BMC Bioinformatics

RESEARCH

Open Access

TransBindpMHCI: a transformer-based model for pan-specific MHC-I peptide binding prediction

Hu Xu[1,2†], Yuanli Ni[1†], Zixuan Chai[1], Xuan Cui[1], Xia Lei[1], Limei Liu[1*], Juanjuan Shan[1*] and Cheng Qian[1*]

†Hu Xu and Yuanli Ni have contributed equally to this work.

*Correspondence:
Limei Liu
limeilliu@cqu.edu.cn
Juanjuan Shan
juanjuansh@gmail.com
Cheng Qian
cqian8634@gmail.com
[1]Center for Precision Medicine of Cancer, Chongqing Key Laboratory of Translational Research for Cancer Metastasis and Individualized Treatment, Chongqing University Cancer Hospital and Chongqing University School of Medicine, Chongqing 400030, China
[2]School of Computer Science, Shanghai Jiao Tong University, Shanghai, China

Abstract

Human leukocyte antigen (HLA) molecules play a pivotal role in antigen presentation. Tumor cells present neoantigens on the cell surface via HLA molecules, thereby activating cytotoxic T cells and eliciting immune responses. This process offers critical opportunities for cancer immunotherapy and tumor vaccine development. However, the identification of tumor neoantigens remains challenging due to limitations in data scale, prediction accuracy, and cross-species compatibility of existing methods. To address these challenges, we developed TransBindpMHCI, a transformer-based pan-specific major histocompatibility complex (MHC) peptide binding prediction model. By employing 1,404,492 mass spectrometry-screened MHC-presented peptides for modeling, the model directly captures the authentic processes of peptide generation and presentation. Its dual-tier transformer encoder architecture significantly enhances feature extraction capabilities for peptide-MHC binding patterns while reducing computational complexity. Furthermore, TransBindpMHCI extends prediction coverage to peptides spanning 8–15 amino acids and achieves cross-species compatibility for both human and murine MHC-I molecules. Comprehensive evaluations demonstrate that TransBindpMHCI outperforms existing methods in accuracy, computational efficiency, and generalizability, enabling the identification of more immunogenic neoantigens. This model holds substantial promise for advancing tumor neoantigen validation and personalized vaccine design.

Keywords MHC-I peptide binding prediction, Transformer, Neoantigen, Cancer immunotherapy, Mass spectrometry, Deep learning

Introduction

In 2014, Science reported a groundbreaking case in which a patient with advanced malignant cholangiocarcinoma was successfully treated using lymphocytes capable of specifically recognizing abnormal proteins resulting from cancer cell gene mutations [1]. Since then, tumor neoantigen-based immunotherapy has garnered significant attention for its remarkable therapeutic potential across various cancer types [2, 3]. Tumor neoantigens are specific short peptides generated by somatic mutations in tumor cells. They are exclusively expressed in tumor cells and absent in normal tissues. Through human

leukocyte antigen (HLA) molecules, tumor neoantigens are presented on the cell surface, activating T cells and inducing specific immune responses, ultimately leading to tumor cells destruction. Due to their potent immunogenicity, tumor neoantigens serve as ideal targets for immunotherapy [4].

However, only a small fraction of the numerous mutated peptides generated by tumor cells can effectively activate anti-tumor immune responses [5, 6]. Therefore, accurately predicting and identifying presented neoantigens is crucial for advancing cancer immunotherapy and vaccine development. Current tumor neoantigen prediction methods primarily rely on genomic sequencing data to identify nonsynonymous mutation sites and computational algorithms to predict potential neoantigen epitopes [7–9]. Additionally, effective prediction algorithms must consider multiple factors involved in antigen presentation, including proteasomal cleavage of mutated proteins, peptide transport, and binding affinity with major histocompatibility complex (MHC) molecules. The primary goal of neoantigen identification is to select peptides capable of eliciting CD8+ T cell responses, with MHC-I molecules playing a crucial role in their presentation [2]. Thus, this study focuses on improving MHC-I peptide binding prediction to enhance tumor neoantigen identification.

In recent years, deep learning models based on the attention mechanism has provided a novel approach to addressing long-range dependencies in sequence data. These methods have demonstrated remarkable feature extraction capabilities and have been successfully applied in various bioinformatics tasks [10, 11]. However, their application in tumor neoantigen identification still faces challenges, including limited HLA type coverage, restricted peptide length ranges, slow prediction speed, and insufficient generalizability to rare or unseen alleles [12, 13]. Additionally, mice are commonly used as experimental models in early-stage vaccine development to evaluate presented candidates and protective effects. Predicting murine MHC binding can facilitate the identification of candidate antigen epitopes capable of eliciting immune responses in mice, thereby optimizing vaccine design. Compared to blind antigen selection, this approach significantly reduces the range and cost of experimental validation, improving vaccine development efficiency. However, most existing methods do not support murine model predictions, limiting their applicability in preclinical research [14–16].

In parallel with these developments, several recent studies have proposed more advanced models for peptide–protein and peptide–MHC interactions. For example, the Sliding Window Interaction Grammar (SWING) model formulates peptide–protein binding as a generalized interaction language model and has demonstrated strong performance across diverse peptide–protein interaction tasks [17]. More recently, HLApollo introduced a high-quality peptide–MHC-I presentation model tailored to the design of improved cancer immunotherapy targets, highlighting the importance of accurate MHC-I presentation modeling for neoantigen prioritization [18]. These approaches further underscore the potential of deep learning and attention-based architectures for modeling complex peptide–protein interactions, but most existing methods still face limitations in cross-species coverage, support for variable peptide lengths, and computational efficiency in large-scale screening pipelines. To address these challenges, this study presents TransBindpMHCI, a Transformer-based pan-specific MHC-I peptide binding prediction model. By employing 1,404,492 mass spectrometry data entries for end-to-end modeling, TransBindpMHCI captures the authentic processes of tumor

neoantigen production and presentation. The model employs a dual-layer Transformer encoder design, integrating single- and multi-layer multi-head attention mechanisms to balance computational complexity and learning capacity. Furthermore, TransBindpMHCI expands its prediction scope to support peptides ranging from 8 to 15 amino acids in length and employs a mixed training strategy to achieve cross-species prediction for human and murine MHC-I molecules. Comprehensive evaluations demonstrate that TransBindpMHCI outperforms existing methods in terms of accuracy, computational efficiency, and generalizability. To validate its predictive capability, we applied the model to experimentally validated datasets from HPV vaccines and tumor neoantigens, where TransBindpMHCI exhibited high accuracy in neoantigen prediction, providing a valuable tool for immunotherapy target screening and vaccine design. Furthermore, we developed a free online platform to facilitate the application of this model in tumor immunotherapy and vaccine development.

Methods

TransBindpMHCI model construction

This study introduces TransBindpMHCI, a Transformer-based pan-specific MHC-I peptide binding prediction model (Fig. 1a), designed to capture the intricate binding mechanisms between peptides and MHC-I molecules. The Transformer architecture [19], with its self-attention mechanism, effectively addresses the challenge of long-range dependencies in sequence data, making it an ideal choice for peptide-MHC binding prediction. The core design of TransBindpMHCI consists of four key modules: the Embedding Module, Fusion Module, Encoding Module, and Mapping Module. The collaborative integration of these modules significantly enhances the model's predictive performance and applicability (Fig. 1b, Supplementary Information).

In the Embedding Module, the input peptide sequence $\boldsymbol{P} = (\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_n)$ and the MHC-I pseudo-sequence [20] $\boldsymbol{M} = (\boldsymbol{m}_1, \boldsymbol{m}_2, \ldots, \boldsymbol{m}_n)$ are embedded into high-dimensional vector spaces, resulting in embedding matrices $\boldsymbol{E_p} \in \boldsymbol{R}^{n \times d_{\mathrm{model}}}$ and $\boldsymbol{E}_M \in \boldsymbol{R}^{k \times d_{\mathrm{model}}}$. To retain positional information, the module incorporates Positional Encoding, utilizing sine and cosine functions to define relative positions within sequences, enhancing sequence feature expressiveness. Unlike existing methods, TransBindpMHCI significantly broadens its input range by supporting peptide lengths of 8–15 amino acids while accommodating both human and murine MHC-I molecules. This design improves the model flexibility and supports more complex antigen presentation scenarios.

Building on the features extracted by the Embedding Module, the Fusion Module integrates peptide sequence features and MHC-I pseudo-sequence features to capture potential interactions between the two. Specifically, the peptide features $\boldsymbol{E_P}$ and MHC-I features $\boldsymbol{E}_M$ are independently encoded using Encoder-1, then concatenated along the sequence dimension: $\boldsymbol{F}_{\mathbf{concat}} = [\boldsymbol{E_P}; \boldsymbol{E}_M]$. A linear mapping is subsequently applied to transform the concatenated features into a unified latent representation $\boldsymbol{F}_{\mathbf{fusion}}$. By retaining individual features and effectively modeling the complex interactions between peptides and MHC-I molecules, the Fusion Module addresses limitations of traditional methods that treat the two components independently.

The Encoding Module, based on the Transformer architecture, is designed with a two-tier encoder (Encoder-1 and Encoder-2) to separately extract local features and model

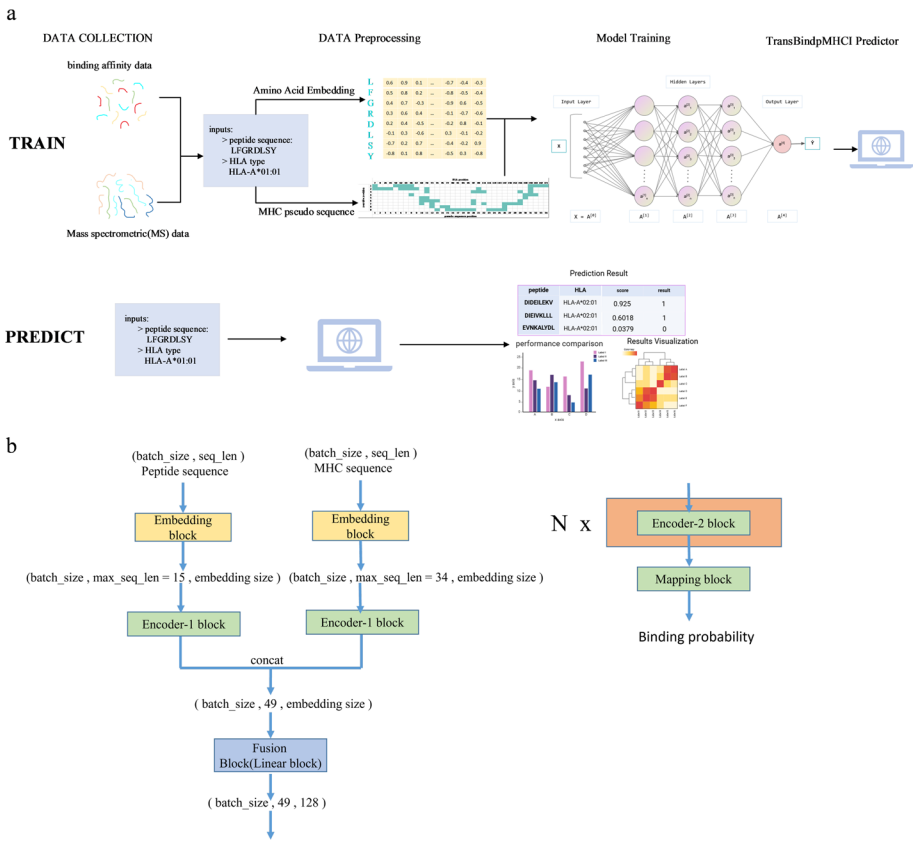

Fig. 1 The framework of TransBindpMHCI and architecture. **a** The training and prediction workflows of TransBindMHCI. The training process involves data collection (binding affinity and mass spectrometry data), data preprocessing (amino acid embedding of peptide sequences and MHC pseudo sequence), and model training using a neural network. The predictor outputs binding probabilities and enables performance comparison and visualization. The prediction process uses trained models to evaluate peptide-MHC binding probabilities. **b** The architecture of the TransBindMHCI model. Peptide and MHC sequences undergo embedding and are processed by encoder blocks. The outputs are concatenated and passed through fusion and subsequent encoder blocks, followed by a mapping block to produce the final binding probability

global interactions. The self-attention mechanism allows each position in the sequence to attend to all other position, capturing dependencies beyond simple forward or backward contexts. Multi-head self-attention (MHSA), an extension of self-attention, operates multiple independent attention mechanisms ("heads") within the same layer. Each head has its own parameter set, enabling the model to capture information from different perspectives, including local and global contexts, structural information, and semantic relationships. Encoder-1 employs a single-layer MHSA mechanism to learn independent peptides and MHC-I features. Residual connections and layer normalization stabilize the training process and reduce computational complexity. Encoder-2, in contrast, employs a multi-layer MHSA mechanism. Its stacked design captures higher-order interactions within the fused features, further enhancing the ability to model peptide-MHC binding patterns. Compared to the encoder proposed by Chu et al. [21], the multi-level encoder in TransBindpMHCI improves computational efficiency by approximately 1.2 times while retaining high feature extraction capability.

Finally, the Mapping Module uses a feed-forward network to map high-dimensional fused features $\boldsymbol{F}_{\mathbf{encoder}}$ extracted by the Encoding Module into a scalar, representing the binding probability between a peptide and an MHC-I molecule. A sigmoid activation

function limits the output value to the range [0,1], ensuring alignment with the numerical range of binding probabilities and improving the biological relevance and experimental applicability of the predictions.

The modular design of TransBindpMHCI not only significantly expands its input range but also demonstrates competitive performance in cross-species predictions, supports non-canonical peptide lengths, and enhances computational efficiency compared to existing methods.

Model training

To construct a high-quality dataset for model training, this study integrates data from multiple publicly available databases, including IEDB [22], SysteMHC [23, 24], EPIMHC [25], SYFPEITHI [26, 27], along with clinical trial data disclosed in relevant literature. The database includes both qualitative MHC-I peptide ligand data, obtained via mass spectrometry, and quantitative peptide-MHC binding affinity data. The final dataset was divided into six subsets for model training, selection, and performance evaluation: Training Dataset, GD (Generalization Dataset), IWBD (IEDB_weekly_benchmark Dataset), DMVD (Diverse MHC-Verified Dataset), HPV Dataset and Neoantigen Dataset (Fig. 2a, Supplementary Table S1). The Training Dataset contains 1,404,492 MHC-presented peptides from both humans and mice, covering 153 human MHC-I subtypes and 7 mouse subtypes, with 655,987 positive samples and 748,505 negative samples (Supplementary Figs. 1a, and 2a). Negative samples were composed of naturally occurring peptides in the database as well as peptides generated from source proteins within the immunopeptidome (Supplementary Information). To explore cross-species prediction capabilities, the Training Dataset was further divided into Training_Mouse and Training_Human subsets (Fig. 2b). Three models were trained based on these subsets: (1) TransBindpMHCI_Mouse (TM) trained on Training_Mouse, (2) TransBindpMHCI_Human (TH) trained on Training_Human, and (3) TransBindpMHCI_Total (TT) trained on the full Training Dataset.

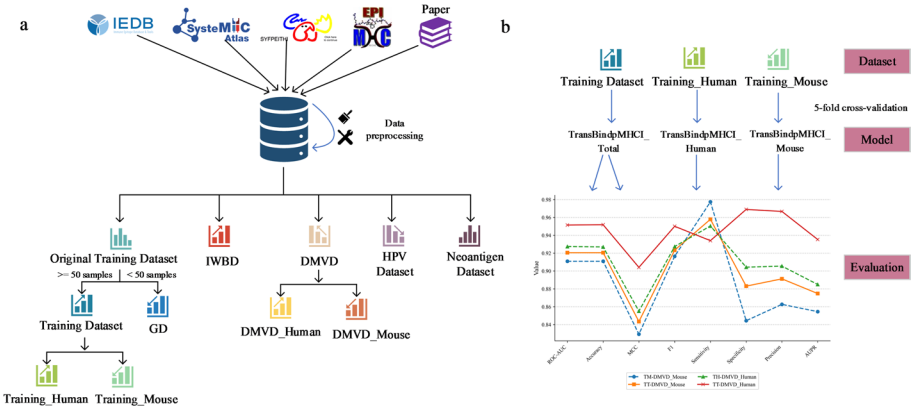

Fig. 2 Dataset curation, Model Evaluation, and Performance Comparison. **a** Overview of data sources and preprocessing steps. Data from IEDB, SysteMHC, EPIMHC, SYFPEITHI and clinical trial dataset published in papers are collected and curated into different datasets, including IWBD, DMVD (with human and mouse subset), HPV Dataset. The original training dataset are split into Training_Mouse based on species. **b** Multiple strategies were employed on the Training Dataset, resulting in the training of TT (trained on Training Dataset), TH (trained on Training_Human), and TM (trained on Trained_Mouse). This figure demonstrate that the TT, trained using mixed data, perform better on DMVD_Human and DMVD_Mouse compared to the independently speices-trained TH and TM

During model training, cross-entropy loss was employed as the optimization objective, with the Adam optimizer used to adjust model parameters, enhancing prediction accuracy and stability. A five-fold cross-validation strategy was implemented to evaluate model performance, where one subset was used for validation while the remaining subsets were used for training in each iteration. The average performance metrics across five iterations were calculated to assess model reliability and generalizability. To comprehensively evaluate model performance, eight performance metrics were employed: ROC-AUC, accuracy, F1 score, MCC, sensitivity, specificity, precision, and AUPR (Supplementary Information). For the HPV dataset, which consists exclusively of experimentally validated binders (positive-only data), standard classification metrics requiring negative samples (e.g., ROC-AUC, Specificity) are not applicable. Furthermore, comparing probability cut-offs directly to IC50 fold-changes can be misleading due to model-specific distribution variances. Therefore, we adopted a rank-based metric, Recall@K (or Top-K Recall), to evaluate the models. This metric measures the proportion of validated peptides successfully identified within the top K predictions when the dataset is ranked by predicted binding probability. This approach provides a fair comparison of the models' ability to prioritize true binders without relying on arbitrary threshold adjustments. To ensure a fair and rigorous comparison, we benchmarked TransBindpMHCI against leading pre-trained models. We adhered to the standard practice in bioinformatics, where new methods are evaluated against the best available versions of baseline methods on independent, non-overlapping test sets.

To verify the effectiveness of the cross-species training strategy, experiments were conducted on DMVD_Human and DMVD_Mouse, demonstrating that the TT model trained on mixed-species data outperformed the TM and TH models in both datasets (Fig. 2b). For instance, on DMVD_Human, the TT model achieved higher scores in 7 out of 8 metrics compared to the TH model, with a precision of 0.967 for TT versus 0.905 for TH. Similarly, on DMVD_Mouse, the TT model outperformed the TM model in 7 out of 8 metrics, achieving a precision of 0.891 compared to 0.863 for TM. These results confirm that the mixed-species training strategy effectively enhanced the model's predictive capability for murine data. For subsequent experiments, the TT model trained on mixed-species data were adopted and formally designated as TransBindpMHCI.

To rigorously evaluate the transfer learning capability of Trans-BindpMHCI and rule out the possibility of the model merely "memorizing" sequences shared between species, we conducted a comprehensive overlap analysis. We identified 7,822 peptides that were present in both the Human Training Set and the Mouse Training Set. To eliminate potential data leakage, we constructed a filtered human dataset by permanently removing these 7822 overlapping peptides. We then retrained the Trans-BindpMHCI model from scratch using this filtered dataset. This ensures that the model's performance on mouse alleles is derived from learned molecular features rather than memorized peptide sequences. The results demonstrate that the retrained model (trained on non-overlapping data) maintains a high performance level comparable to the original model. The performance drop was negligible, confirming that Trans-BindpMHCI relies on robust feature extraction and transfer learning mechanisms rather than data memorization. Detailed performance comparisons between the original and retrained models are provided in Supplementary Information Section 9. This result strongly supports the validity of our transfer learning strategy for low-resource species.

Results

Performance analysis

Generating to evaluate the effectiveness of TransBindpMHCI, we systematically compared its performance with five benchmark methods from IEDB (ANN [28, 29], NetMHCpan_BA [30, 31], SMM [32], SMMPMBEC [33]), the recommended IEDB method (NetMHCpan_EL [30, 31]), the widely used MHCflurry2 [34, 35] (including MHCflurry_BA and MHCflurry_AP), and four recently developed attention-based models (BigMHC [36], TransPHLA [21], DeepNetBim [37], and DeepAttentionPan [38]). These methods use different scoring standards to assess peptide-MHC (pHLA) binding, such as predicted score, predicted classification, predicted IC50 values, and percentile rank. In this study, predicted IC50 values and predicted scores were adopted as the evaluation criteria for regression and classification tasks, respectively.

Notably, not all methods are capable of predicting all data in the following comparison analyses. While models such as TransBindpMHCI, TransPHLA, NetMHCpan_BA, NetMHCpan_EL, MHCflurry_BA, and MHCflurry_AP support multiple MHC-I types and variable peptide lengths, other methods have limitations in terms of MHC-I types and peptide length range. For instance, DeepNetBim can only predict 9-mer peptides, while SMM and SMMPMBEC are limited to peptide lengths ranging from 8 to 11 amino acids. To ensure a fair and unbiased comparison across methods, we adhered to the Restricted Domain Comparison (RDC) principle, with detailed criteria outlined in the Supplementary Information section.

To comprehensively assess model performance, we used two datasets and eight performance metrics. To simulate real-world inconsistencies in data, we collected the IWBD dataset (Supplementary Fig. 1c, Supplementary Fig. 2e) from IEDB's weekly benchmark data, spanning from March 21, 2014 to October 19, 2023. This dataset exhibits significant differences in data distribution and features due to the varying release times of each batch. Results indicate that TransBindpMHCI significantly outperforms other methods in ROC-AUC, achieving a score of 0.708. While some methods excel in specific metrics, for instance, TransPHLA achieves a higher sensitivity (0.811 compared to 0.747 for TransBindpMHCI), its specificity is relatively poor (0.598 compared to 0.670 for TransBindpMHCI) (Fig. 3a, Supplementary Table S2). Overall, TransBindpMHCI demonstrates the best performance on this dataset. It is worth noting that all methods show relatively poor overall performance on the IWBD dataset, likely due to significant differences in data distribution.

To eliminate potential data contamination and further ensure a rigorous comparison, we additionally evaluated model performance using the DMVD_Human dataset, which includes 112 common MHC-I allele types spanning HLA-A, HLA-B, and HLA-C. This dataset is large-scale and covers a wide range of allele types, providing a robust benchmark for assessing model generalizability. Among the 11 compared methods, only six were capable of predicting the full DMVD_Human dataset, while the remaining five could only predict specific subsets. Results demonstrate that TransBindpMHCI achieves outstanding performance across multiple evaluation metrics, particularly in accuracy (0.952), ROC-AUC (0.952), and specificity (0.969) (Fig. 3b, Supplementary Table S2). These results highlight its ability to maintain high identification rates while minimizing false positives. Conversely, the NetMHCpan and MHCflurry series excel in specificity (NetMHCpan_BA: 0.991, NetMHCpan_EL: 0.996, MHCflurry_BA: 0.975,

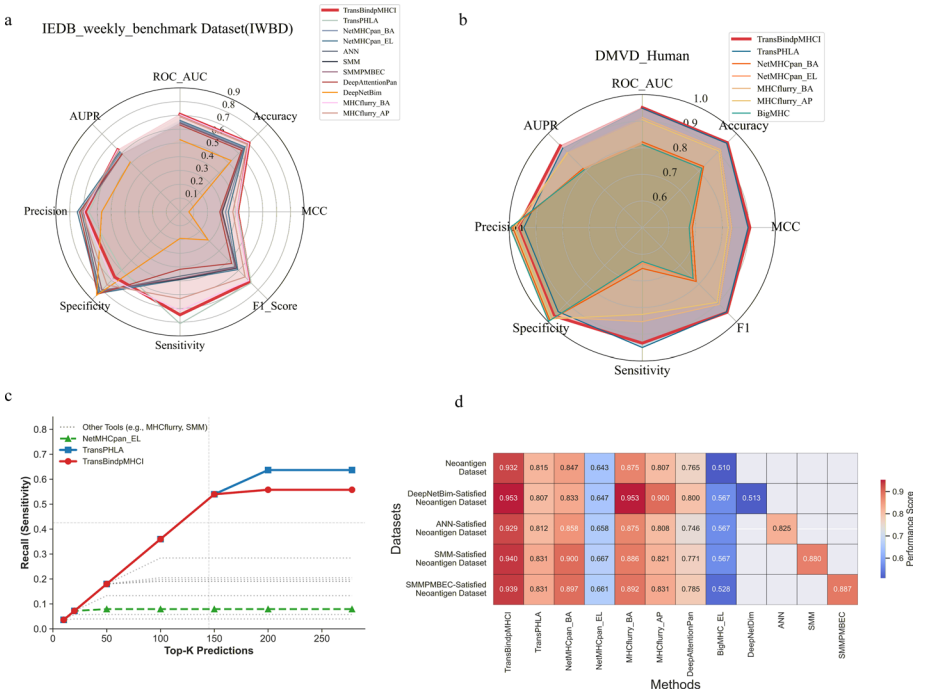

Fig. 3 Comprehensive Performance Evaluation of TransBindpMHCI on IWBD, DMVD, HPV and Neoantigen Dataset. **a** Each axis represent a different metric, including ROC_AUC, AUPR, Accuracy, Precision, MCC, Specificity, E4_Score, and Sensitivity. The different colored lined correspond to distinct prediction methods, such as TransBindpMHCI, NetMHCpan_BA, DeepAttentionPan and others. The data is sourced from the IWBD, with scores ranging from 0 to 1. The legend provides a detailed explanation of the color coding for each method. **b** This figure is the comparison result on the complete DMVD_Human. The other meanings are the same as those in (**a**). **c** Recall performance across Top-K predictions. The curves illustrate the sensitivity of TransBindpMHCI (red solid line) compared to state-of-the-art method, including TransPHLA (blue square), NetMHCpan_EL (green triangle), and other baseline tools (grey doted lines, e.g., MHCflurry, SMM). The x-axis represents the number of top-ranked prediction considered (K), while the y-axis shows the corresponding recall rate. TransBindpMHCI demonstrate a rapid increase in recall, effectively retrieving a majority of true positives within the top 150 predictions. **d** Heatmap illustrating the prediction accuracy of different methods on the Neoantigen dataset and its subset (because some methods cannot predict all data). The color intensity represents the accuracy, red represents high accuracy, and blue represent low accuracy. The rows represent the datasets used, and the columns represent different methods. The data comes from the Neoantigen Dataset. DeepNetBim-Satisfied Neoantigen Dataset means comparing TransBindMHCI with other methods on the subset of Neoantigen that can be predicted by DeepNetBim. Other subsets have similar meanings to the DeepNetBim-Satisfied Neoantigen Dataset

MHCflurry_AP: 0.985) but show reduced overall recognition rates (e.g., ROC-AUC scores: NetMHCpan_BA: 0.822, NetMHCpan_EL: 0.815, MHCflurry_BA: 0.914, MHCflurry_AP: 0.906). In further subset comparisons, TransBindpMHCI continues to exhibit significant advantages across multiple evaluation metrics (Supplementary Fig. 3a–e, Supplementary Table S2-DMVD_Human).

Clinical data evaluation

To assess the clinical applicability of TransBindpMHCI, we designed two evaluation tasks based on clinical datasets: HPV vaccine screening and tumor neoantigen identification. For the HPV vaccine screening task, we used the experimentally validated HPV Dataset published by Bonsack et al. [39] (Supplementary Fig. 1e, Supplementary Fig. 2f). This dataset consists of 278 experimentally confirmed pHLA complexes derived from HPV16 proteins E6 and E7 [40], with peptide lengths ranging from 8 to 11 amino acids.

Notably, the binding threshold for this dataset is defined as IC50 < 100 μm, which is 200 times more lenient than the typical binding threshold of 500 nm.

The HPV Dataset represents a challenging scenario containing viral neoantigens with varying binding affinities (IC50 < 100 μM). Since this dataset contains only positive samples, we evaluated the models using the Recall@K metric to assess their ability to retrieve validated epitopes. As shown in Fig. 3c, Transformer-based models demonstrated a significant advantage over traditional methods in identifying these viral peptides. Specifically, TransPHLA and TransBindpMHCI achieved the highest recall rates, identifying 63.7% and 55.8% of the validated peptides within the total ranked list (Top-278), respectively. In stark contrast, the widely used NetMHCpan_EL and NetMHCpan_BA identified only 7.9% and 19.1% of the epitopes, respectively.

This result highlights a critical limitation in current standard tools. They are often overly conservative and may filter out potential immunogenic neoantigens that exhibit weaker binding signals. TransBindpMHCI, by leveraging its attention mechanism, effectively captures these subtle binding patterns, offering a recall rate comparable to the state-of-the-art TransPHLA and significantly outperforming conventional tools. This capability is particularly vital for vaccine development, where missing a potential epitope (False Negative) is often more detrimental than including a non-binder. In our supplemental information, we further explored the performance variations of our method under different thresholds for HPV.

For the tumor neoantigen identification task, we used the Neoantigen Dataset collected by Chu et al. [21] (Supplementary Fig. 1f, Supplementary Fig. 2d), which includes 249 experimentally validated pHLA complexes from patients with non-small cell lung cancer, melanoma, ovarian cancer, and pancreatic cancer. Among the 11 compared methods, 7 were able to predict the full Neoantigen Dataset, while the remaining 4 could only predict subsets. Experimental results demonstrate that TransBindpMHCI exhibits significant advantages on this dataset, achieving an overall accuracy of 0.932, outperforming TransPHLA (0.815), NetMHCpan_BA (0.847), and MHCflurry_BA (0.876). In subset comparisons, TransBindpMHCI consistently maintained high accuracy (≥ 0.929). Only on the DeepNetBim subset did MHCflurry_BA achieve comparable performance to TransBindpMHCI. However, across all other subsets, TransBindpMHCI exhibited clear performance advantages (Fig. 3d). These results confirm that TransBindpMHCI reliably identifies presented, tumor-specific peptides across complete and subset datasets, demonstrating stable and superior performance compared to other methods.

Although TransBindpMHCI performs exceptionally well on the above two clinical datasets, further validation on larger-scale clinical datasets is necessary to ensure its reliability. Future research should focus on expanding dataset size and diversity, including data from various cancer types, infectious diseases, and diverse populations, to further enhance the model's generalizability and clinical applicability.

Other advantages of the model

To evaluate the practical accessibility of our tool, efficiency benchmarks were conducted on a standard consumer-grade device equipped with an Intel Core i5-10300H CPU and 16 GB RAM, without utilizing GPU acceleration. We adopted Throughput (pHLA/s) as the standardized metric to measure inference speed, defined as the number of peptide-HLA pairs processed per second. We focused our reporting on Batch Mode

performance, as this represents the standard workflow for high-throughput neoantigen screening tasks where millions of candidates are evaluated simultaneously. Single-peptide inference speed was excluded from the primary comparison as it is typically dominated by system overhead (e.g., Python interpretation) rather than model latency.

We compared the inference efficiency of TransBindpMHCI against state-of-the-art methods on a dataset of 164,702 sequences (DMVD_Human). As shown in Fig. 4a, TransBindpMHCI demonstrated superior performance even on a standard CPU. Our model achieved a throughput of 935.8 pHLA/s, completing the screening of the entire dataset in just 176 s (under 3 min). This represents a 1.2 × speedup over TransPHLA (776.9 pHLA/s) and a significant improvement over MHCflurry (115.2 pHLA/s) and DeepAttentionPan (32.7 pHLA/s).

Notably, TransBindpMHCI is extremely lightweight with a model size of only 19.3 MB. This compact architecture ensures a minimal memory footprint during inference, well within the 16 GB capacity of standard laptops, thus avoiding Out-of-Memory (OOM) issues during large-scale batch processing. These results demonstrate that TransBindpMHC is highly optimized and does not require specialized GPU hardware to achieve high-throughput performance, offering a significant advantage for clinical or experimental laboratories lacking high-end computational resources.

To further verify the model generalizability, we conducted experiments on the GD dataset (Supplementary Fig. 1b, Supplementary Fig. 2c), which consists of low-binding probability samples that were excluded from the training set. This dataset covers 53 MHC-I types (Supplementary Fig. 5), with each type containing fewer than 50 peptide samples, making it particularly challenging for prediction. Using eight evaluation metrics, the results indicate that TransBindpMHCI achieves an ROC-AUC score of 0.905, showcasing its strong ability to classify positive and negative samples (Fig. 4b). The model also achieved an accuracy of 0.906, demonstrating high prediction accuracy, and an F1 score of 0.899, indicating a balanced trade-off between precision and recall. These results highlight the robustness and reliability of TransBindpMHCI in predicting peptide-MHC binding for less common MHC-I alleles.

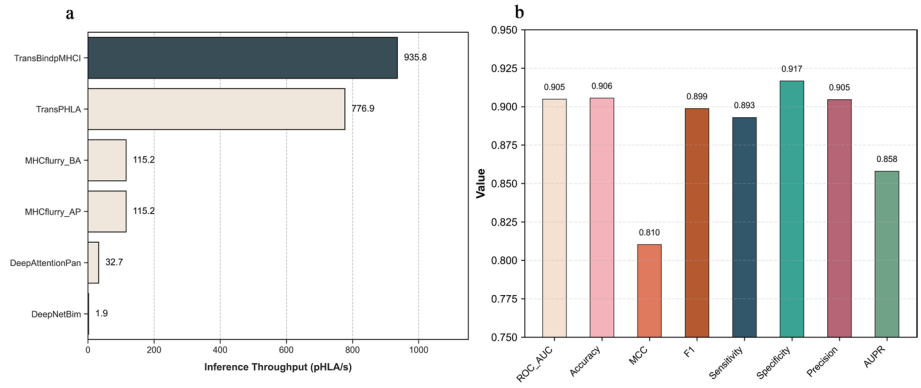

Fig. 4 Comparison of computational efficiency and performance evaluation of TransBindpMHCI on GD. **a** Efficiency comparison in terms of Throughput (pHLA/s). The bar chart displays the processing speed of TransBindpMHC compared to other tools on a standard CPU (Intel i5). TransBindpMHC achieves the highest throughput (~ 936 pHLA/s), enabling rapid screening of large-scale datasets. **b** Performance evaluation metrics of TransBindpMHCI on GD, demonstrating high scores in ROC-AUC, accuracy, sensitivity, precision, and other evaluation criteria. The result show that TransBindpMHCI is also robust and reliable in predicting the binding of some uncommon MCH-I subtypes

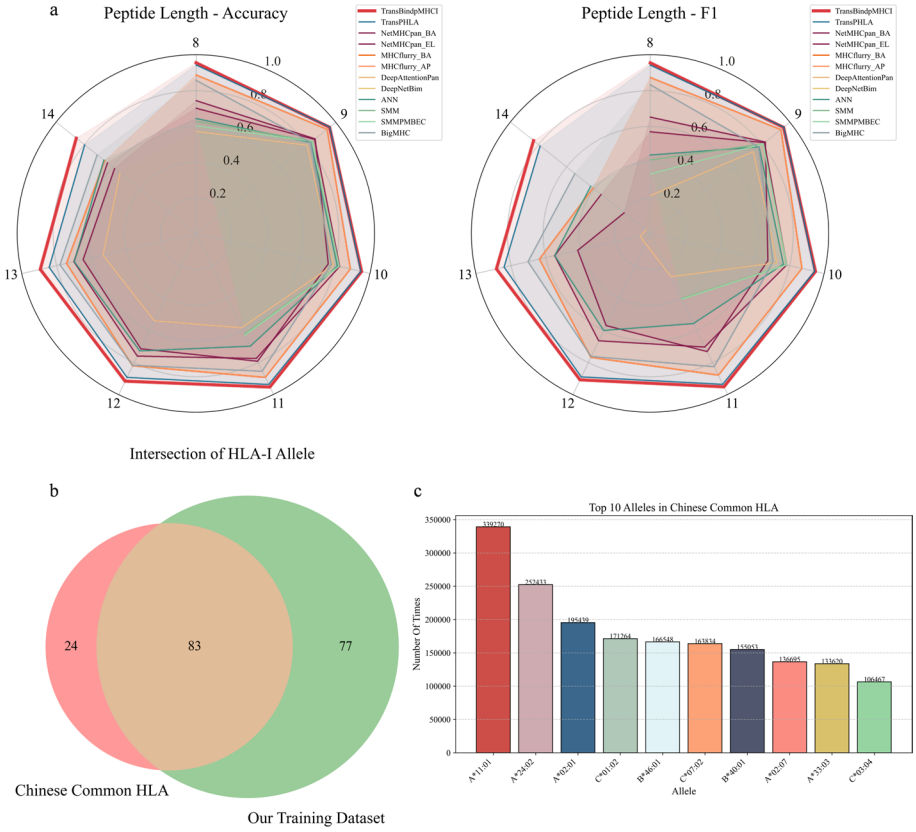

Fig. 5 Evaluation of the model's preference for peptide length and MHC-I type. **a** Radar chart comparing Accuracy and F1 scores of TransBindpMHCI and other baseline methods in different peptide lengths (8–14 amino acids) on DMVD_Human. TransBindpMHCI consistently outperforms baseline methods, with higher scores across all peptide lengths, demonstrating its robustness in handling varying peptide lengths. **b** Venn diagram illustrating the overlap between HLA types in the Training Dataset and those commonly found in the Chinese population. Of the total HLA types, 83 are shared, while 77 are exclusive to the training dataset and 24 are unique to the Chinese population, indicating adequate but improvable coverage of population-specific HLA alleles. **c** Bar chart showing the top 10 most common HLA alleles in the Chinese population based on occurrence frequency. Allele A*11:01 is the most prevalent, followed by A*24:02 and A*02:01, highlighting the importance of prioritizing these alleles in population-specific binding predictions

To further optimize the model for different MHC-I molecules and peptide lengths, which are critical for neoantigen recognition, we systematically compared the prediction performance of 11 different methods across peptide lengths and MHC-I types (Supplementary Table S3). Given that the DMVD dataset (Supplementary Fig. 1d, Supplementary Fig. 2b) includes diverse MHC-I types and was excluded from training, we used this dataset for further evaluations. However, since other methods do not support murine predictions, only DMVD_Human was used for these assessments.

Prediction Preference by Peptide Length. The model's performance across different peptide lengths reveals several key characteristics (Fig. 5a, Supplementary Table S4). First, TransBindpMHCI demonstrates remarkable stability and prediction performance across short to medium peptide lengths, ranging from 8 to 14 amino acids. Specifically, for short peptides (8-mer and 9-mer), the model achieves accuracy scores of 0.957 and 0.956, ROC-AUC scores of 0.957 and 0.956, AUPR scores of 0.938 and 0.942, and precision and specificity values exceeding 0.960. Second, prediction performance declines for longer peptides. When peptide lengths exceed 12 amino acids, prediction performance decreases significantly. For instance, the ROC-AUC and AUPR scores for 14-mer

peptides drop to 0.852 and 0.802, respectively. Third, comparison with other methods shows that most methods perform well for 9-mer peptides. TransBindpMHCI achieves an AUPR score of 0.941, significantly outperforming NetMHCpan_BA (0.834) and MHCflurry_BA (0.920) while slightly surpassing TransPHLA (0.936). Fourth, the range of supported peptide lengths varies across methods. Methods such as TransPHLA, NetMHCpan_BA, NetMHCpan_EL, MHCflurry_BA, and MHCflurry_AP perform well for peptide lengths of 8–11 amino acids, with TransPHLA approaching TransBindpMHCI's performance on certain metrics. Fifth, TransBindpMHCI demonstrates greater generalizability for long peptides, whereas most methods exhibit significant instability for peptides longer than 11 amino acids. Despite the inherent challenges in predicting longer peptide-MHC interactions, TransBindpMHCI maintains relatively stable performance. Finally, other methods have strict length limitations. DeepNetBim supports only 9-mer peptide predictions, while SMM and SMMPMBEC are limited to lengths of 8–11 amino acids. Although these methods exhibit competent performance within their supported ranges, they are generally outperformed by TransBindpMHCI.

Prediction Performance by MHC-I Type. We next evaluated the prediction performance of TransBindpMHCI across different MHC-I types in antigen recognition tasks and investigated potential influencing factors. Current prediction methods are often trained on datasets derived from European populations, leading to uncertainty in peptide-MHC binding predictions for non-European populations, such as Chinese populations. To address this, we further assessed the prediction capability of TransBindpMHCI for common HLA molecules in the Chinese population. A dataset compiled by Zhang [41] contains 107 common HLA-I allele types in the Chinese population. Of these, 83 alleles overlap with the 160 MHC-I types in our training dataset, achieving a coverage of 77.57% of Chinese HLA types (Fig. 5b).

We compared the performance of multiple methods in predicting peptide binding for the 10 most common HLA alleles in the Chinese population (Fig. 5c, Supplementary Table S5). The results indicate that for HLA-A*11:01, HLA-A*24:02, HLA-A*02:01, and HLA-C*03:04, TransBindpMHCI consistently ranks among the top three methods and outperforms other methods in most evaluation metrics. Specifically, for HLA-A11:01 and HLA-C03:04, TransBindpMHCI achieves ROC-AUC scores of 0.957 and 0.964, AUPR scores of 0.938 and 0.937, and strong performance in accuracy, F1 score, and sensitivity. For HLA-A24:02 and HLA-A02:01, while precision is slightly lower than that of NetMHCpan_BA and MHCflurry_BA, TransBindpMHCI shows significant advantages in comprehensive metrics, such as F1 score and AUPR. In contrast, NetMHCpan_EL and DeepAttentionPan perform poorly across all HLA allele types (Supplementary Fig. S6).

Overall, TransBindpMHCI demonstrates robust prediction accuracy across most MHC-I types, with ROC-AUC scores ranging from 0.900 to 0.990 and precision, MCC, and F1 scores exceeding 0.900 in most cases. For specific MHC-I types such as HLA-A02:04 and HLA-A02:05, TransBindpMHCI achieves near-perfect ROC-AUC scores of 1.000, indicating the presence of well-defined peptide binding patterns. However, for MHC-I types with more complex binding patterns (e.g., HLA-A30:01 and HLA-A30:02), ROC-AUC scores are slightly lower, suggesting challenges in prediction accuracy.

In summary, TransBindpMHCI consistently delivers high performance across different MHC-I types, with minor variations between specific alleles. Its overall reliability

and strong generalizability highlight its utility in peptide-MHC binding predictions, making it a valuable tool for immunotherapy and vaccine development.

Discussion

The TransBindpMHCI model presented in this study demonstrates exceptional performance and broad applicability in the field of MHC-I peptide binding prediction. By integrating large-scale mass spectrometry data and employing an innovative model architecture and training strategy, it effectively addresses key challenges in this field, including limited data diversity, insufficient support for variable peptide lengths, and inadequate cross-species generalization.

Compared to existing methods, the superiority of TransBindpMHCI stems from several key innovations. First, at the data and training level, we constructed a comprehensive dataset of over 1.4 million mass spectrometry entries, which more authentically reflects the processes of antigen processing and presentation. More importantly, we pioneered a mixed-species training strategy using both human and murine data, enabling the model's direct application in preclinical mouse models—a capability lacking in many existing tools. Second, in terms of model architecture, the unique dual-tier Transformer encoder not only extracts local and global features efficiently but also significantly enhances computational speed. This architectural advantage allows it to effectively handle variable-length peptides (8–15 amino acids), addressing a major limitation of traditional models like the CNN-based MHCflurry and proving crucial for discovering neoantigens derived from gene fusions or alternative splicing. Furthermore, the model exhibits strong generalization to rare alleles and non-European populations, such as common HLA alleles in the Chinese population, thereby addressing the population bias common in existing models.

At the clinical level, the model's successful validation on HPV vaccine and tumor neoantigen datasets highlights its practical utility. It not only accurately identifies experimentally validated pHLA complexes but also demonstrates immense potential for personalized cancer vaccine design and immunotherapy. By bridging the gap between preclinical (murine) and clinical (human) research, TransBindpMHCI is poised to significantly accelerate the screening of vaccine candidates and reduce development costs. The freely accessible online platform we provide further promotes the adoption and application of this tool within the research community.

Furthermore, the analysis of the HPV Dataset revealed that traditional probability thresholds (e.g., 0.5) might be too stringent for certain clinical applications, such as identifying viral neoantigens with lower binding affinities. Our results showed that while standard tools like NetMHCpan missed the majority of these "weak binders", Transformer-based architectures (TransBindpMHCI and TransPHLA) successfully recovered a substantial portion of them. This suggests that TransBindpMHCI is well-suited for screening tasks where sensitivity is paramount, allowing researchers to explore a broader landscape of potential vaccine candidates.

Despite these significant advancements, we acknowledge several limitations and future directions. The datasets used for clinical validation remain relatively small, and the model's reliability must be further tested on larger, more heterogeneous clinical datasets covering a wider range of cancer types, diseases, and populations. Future research could focus on three areas: first, integrating multimodal data, such as genomic

and protein structural information, to enhance predictive accuracy; second, exploring optimized fine-tuning strategies; and third, constructing a large-scale pre-trained model tailored specifically for neoantigen prediction to achieve greater adaptability and predictive capacity in broader applications.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-026-06423-1.

Supplementary Material 1

Supplementary Material 2

Supplementary Material 3

Supplementary Material 4

Supplementary Material 5

Supplementary Material 6

Supplementary Material 7

Supplementary Material 8

Supplementary Material 9

Supplementary Material 10

Supplementary Material 11

Supplementary Material 12

Acknowledgements

We thank all the members of Center for Precision Medicine of Chongqing University Cancer Hospital for their kind technique support.

Author contributions

Hu Xu: conceptualized the study, performed experiments, collected data, developed the core al-gorithms, and drafted the original manuscript; Yuanli Ni: Conducted data analysis and visualization, performed literature review, supervised ex-perimental quality control, and revised the manuscript; Zixuan Chai: Collected and organized literature for the review section, summarizing both domes-tic and international research progress to provide a solid theoretical foundation for the study; Xuan Cui: Assisted in data analysis and figure production, contributing to the enhancement and re-finement of the manuscript content; Xia Lei: Contributed to data analysis and figure production, supplementing and improving the over-all presentation of results; Limei Liu: Oversaw experimental quality control, secured research resources, coordinated inter-disciplinary collaboration; Juanjuan Shan: Project Administration, data visualization, revised the manuscript, provided meth-odological guidance, and resolved academic discrepancies; Cheng Qian: Designed the research framework, provided methodological guidance, coordinated in-terdisciplinary collaboration, funding Acquisition, and managed manuscript submission and peer-review communication.

Funding

This work was supported by National Key R&D Program of China (2022YFC3401600), National Natural Science Foundation of China (82120108019), and the Fundamental Research Funds for the Central Universities (2023CDJKYJH001).

Data availability

Codes and data are available at https://github.com/xuhu0115/TransBindpMHC-V1/. All publicly available datasets used in this study are listed in Additional file 1: Supplementary Table S1, which includes the full names of the data repositories, permanent accession numbers or direct URLs, and the corresponding references. To facilitate usability, we have also provided a simplified system deployed on the Streamlit Cloud plat-form (https://xuhu-transbindpmhc.streamlit.app) for demonstration purposes, although it is unlikely to support heavy traffic.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 27 October 2025 / Accepted: 9 March 2026

Published online: 20 April 2026

References

1. Tran E, Turcotte S, Gros A, et al. Cancer immunotherapy based on mutation-specific CD4+ T cells in a patient with epithelial cancer. Science. 2014;344(6184):641–5.
2. Blass E, Ott PA. Advances in the development of personalized neoantigen-based therapeutic cancer vaccines. Nat Rev Clin Oncol. 2021;18(4):215–29. https://doi.org/10.1038/s41571-020-00460-2.
3. De Mattos-Arruda L, Vazquez M, Finotello F, et al. Neoantigen prediction and computational perspectives towards clinical benefit: recommendations from the ESMO precision medicine working group. Ann Oncol. 2020;31(8):978–90.
4. Xie N, Shen G, Gao W, et al. Neoantigens: promising targets for cancer therapy. Signal Transduct Target Ther. 2023;8(1):9. https://doi.org/10.1038/s41392-022-01270-x.
5. Tran E, Robbins PF, Rosenberg SA. Final common pathway' of human cancer immunotherapy: targeting random somatic mutations. Nat Immunol. 2017;18(3):255–62.
6. Tan X, Xu L, Jian X, et al. PGNneo: a proteogenomics-based neoantigen prediction pipeline in noncoding regions. Cells. 2023;12(5):782.
7. Gopanenko AV, Kosobokova EN, Kosorukov VS. Main strategies for the identification of neo-antigens. Cancers Basel. 2020;12(10):2879.
8. Hundal J, Kiwala S, McMichael J, et al. pVACtools: a computational toolkit to identify and visualize cancer neoantigens. Cancer Immunol Res. 2020;8(3):409–20.
9. Kim S, Kim HS, Kim E, et al. Neopepsee: accurate genome-level prediction of neoantigens by harnessing sequence and amino acid immunogenicity information. Ann Oncol. 2018;29(4):1030–6.
10. Yang W, Wang P, Xu S, et al. Deciphering cell–cell communication at single-cell resolution for spa-tial transcriptomics with subgraph-based graph attention network. Nat Commun. 2024;15(1):7101.
11. Zhao Q, Zhao H, Zheng K, et al. HyperAttentionDTI: improving drug–protein interaction predic-tion by sequence-based deep learning with attention mechanism. Bioinformatics. 2022;38(3):655–62.
12. Feng M, Liu L, Su K, et al. 3D genome contributes to MHC-II neoantigen prediction. BMC Genomics. 2024;25(Suppl 2):889.
13. Luo R, Chyr J, Wen J, et al. A novel integrated approach to predicting cancer immunotherapy effi-cacy. Oncogene. 2023;42(23):1913–25. https://doi.org/10.1038/s41388-023-02670-1.
14. Jung K, Son MJ, Lee SY, et al. Antibody-mediated delivery of a viral MHC-I epitope into the cytosol of target tumor cells repurposes virus-specific CD8+ T cells for cancer immunotherapy. Mol Cancer. 2022;21(1):102. https://doi.org/10.1186/s12943-022-01574-0.
15. Versteeg L, Adhikari R, Robinson G, et al. Immunopeptidomic MHC-I profiling and immunogenicity testing identifies Tcj2 as a new Chagas disease mRNA vaccine candidate. PLoS Pathog. 2024;20(12):e1012764.
16. Liu Y, Sun B, Wang J, et al. In silico analyses and experimental validation of the MHC class-I restricted epitopes of Ebolavirus GP. Int Immunol. 2022;34(6):313–25.
17. Siwek JC, Omelchenko AA, Chhibbar P, et al. Sliding Window Interaction Grammar (SWING): a generalized interaction language model for peptide and protein interactions. Nat Methods. 2025;22(8):1707–19.
18. Thrift WJ, Lounsbury NW, Broadwell Q, et al. Towards designing improved cancer immunotherapy targets with a peptide-MHC-I presentation model, HLApollo. Nat Commun. 2024;15(1):10752.
19. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Advances in neural information processing systems. 2017;30.
20. Nielsen M, Lundegaard C, Blicher T, et al. NetMHCpan, a method for quanti-tative predictions of peptide binding to any HLA-A and-B locus protein of known sequence. PLoS ONE. 2007;2(8):e796.
21. Chu Y, Zhang Y, Wang Q, et al. A transformer-based model to predict peptide–HLA class I bind-ing and optimize mutated peptides for vaccine design. Nat Mach Intell. 2022;4(3):300–11.
22. Vita R, Overton JA, Greenbaum JA, et al. The immune epitope database (IEDB) 3.0. Nucleic Acids Res. 2015;43(D1):D405–12.
23. Huang X, Gan Z, Cui H, et al. The SysteMHC Atlas v2. 0, an updated resource for mass spectrom-etry-based immunopeptidomics. Nucleic Acids Res. 2024;52(D1):D1062–71.
24. Shao W, Pedrioli PGA, Wolski W, et al. The SysteMHC atlas project. Nucleic Acids Res. 2018;46(D1):D1237–47.
25. Reche PA, Zhang H, Glutting JP, et al. EPIMHC: a curated database of MHC-binding peptides for customized computational vaccinology. Bioinformatics. 2005;21(9):2140–1.
26. Rammensee HG, Bachmann J, Emmerich NPN, et al. SYFPEITHI: database for MHC ligands and peptide motifs. Immunogenetics. 1999;50:213–9.
27. Schuler MM, Nastke MD, Stevanović S. SYFPEITHI: database for searching and T-cell epitope prediction. In: Immunoinformatics: predicting immunogenicity in silico. 2007. p. 75–93.
28. Andreatta M, Nielsen M. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. Bioinformatics. 2016;32(4):511–7.
29. Nielsen M, Lundegaard C, Worning P, et al. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. Protein Sci. 2003;12(5):1007–17.
30. Reynisson B, Alvarez B, Paul S, et al. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predic-tions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. Nucleic Acids Res. 2020;48(W1):W449–54.
31. Lundegaard C, Lamberth K, Harndahl M, et al. NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8–11. Nucleic Acids Res. 2008;36(suppl_2):W509–12.
32. Peters B, Sette A. Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method. BMC Bioinformatics. 2005;6(1):132.
33. Kim Y, Sidney J, Pinilla C, et al. Derivation of an amino acid similarity matrix for peptide: MHC binding and its application as a Bayesian prior. BMC Bioinformatics. 2009;10:1–11.
34. O'Donnell T, Rubinsteyn A. High-throughput MHC I ligand prediction using MHCflurry. In: Bioinformatics for cancer immunotherapy: methods and protocols. 2020. p. 113–27.
35. 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(1):42–8.
36. Albert BA, Yang Y, Shao XM, et al. Deep neural networks predict class I major histocompatibility complex epitope presentation and transfer learn neoepitope immunogenicity. Nat mach intell. 2023;5(8):861–72.

37. Yang X, Zhao L, Wei F, et al. DeepNetBim: deep learning model for predict-ing HLA-epitope inter-actions based on network analysis by harnessing bind-ing and immunogenicity information. BMC Bioinform. 2021;22(1):231.
38. Jin J, Liu Z, Nasiri A, et al. Deep learning pan-specific model for interpretable MHC-I peptide binding prediction with improved attention mechanism. Proteins Struct Funct Bioinform. 2021;89(7):866–83.
39. O'Donnell TJ, Rubinsteyn A, Bonsack M, et al. MHCflurry: open-source class I MHC binding affinity prediction. Cell Syst. 2018;7(1):129-132. e4.
40. Bonsack M, Hoppe S, Winter J, et al. Performance evaluation of MHC Class-I binding prediction tools based on an experimentally validated MHC–peptide binding data set. Cancer Immunol Res. 2019;7(5):719–36.
41. He Y, Li J, Mao W, et al. HLA common and well-documented alleles in China. HLA. 2018;92(4):199–205.

Publisher's Note