scenarios.

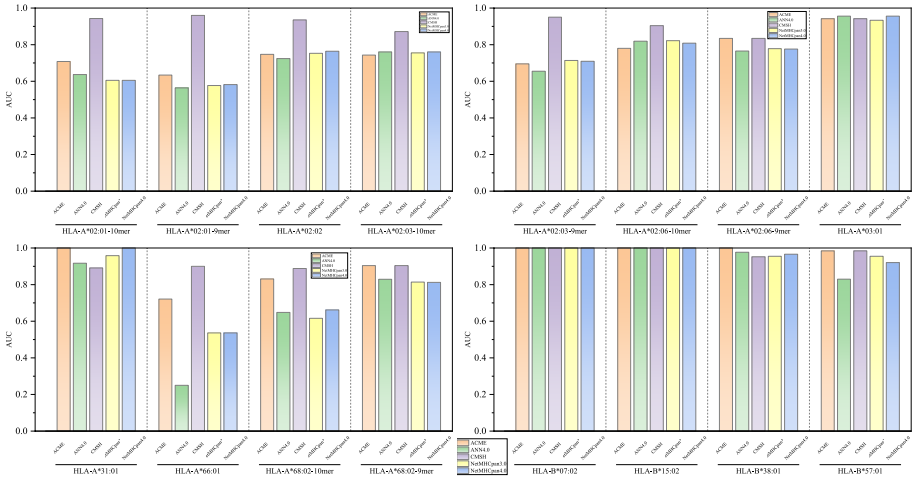

Fig. 6 Testing result for CMHS. AUC Performances for CMHS on IEDB comparing with SOTA Model. Purple line corresponding to the CMHS

Table 2 Performance comparison of CMHS and its variants under different ablation settings

Ablation	Performance	
	SRCC	ACC
CMHS	0.813	0.941
CMHS -without structure	0.7403	0.932
CMHS -without squence	0.4503	0.7658
CMHS -without LoRA	0.796	0.834
CMHS -without contrastive	0.665	0.873
CMHS -without noise layer	0.73	0.889

Ablation of CMHS

We introduce ablation experiment to shows the contribution of each module. We divided into two parts, including sequence-only: remove all the module of graph; and graph-only: remove the sequence module, including LLMs and CNN-extract and contrast learning. The result is shown in Table 2.

The ablation study results is shown in Table 2. The full CMHS model achieves best overall performance, with SRCC reaching 81.3% and AUC at 94.1%, demonstrating the effectiveness of the proposed design.

Removing the structural representation leads to SRCC degradation drops from 81.3% to74%, indicating that MHC structural modeling plays an important role in capturing spatial constraints. In contrast, removing the sequence-based representation results in a much more SRCC performance drop with 45%, shown that sequence information remains a fundamental determinant of binding affinity.

The exclusion of LoRA-based fine-tuning causes a moderate decline in performance SRCC 79.6%, ACC 83.4%), suggesting that parameter-efficient adaptation of the ESM improves feature expressiveness while avoiding overfitting. This result supports the use of LoRA as an effective lightweight fine-tuning strategy within CMHS.

When contrastive learning is removed, the SRCC decreases to 66.5%, indicating that contrastive objectives are essential for aligning multimodal representations and enhancing the discriminability between high- and low-affinity peptide-MHC pairs.

Finally, removing the noise-aware layer also leads to a clear performance drop with SRCC 73.0%, confirming that modeling structural uncertainty contributes to more robust affinity prediction, especially in the presence of noisy or predicted structures.

Overall, these ablation results demonstrate that each component of CMHS contributes positively to the final performance, and that the integration of sequence, structure, contrastive learning, LoRA-based adaptation, and noise modeling is critical for achieving optimal predictive accuracy.

Discussion

In this study, the CMHS provide a framework for MHC-peptide binding prediction by integrating hierarchical graph representation learning, biophysical noise modeling, and LLM-based sequence representations. CMHS decoupling intra-MHC conformational learning from peptide-MHC binding dynamics, together with effective multimodal feature fusion, substantially enhances prediction performance and addresses several limitations of existing approaches. Compared with pseudo sequence based methods, CMHS preserves the spatial topology critical for allosteric regulation, such as long-range communication between the α1 helix and β2-microglobulin in HLA-I molecules, whereas pseudo-sequences tend to fragment these structural pathways. In addition, the proposed B-factor-guided noise-aware layer explicitly models structural uncertainty, enabling the network to account for peptide-induced conformational flexibility. For graph neural network-based approaches, the two-stage hierarchical design mitigates the over-smoothing problem by first stabilizing global MHC structural coherence and subsequently modeling local binding-induced perturbations. Extensive external evaluations demonstrate that CMHS consistently outperforms state-of-the-art methods, including NetMHCpan, MHCflurry, and ACME, achieving superior accuracy and stability.

Our methods involves the introduction of hierarchical graph networks and multimodal feature inevitably increases computational complexity and training time compared with sequence-only methods, which may limit large-scale or real-time applications.

Author contributions

R. L. contributed to the work conception and model design, and also conducted the experiments. Y. W. contributed to analyze the experimental data and with the help of validation. H. L. and B. Z. contributed to data collection. Q. P. contributed to the supervision, funding acquisition, and writing review. All authors reviewed and approved the final manuscript.

Funding

This work is supported by the National Nature Science Foundation of China (61872288).

Data availability

Data was deposited into the IEDB and are available including the training MHC alleles datasets downloaded from https://www.iedb.org/database_export_v3.php, the weekly datasets downloaded from http://tools.iedb.org/auto_bench/mhci/weekly/. All externally submitted data are from NIH epitope contracts whose projects undergo ethical screening prior to data collection, especially when human or live specimens are in question. Therefore, there is no further assessment to be done by the IEDB.

Declarations

Consent to participate

Not applicable. The datasets in our research are from IEDB datasets, including the training MHC alleles datasets downloaded from https://www.iedb.org/database_export_v3.php, the weekly datasets downloaded from http://tools.iedb.org/auto_bench/mhci/weekly/. All externally submitted data are from NIH epitope contracts whose projects undergo ethical screening prior to data collection, especially when human or live specimens are in question. Therefore, there is no further assessment to be done by the IEDB.

Consent for publication

Not applicable.

Competing interest
The authors declare that they have no competing interest.

Received: 10 November 2025 / Accepted: 13 February 2026
Published online: 28 February 2026

References

1. Yewdell JW, Bennink JR. Immunodominance in major histocompatibility complex class I-restricted t lymphocyte responses. Annu Rev Immunol. 1999;17(1):51–88.
2. Huang H, Wang C, Rubelt F, Scriba TJ, Davis MM. Analyzing the mycobacterium tuberculosis immune response by T-cell receptor clustering with GLIPH2 and genome-wide antigen screening. Nat Biotechnol. 2020;38(10):1194–202.
3. Lundegaard C, Lund O, Keşmir C, Brunak S, Nielsen M. Modeling the adaptive immune system: predictions and simulations. Bioinformatics. 2007;23(24):3265–75.
4. Zhang W, Liu J, Niu Y. Quantitative prediction of MHC-II binding affinity using particle swarm optimization. Artif Intell Med. 2010;50(2):127–32.
5. Gulbakan B, Yasun E, Shukoor MI, Zhu Z, You M, Tan X, et al. A dual platform for selective analyte enrichment and ionization in mass spectrometry using aptamer-conjugated graphene oxide. J Am Chem Soc. 2010;132(49):17408–10.
6. Luo Y, Kanai M, Choi W, Li X, Sakaue S, Yamamoto K, et al. A high-resolution HLA reference panel capturing global population diversity enables multi-ancestry fine-mapping in HIV host response. Nat Genet. 2021;53(10):1504–16.
7. Nielsen M, Lundegaard C, Blicher T, Lamberth K, Harndahl M, Justesen S, et al. NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and-B locus protein of known sequence. PLoS ONE. 2007;2(8):796.
8. Nielsen M, Andreatta M. NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. Genome Med. 2016;8:1–9.
9. Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. NetMHCpan-4.0: improved peptide-MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. J Immunol. 2017;199(9):3360–8.
10. 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):449–54.
11. O'Donnell TJ, Rubinsteyn A, Bonsack M, Riemer AB, Laserson U, Hammerbacher J. MHCflurry: open-source class I MHC binding affinity prediction. Cell Syst. 2018;7(1):129–32.
12. Ma J, Wang Z, Tong C, Yang Q, Zhang L, Liu H. pMHChat, characterizing the interactions between major histocompatibility complex class II molecules and peptides with large language models and deep hypergraph learning. Brief Bioinform. 2025;26(4):321.
13. Hu Y, Wang Z, Hu H, Wan F, Chen L, Xiong Y, et al. Acme: pan-specific peptide-MHC class I binding prediction through attention-based deep neural networks. Bioinformatics. 2019;35(23):4946–54.
14. Phloyphisut P, Pornputtapong N, Sriswasdi S, Chuangsuwanich E. MHCSeqNet: a deep neural network model for universal MHC binding prediction. BMC Bioinform. 2019;20:1–10.
15. Sefer E. Anomaly detection via graph contrastive learning. In: Proceedings of the 2025 SIAM International Conference on Data Mining (SDM). Society for Industrial and Applied Mathematics. 2025;51–60.
16. Bi X, Ma W, Jiang H, Lu W, Nie J, Wei Z, et al. Protein interaction pattern recognition using heterogeneous semantics mining and hierarchical graph representation. Pattern Recogn. 2025;172:112563.
17. Bi X, Ma W, Jiang H, Lu W, Wei Z, Zhang S. SSPPI: cross-modality enhanced protein-protein interaction prediction from sequence and structure perspectives. IEEE Trans Neural Netw Learn Syst. 2025;37:22–36.
18. 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):785.
19. Zhang L, Song W, Zhu T, Liu Y, Chen W, Cao Y. ConvNeXt-MHC: improving MHC-peptide affinity prediction by structure-derived degenerate coding and the ConvNeXt model. Brief Bioinform. 2024;25(3):133.
20. McShan AC, Devlin CA, Morozov GI, Overall SA, Moschidi D, Akella N, et al. TAPBPR promotes antigen loading on MHC-I molecules using a peptide trap. Nat Commun. 2021;12(1):3174.
21. Trowsdale J, Knight JC. Major histocompatibility complex genomics and human disease. Annu Rev Genomics Hum Genet. 2013;14(1):301–23.
22. Soylu NN, Sefer E. DeepPTM: protein post-translational modification prediction from protein sequences by combining deep protein language model with vision transformers. Curr Bioinform. 2024;19(9):810–24.
23. Soylu NN, Sefer E. Bert2ome: Prediction of 2 -o-methylation modifications from RNA sequence by transformer architecture based on bert. IEEE/ACM Trans Comput Biol Bioinf. 2023;20(3):2177–89.
24. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science. 2023;379(6637):1123–30.
25. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang L, Chen W. Lora: low-rank adaptation of large language models. 2021. arXiv:2106.09685 [cs.CL].
26. McMaster B, Thorpe C, Ogg G, Deane CM, Koohy H. Can Alphafold's breakthrough in protein structure help decode the fundamental principles of adaptive cellular immunity? Nat Methods. 2024;21(5):766–76.
27. Schwab J, Kimanius D, Burt A, Dendooven T, Scheres SHW. DynaMight: estimating molecular motions with improved reconstruction from cryo-EM images. Nat Methods. 2024;21:1855–62.
28. Bahar I, Atilgan AR, Erman B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. Fold Des. 1997;2(3):173–81.
29. Dehghan A, Abbasi K, Razzaghi P, Banadkuki H, Gharaghani S. CCL-DTI: contributing the contrastive loss in drug-target interaction prediction. BMC Bioinform. 2024;25(48):1471–2105.
30. Andreatta M, Nielsen M. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. Bioinformatics. 2016;32(4):511–7.

31. O'Donnell TJ, Rubinsteyn A, Laserson U. MHCflurry 2.0: improved pan-allele prediction of MHC class I-presented peptides by incorporating antigen processing. Cell Syst. 2020;11(4):418–9.
32. Trolle T, Metushi IG, Greenbaum JA, Kim Y, Sidney J, Lund O, et al. Automated benchmarking of peptide-MHC class I binding predictions. Bioinformatics. 2015;31(13):2174–81.
33. Karosiene E, Lundegaard C, Lund O, Nielsen M. NetMHCcons: a consensus method for the major histocompatibility complex class I predictions. Immunogenetics. 2012;64(3):177–86.
34. Zhang H, Lund O, Nielsen M. The pickpocket method for predicting binding specificities for receptors based on receptor pocket similarities: application to MHC-peptide binding. Bioinformatics. 2009;25(10):1293–9.
35. Peters B, Tong W, Sidney J, Sette A, Weng Z. Examining the independent binding assumption for binding of peptide epitopes to MHC-I molecules. Bioinformatics. 2003;19(14):1765–72.
36. Kim Y, Sidney J, Pinilla C, Sette A, Peters B. Derivation of an amino acid similarity matrix for peptide: MHC binding and its application as a Bayesian prior. BMC Bioinform. 2009;10(1):394.

Publisher's Note