*Briefings in Bioinformatics*, 2025, **26(1)**, bbae595
https://doi.org/10.1093/bib/bbae595
**Problem Solving Protocol**

# CapHLA: a comprehensive tool to predict peptide presentation and binding to HLA class I and class II

Yunjian Chang and Ligang Wu*

Key Laboratory of RNA Science and Engineering, Shanghai Institute of Biochemistry and Cell Biology, Center for Excellence in Molecular Cell Science, Chinese Academy of Sciences, University of Chinese Academy of Sciences, 320 Yueyang Road, Shanghai 200031, China

*Corresponding author. Ligang Wu, Key Laboratory of RNA Science and Engineering, Shanghai Institute of Biochemistry and Cell Biology, Center for Excellence in Molecular Cell Science, Chinese Academy of Sciences, University of Chinese Academy of Sciences, 320 Yueyang Road, Shanghai 200031, China.
E-mail: lgwu@sibcb.ac.cn

## Abstract

Human leukocyte antigen class I (HLA-I) and class II (HLA-II) proteins play an essential role in epitope binding and presentation to initiate an immune response. Accurate prediction of peptide-HLA (pHLA) binding and presentation is critical for developing effective immunotherapies. However, current tools can predict antigens exclusively for pHLA-I or pHLA-II, but not both; have constraints on peptide length; and commonly show unsatisfactory predictive accuracy. Here, we developed a convolution and attention-based model, CapHLA, trained with eluted ligand and binding affinity mass spectrometry data, to predict peptide presentation probability (PB) and binding affinities (BA) for HLA-I and HLA-II. In comparison with 11 other methods, CapHLA consistently showed improved performance in predicting pHLA BA and PB, particularly in HLA-II and non-classical peptide length datasets. Using CapHLA PB and BA predictions in combination with antigen expression level (EP) from transcriptomic data, we developed a neoantigen quality model for predicting immunotherapy response. In analyses of clinical response among 276 cancer patients given immunotherapy and overall survival in 7228 cancer patients, our neoantigen quality model outperformed other genetics-based models in predicting response to checkpoint inhibitors and patient prognosis. This study provides a versatile neoantigen screening tool, illustrating the prognostic value of neoantigen quality.

**Keywords**: neoantigen presentation; binding affinity; immunotherapy; biomarker; machine learning

## Introduction

Peptide fragments produced from cancer-associated gene variants can serve as tumor neoantigens that bind with human leukocyte antigen (HLA) molecules to form peptide-HLA (pHLA) complexes which are eventually presented on the cell surface [1]. This process is a crucial prerequisite for proficient T-cell recognition that ultimately results in initiating a potent immune response [2]. HLAs are generally categorized as either HLA class I (HLA-I) or HLA class II (HLA-II). HLA-I binds peptides 8–15 amino acids in length derived from intracellular proteins following proteasomal degradation. These peptides are recognized on the cancer cell surface by CD8+ T cells. By contrast, HLA-II binds peptides 12–20 amino acids in length generated by protease-mediated digestion of extracellular proteins, which are subsequently recognized by CD4+ T cells. CD4+ and CD8+ T cells can collaborate to eradicate tumors [1].

Due to the high specificity, along with polymorphisms in HLA in the human population, only a limited subset of peptides can strongly bind to HLA for presentation on the cell surface. Mass spectrometry (MS) analysis of HLA-eluted ligands (EL) can provide valuable insight into peptide processing via HLA-mediated presentation that enable estimation of the probability a given peptide will be presented in vivo [3] [4]. Alternatively, binding affinity (BA) data can be used to verify peptides will indeed bind with HLAs. Thus, a number of tools have been developed to predict the likelihood a peptide fragment will serve as an effective antigen, but such tools trained only with EL data [5] [6] [7] or incorporating EL and BA data by pseudo-labeling [8] [9] [10] (Table 1), may lead to inadvertent loss of critical affinity-related information vital for T-cell recognition of a peptide. In addition, although various methods have been proposed for predicting peptide binding to HLA-I or HLA-II, these methods cannot accommodate both complexes, while those that do accommodate both remain unsatisfactory in prediction accuracy [11] [12]. These issues thus present the need for a versatile and reliable approach to predicting presentation probability and BA for both HLA-I and HLA-II, regardless of peptide length.

Most immunotherapies, including checkpoint inhibitors, exhibit efficacy in only a limited subset of patients [13]. The variability of response to immune checkpoint blockade (ICB) underscores the important to identify predictive biomarkers. Promising predictive biomarker is the tumor mutation/neoantigen burden (TMB/TNB). Nevertheless, relying solely on the number of mutant or neoantigen does not fully exploit the wealth information embedded in the entire repertoire of neoantigens and has demonstrated success in certain studies [14] [15] [16] while yielding inconclusive results in others [17] [18] [19]. Many studies have demonstrated that some neoantigens are more immunogenic than others [20] [21], a fact could prove critical in comprehending the responsiveness of cancer patients to

**Received:** June 29, 2024. **Revised:** September 13, 2024. **Accepted:** December 14, 2024

Table 1. Method and feature comparison of CapHLA and prior works.

| Method | Reference | Predict BA | Predict EL | HLA-I | HLA-II | Peptide length | Web server | Core module |
|---|---|---|---|---|---|---|---|---|
| CapHLA | This paper | ✓ | ✓ | ✓ | ✓ | 7–25 | ✓ | Convolution +attention |
| TripHLApan | Wang et al., 2024 | | ✓ | ✓ | ✓ | 7–25 | | BiGRU+ attention |
| NetMHCIIpan4.3 | Jonas et al., 2023 | ✓ | ✓ | | ✓ | 8-None | ✓ | MLP |
| BigMHC | Albert et al., 2023 | | ✓ | ✓ | | 8-None | | Ensemble of other methods |
| MixMHCpred | Gfeller et al., 2023 | | ✓ | ✓ | | 8–14 | | Motif deconvolution |
| MixMHC2pred | Guillaume et at., 2023 | | ✓ | | ✓ | 12–21 | ✓ | Motif deconvolution |
| DeepNeo | Kim et al., 2023 | | ✓ | ✓ | ✓ | 9,15 | ✓ | Convolution |
| TransPHLA | Chu et al., 2022 | | ✓ | ✓ | | 8–15 | ✓ | Transformer |
| DeepSeqPanII | Liu et al., 2022 | ✓ | | | ✓ | 8–26 | | RNN |
| NetMHCpan4.1 | Reynisson et al., 2020 | ✓ | ✓ | ✓ | | 8-None | ✓ | MLP |
| MHCflurry2.0 | Timothy et al., 2020 | ✓ | ✓ | ✓ | | None-15 | | MLP |

Cells with '✓' indicate that the method has the given feature. 'None' indicate no limitation of peptide length. MLP, multilayer perceptron; RNN, recurrent neural network.

immunotherapy treatment but has not been adequately captured by the basic mutation/neoantigen load approach.

In this study, we developed a novel convolution and attention-based model, CapHLA, in order to provide a comprehensive and accurate algorithm for predicting peptide BA and presentation probability with both HLA-I and HLA-II. Through analysis of immunogenic neoantigen validation data, results obtained with the CapHLA model demonstrate that the probability of peptide presentation and peptide BA are both essential for determining neoantigen immunogenicity. Based on those findings, we then developed a neoantigen quality scoring model that incorporates neoantigen presentation probability, with BA, and antigen expression level (i.e., a PAE score) to predict response to ICB treatment and patient survival in several cancers, which demonstrated significantly improved performance than previously described methods.

# Methods

## Dataset

The training dataset consists of 518,806 positive pHLA-I and 520 831 positive pHLA-II MS experimental validation data from the Immune Epitope Database (IEDB) [22] up to 2020, as well as data from published paper [6] [23] (Fig. S1A). The external validation dataset includes 51,984 pHLA-I data from published paper [23] and 48,238 pHLA-II data from the IEDB, spanning post-2020 to December 12, 2022 (Fig. S1C). The negative dataset comprises sequence segments randomly selected from the source proteins of the IEDB HLA immunopeptidomes. The length distribution and number of negative pMHC EL data corresponding to each allele are matched to those in the positive dataset. While false-negative peptides may potentially occur, their frequency and proportion are extremely low and can be considered negligible. This strategy for constructing negative samples ensures the dataset remains balanced.

Similarly, we compiled a dataset of 114,055 pHLA-I and 52,020 pHLA-II BA data from the IEDB, all predating 2020, for training the CapHLA-BA model (Fig. S1B). For external validation, we used a separate dataset containing 699 pHLA-I and 1743 pHLA-II BA data from the IEDB's records post-2020 to December 13, 2022 (Fig. S1D). BA values are measured as EC50 values in nanomolar (aff) and were rescaled to the interval [0,1] by applying $1-\frac{\log(aff)}{\log(50{,}000)}$, representing continuous target values [24].

## CapHLA

The fundamental concept behind CapHLA revolves around integrating of convolutional and attention-based mechanisms. The architectural framework consists of a series of techniques for model construction and optimization, incorporating four key sub-modules:

(1) The encoding block: In this block, both HLA proteins and peptides undergo one-hot encoding, resulting in numerical matrices. To accommodate variable input lengths, peptides are padded to a maximum length of 25 with X. The HLA molecule is represented as a pseudo-sequence of 34 amino acids [25]. Each amino acid in a peptide sequence is encoded as a 21-number one-hot vector (20 common amino acids + X). The resulting HLA and peptide one-hot matrices are concatenated to form a 59 × 21 matrix. While attempts were made with BLOSUM [26] and ProtVec [27] amino acid encodings, these had little influence on model performance.

(2) The convolution block: This block includes a pointwise convolution with an expansion factor of 2 to project the number of channels, followed by a GLU activation layer, a 1-D depthwise convolution, and batch normalization immediately after the 1-D depthwise convolution to aid in training deep models, then a SiLU activation layer. Finally, the block concludes with a pointwise convolution. The configuration specifies 3200 channels, a kernel size of 9, and a stride of 1.

(3) The attention block [28]: This block employs a multi-head self-attention mechanism to capture the intricate interactions between HLA molecules and peptides. It operate by mapping the query (Q) to a set of key-value (K-V) pairs and deriving an output. The K-V pairs are used to store sequence elements in memory, with the attention score (weight) calculated based on the correlation or similarity between Q and K. Model selection was conducted for both the layer and head of the multi-head attention mechanism, with final parameters specifying one layer and nine heads.

(4) The feature selection block: This block processes the features extracted by the preceding attention block via a dense layer. The convolution module first flattens the 21x59 matrix into 1239 linear features. A linear layer with 800 hidden variables is applied, followed by a SiLU activation layer, batch normalization, another linear layer with 64 hidden variables, and a ReLU activation layer. Finally, a linear layer generates a 2D vector for EL and a one-dimensional BA. A softmax function is used to convert 2D vector

into presentation probabilities. pHLA presentation probabilities higher than 0.5 are considered to positive pHLA.

Notably, the convolution block, attention block, and feature Selection block are interconnected through prenorm residual units with a dropout ratio 0.2, facilitating the training and regularization of deeper models. The sole distinction between the architecture of CapHLA-EL and CapHLA-BA models lies in the format of their final output data.

## CapHLA training

Model training was conducted on the RedHat Linux release 7.7 system. The GPUs used were two NVIDIA Tesla V100-PCIE with CUDA 11.4. The programming language utilized was Python 3.8.13, and the model was implemented using PyTorch 1.7.0. The Adam optimizer was used to minimize binary cross entropy loss during training. In this study, model evaluation during the training phase was performed using a fivefold cross-validation (CV) approach. This methodology involves dividing the training dataset into five equal partitions, with four of these partitions utilized for model training and the remaining partition reserved for evaluation under the same parameter settings. The training and evaluation process is repeated five times, ensuring that each portion of the data is used four times for training and once for evaluation. In each iteration, different negative data is introduced to enhance the model's generalization. Finally, the average value of the five models was taken as the final output, yielding results that outperform any individual model.

## Predictive performance metric calculation

For each EL data predictive model, the following metrics were calculated:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$Sensitivity(Recall) = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{FP + TN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F1\ socre = \frac{2 * Precision * Recall}{Precision + Recall}$$

$$MCC = \frac{(TP * TN) - (FN * FP)}{\sqrt{(TP + FN) * (TN + FP) * (TP + FP) * (TN + FN)}}$$

where TP is true positive, FP is false positive, FN is false negative and TN is true negative. Area under the receiver operating characteristic curve (AUROC) scores were calculated based on the area under sensitivity and 1 – specificity curves. AUPRC scores were calculated based on the area under precision and recall curves.

For each BA data predictive model, the following metrics were calculated:

$$MSE\ (Mean\ Squared\ Error) = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2$$

$$MAE\ (Mean\ Absolute\ Error) = \frac{1}{m}\sum_{i=1}^{m}\left|(y_i - \hat{y}_i)\right|$$

$$R2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (\bar{y} - y_i)^2}$$

where $y_i$ is the ith pHLA experimental score, $\hat{y}_i$ is the ith pHLA predicted score, $\bar{y}$ is the experimental mean score of all pHLA. Various indexes were implemented using Python scikit-learn.

## Immunogenic-validated peptide dataset

The immunogenic-validated peptide dataset from Wells was obtained from the supplementary table in the publication by Wells et al. [29] Peptides without expression level were excluded from the analysis. Similarly, the immunogenic-validated peptide dataset from Puig-Saus was sourced from the supplementary table in the paper by Puig-Saus et al. [21]. The expression levels of peptides were quantified using gene abundance calculated through our pipeline. The quality control and mapping procedures for raw data were the same as those used in ICB cohort pipeline. Gene abundance was quantified by featureCounts (v2.0.2) [30].

## PAE score

The foundation of the PAE score lies in its integration of pHLA presentation probability, BA, and peptide expression level. If the HLA gene remains unmutated, its expression level is usually sufficient, requiring only consideration of the peptide's expression level. To incorporate both pHLA presentation probability (PB) and BA into a unified measure, we adopted the composite score $(PB - 0.5) * BA$, thereby amplifying the influence of PB within the scoring framework. Incorporating peptide expression levels into the model required careful consideration. We hypothesized that at high expression levels, peptide variation was minimal, likely due to the saturation of HLA complexes on the cell surface. However, at low expression levels, significant differences became apparent. Therefore, we introduced an upper limit of 100 Transcripts Per Million (TPM) for expression levels and applied a logarithmic transformation. The PAE score of a peptide is calculated as $PAE = (PB - 0.5) * BA * \log(EP + 1)$. When integrating all neoantigens from a patient, we also incorporated variant allele frequency (VAF) and adjusted the score for each neoantigen. First, we combined all peptides of a mutation: $\frac{\sum_{\substack{i=1..n \\ PB_i>0.5 \\ BA_i>0.5}}((PB_i-0.5)*(BA_i)\log(EP_i+1))}{\sqrt{I(PB_i>0.5)}}$, which I(j) evaluates to 1 if the statement j is true, 0 otherwise. Only peptide with PB greater than 0.5 and BA greater than 0.5 are considered. Finally, we combined all mutation from a patient with VAF:

$$PAE = \sum_{k=1..m}\frac{VAF_k}{\overline{VAF}}\left(\frac{\sum_{\substack{i=1..n \\ PB_i>0.5 \\ BA_i>0.5}}((PB_i - 0.5) * (BA_i)\log(EP_i + 1))}{\sqrt{I(PB_i > 0.5)}}\right)$$

where $\overline{VAF}$ is the mean vaf of all mutaions.

## ICB cohort

To validate the performance of the PAE model, we collected data from five publicly available cohorts receiving ICB across three cancer types. Specifically, three melanoma cancer cohorts treated with anti-PD(L)1: the Hugo dataset [31], Riaz dataset [16], and Liu dataset [32]. One non-small cell lung cancer cohort treated with anti-PD(L)1, the Ravi dataset [33], and one clear cell renal cell carcinoma (ccRCC) cohort treated with anti-PD(L)1, the Miao dataset [34]. Due to the requirement for RNA-seq data, the available data is relatively limited. For all patient cohorts, where approval for access to raw exome-seq and RNA-seq data was obtained, we predicted somatic mutations and neoantigens using our in-house pipelines. Whole-exome sequencing data were aligned to the hg19 reference genome using Burrows-Wheeler Aligner (v0.7.15) [35]. Duplicate reads were removed with Picard (v2.23.3), and we performed fix-mate information and base quality score recalibration with Genome Analysis Toolkit (GATK, v4.1.9) [36]. MuTect2 [37], Strelka2 [38], VarScan2 [39], Vardict [40] and SomaticSniper [41]

were used to call somatic mutations. A mutation that was repeatedly called by any two of these software programs was retained. We used Oncotator (v1.9.9) [42] to annotate protein sequence change in somatic mutations. Peptides of fixed lengths containing mutated amino acids were extracted with homemade scripts. HLA-I and HLA-II genotypes were called using HD-HLA (v1.7.0) [43]. All tools used in processing and calling mutant peptides were run with default settings. RNA-seq data were first processed by fastp (v0.20.1) [44] to remove low quality reads from raw data using default parameters. High-quality reads were aligned to human genome by STAR (v2.7.5) [45] using the parameters—outFilterMultimapNmax 1 and ENCODE options. Transcript abundance was quantified by StringTie (v2.1.4) [46] using the default parameters. The overall expression of each somatic variant was calculated as the product of the transcript expression (summed across all overlapping protein-coding transcripts).

## TCGA data processing

Variant data, copy number variation, gene expression, and clinical data were obtained from the Pan-Cancer Atlas consortium's publications. These datasets were retrieved from the publication page at https://gdc.cancer.gov/about-data/publications/pancanatlas. Additionally, HLA-I genotypes for each sample were sourced from the pan-cancer landscape paper [47]. To ensure the confidentiality of the results obtained through Cox proportional hazards analysis, we applied a filter to exclude cancer types with fewer than 100 samples and a mortality ratio lower than 0.1. After filtration, we obtained 7228 samples across 23 tumor types.

## Statistical analyses

All computations and statistical analyses were conducted within the python computing environment. For the analysis of patient survival data, the Kaplan–Meier estimator was employed, and *P* values for log-rank tests were calculated using the lifelines package. To facilitate model comparison, we generated 5000 bootstrap resamples for each cohort. Each resample was then used to assess the predictive or prognostic performance of five distinct methods. To determine statistical significance, the P values derived from the 5000 bootstraps of each approach were compared using a two-sided Wilcoxon signed-rank test. To assess the association between dichotomized PAE scores and other methods with ordered response categories (e.g. complete response→partial response→stable disease→progressive disease), an ordinal $\chi 2$ test was used.

# Results

## Design and training of the CapHLA deep learning framework

In order to construct an accurate and comprehensive tool for predicting peptide presentation and binding with HLA molecules, we sought to integrate convolutional and attention-based mechanisms in a deep learning framework, which we termed CapHLA (Fig. 1). The overall model architecture includes four main submodules (1): an encoding block, in which both HLA proteins and peptides are subjected to one-hot encoding, resulting in numerical matrices used in later process (2); a convolution block designed to identify functional motifs or patterns inherent in the HLA and peptide sequences (3); an attention block that uses a multi-head self-attention [28] mechanism to detect weak and complicated interplay between HLA molecules and peptides (4); and a feature selection block that uses a fully connected layer to process features extracted by the preceding attention block, cumulatively resulting in peptide presentation probability and pHLA BA prediction outputs. In this architecture, the convolution block, attention block, and feature selection block are interlinked through residual connections. The CapHLA system features two modeling paradigms that both use the same deep-learning framework, CapHLA-EL for predicting EL and CapHLA-BA for BA prediction. The ablation analysis demonstrated that incorporating convolutional block with attention block yielded the highest AUROC on external validation sets for both HLA-I and HLA-II (Fig. S2A-B). Moreover, the complete model consistently outperformed others across various peptide lengths (Fig. S2C-D), highlighting the significance of each submodule.

## Comparison of CapHLA performance with existing methods

To validate the predictive performance of CapHLA, we performed a comparative analysis with other current tools tailored for each respective data type. Beginning with the EL HLA-I dataset, we compared CapHLA with five methods, including the method recommended by IEDB (NetMHCpan_EL [9]) and six other state-of-the-art methods (TripHLApan [48], BigMHC [5], MixMHC2pred [7], TransPHLA [10], DeepNeo-MHC [8], MHCflurry [49]) (Table 1). We found that CapHLA reached an area under the curve of 0.980 in receiver operating characteristic (AUROC) analysis and 0.979 in precision-recall curves (AUPRC) in predictions with HLA-I EL data (Fig. 2A, Fig. S3A). By contrast, TripHLApan achieved the highest performance among the comparison methods, with AUROC and AUPRC values of 0.962 and 0.964, respectively. We also compared accuracy, F1 score and Matthews correlation coefficient (MCC) across all methods (Fig. 2B), with CapHLA demonstrated the best overall performance. After further stratifying for peptide length, CapHLA had mean AUROC and AUPRC values of 0.958 and 0.961, respectively (Fig. 2C, Fig. S3B). These results indicated that CapHLA had higher predictive accuracy than other current methods, and across all tested peptide lengths. It should be noted that all methods except CapHLA exhibited a sharp decrease in accuracy in peptides longer than 11 residues.

For EL HLA-II data, we compared CapHLA with four methods, including the method recommended by IEDB (NetMHCIIpan_EL [50]) and three other recently published tools (TripHLApan [48], MixMHC2pred [6] and DeepNeo-MHC [8]). CapHLA had AUROC and AUPRC values of 0.967 and 0.970, respectively, whereas NetMHCIIpan_EL had the next highest accuracy, with AUROC and AUPRC values of 0.937 and 0.950, respectively (Fig. 2D, Fig. S3C). When considering accuracy, F1 score and MCC together, these results indicated that CapHLA could provide superior performance than other available models in EL prediction for both HLA-I and HLA-II (Fig. 2E). When testing accuracy across different peptide lengths, CapHLA consistently showed the highest accuracy among the tested methods for all peptide lengths, reaching mean AUROC and AUPRC scores of 0.969 and 0.972 (Fig. 2F, Fig. S3D). These results indicated that CapHLA could analyze complicated sequence features, resulting in higher predictive accuracy for longer peptides predictions and greater consistency across peptides of varying length.

It should be mentioned that since not all methods are compatible with every HLA allele or all peptide lengths (e.g. DeepNeo-MHC only supports nine and 15 aa peptides), we therefore compared performance using subsets of the external validation data. To ensure fair evaluations, we performed pairwise comparisons between CapHLA and other methods using the corresponding

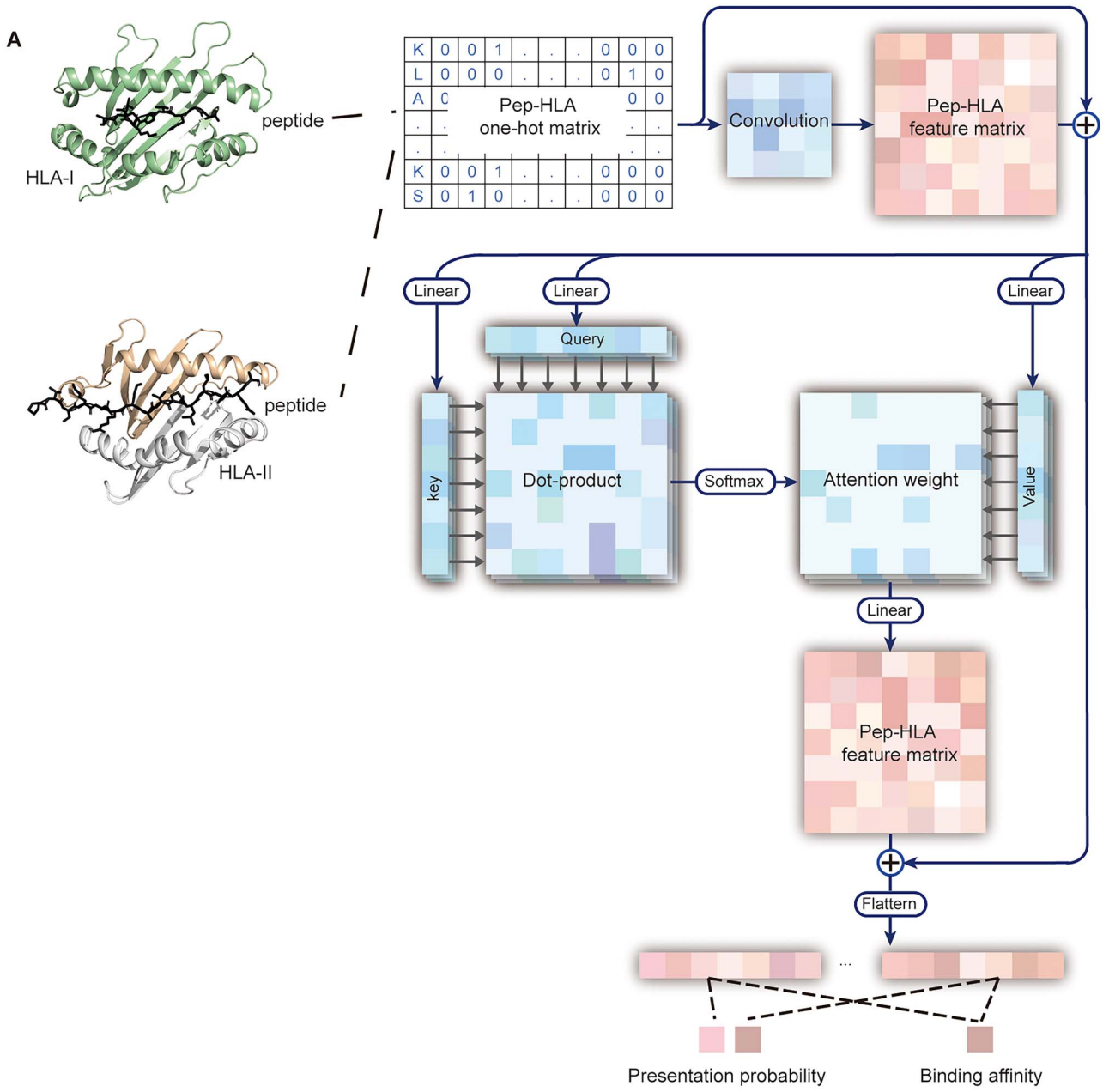


Figure 1. Schematic of the CapHLA model. CapHLA predominantly consists of an encoding layer, a convolutional layer, an attention layer, and a feature selection layer.

subset data (Fig. S3E-F). In these comparisons, CapHLA had consistently higher accuracy in EL predictions for both HLA-I and HLA-II, across a wide spectrum of alleles. Detailed analysis of CapHLA performance for each peptide length in the validation dataset showed that CapHLA achieve an AUROC of >0.95 across 8–15 aa peptides for HLA-I and 12–20 aa peptides for HLA-II (Fig. S4A). Analysis of CapHLA predictive accuracy across all HLA alleles in the validation dataset revealed that it could achieve AUROC values greater than 0.90 for 184 of 200 alleles (Fig. S4B), despite the limited availability of pHLAs for several alleles in the training data.

In comparative evaluations of BA prediction between CapHLA and several IEDB baseline methods (including NetMHCpan_BA [9] and MHCflurry [49] for HLA-I or NetMHCIIpan_BA [50] and DeepSeqPanII [51] for HLA-II), we found that CapHLA exhibited consistently higher accuracy in predicting pHLA BA than other methods, regardless of HLA-I or HLA-II context, with a mean absolute error (MAE) of <0.1 and a coefficient of determination (R2) exceeding 0.8 (Fig. 2G-H). Pairwise comparison between CapHLA-BA predictions of BA with that of experimentally determined BA scores showed that 93.13% (pHLA-I) and 97.70% (pHLA-II) fell within 0.2 of the experimental value in external validation dataset (Fig. 2I-J), whereas predictions by the comparison methods ranged from 58.02% to 62.03% (pHLA-I) and range from 53.81% to 63.05% (pHLA-II) within 0.2 of the experimental BA. These results showed that CapHLA had lower prediction error than other methods.

## The attention score of CapHLA uncovers the underlying patterns of pHLA

The attention block of CapHLA provides biological interpretability for the model. We initially aggregated the attention scores for each amino acid at various positions within peptides across different

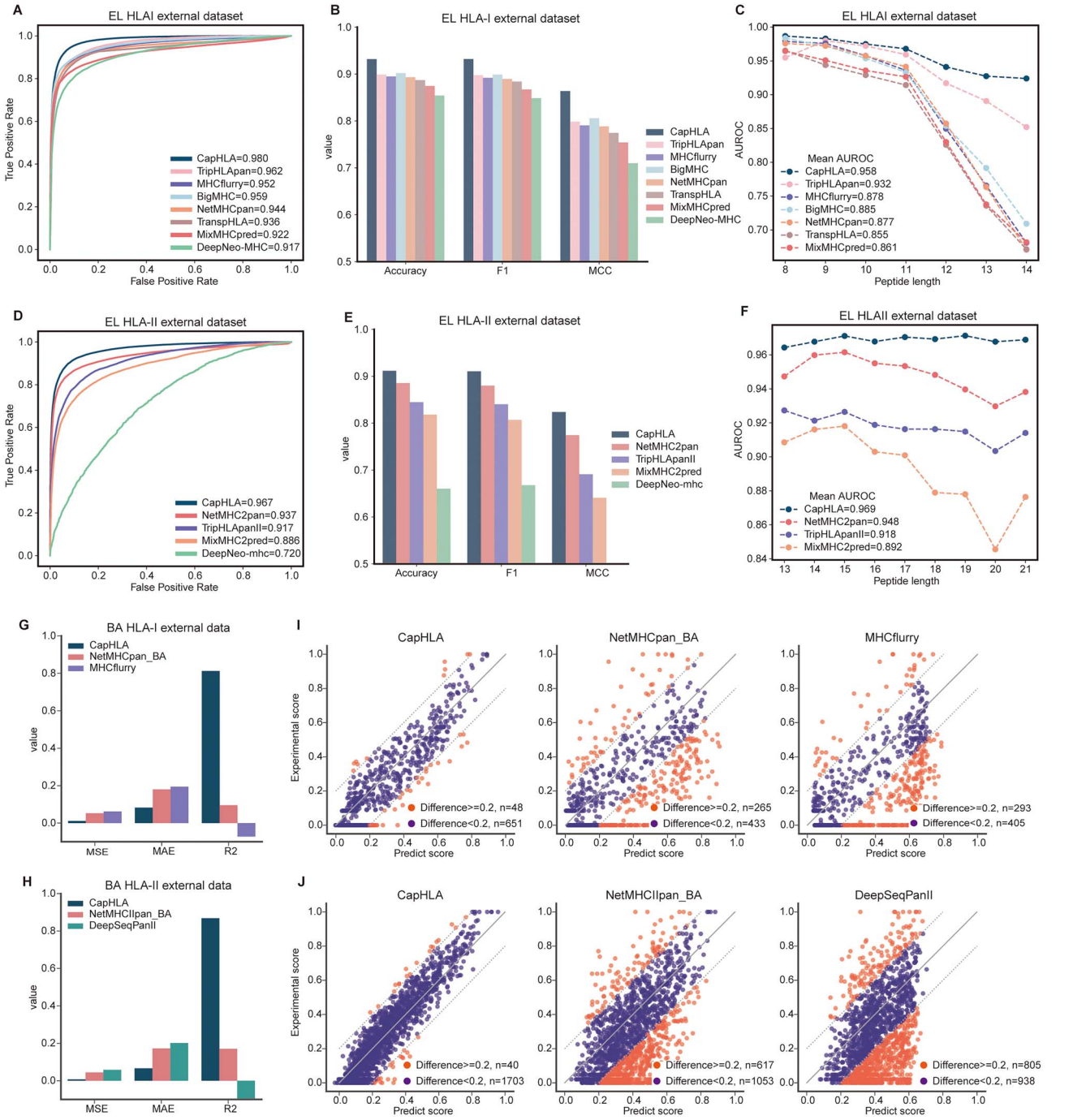


Figure 2. Comparison of CapHLA performance with other methods. (A-C) evaluation of CapHLA and other methods by ROC curve (A), accuracy, F1 score and MCC (B), AUROC per peptide length (C) using the EL HLA-I external dataset. (D-F) the same analysis as in A-C, but using the EL HLA-II external dataset. (G-H) comparison among methods by MSE, MAE, and R2 in the BA HLA-I (G) and HLA-II (H) external datasets. (I-J) experimental scores and predicted scores by CapHLA and other methods in the BA HLA-I (I) and HLA-II (J) external datasets. Orange dots represent points for which the prediction value deviates from the experimental value by >0.2.

sample types (Fig. S5). This analysis revealed that positive pHLA attention scores were focused on specific positions and amino acids, whereas negative pHLA scores remained consistent across all positions and amino acids. Further, the distribution of positive pHLA-II attention scores was obviously more diffuse than that of positive pHLA-I scores. This discrepancy could be attributed to the greater variability in the length of peptides binding with HLA-II compared to HLA-I-binding peptides, and the relatively high diversity of core binding region positions across the entire peptide.

To improve our understanding of the mechanism of pHLA binding, we next analyzed the patterns of binding and non-binding peptides for each HLA allele to identify the key amino acids at critical positions for pHLA binding. Attention scores for all HLA alleles and corresponding heatmaps are available for download from our Github repository. As expected, CapHLA identified patterns similar to those described in previous studies [10]. For HLA-A*11:01, CapHLA recognized that K was the predominant residue at the C-terminal position, while V or T were the most frequent residues

at position 2 (P2). This is consistent with previous studies showing that P2 and the C-terminus served as anchor residues [52]. (Fig. 3A). We observed that peptides of varying lengths maintained a hydrogen bonding network by conserved anchoring sites near N- or C-terminus, while the central residues of such peptides were typically showed high variability, which was also demonstrated by previous studies [53] [54]. Additionally, we noted that other HLA-I allelic peptide patterns were consistent with structures in protein data bank (2FYY for B*35:01 [55], 1M6O for B*44:02 [56], 5VGD for C*05:01 [57]) (Fig. 3B). For HLA-II-binding peptides, the amino acid attention patterns are generally more complex than for HLA-I-binding peptides, due to the extended conformation of the peptides within the groove [58]. In the case of HLA-DRB1*03:01, CapHLA identified the first anchor residue (corresponding to the residue at position 4 or 5, which aligns with the position 1 in the binding core) as typically being I, L, or V, while the second to fourth residues were D, K, and R, respectively (Fig. 3C). Along with the attention patterns defined for other HLA-II alleles, these results align well with previous studies [6] and demonstrate the effectiveness of CapHLA attention block.

## Immunogenic peptide features

T cell responses are typically elicited by a relatively small subset of peptides among numerous potential candidates. For instance, even when using a comprehensive consortium approach, only 37 out of 608 tested pMHCs were found to be engaged by patient-matched T cells [29]. To examine whether CapHLA could be used to predict peptide immunogenicity, we analyzed the immunogenicity characteristics of experimentally validated peptides from a dataset published by Wells and co-workers (Wells dataset), including presentation probability (PB) which predicted by CapHLA-EL, BA that predicted by CapHLA-BA and expression level of each peptide by tumor cells (EP) (Fig. 4A). As anticipated, CapHLA-EL predicted high PB values for almost all of these immunogenic peptides, with only one peptide displaying a predicted presentation probability <0.8. Intriguingly, 191 of 591 non-immunogenic peptides were predicted to have high PB but low BA. By contrast, 39 0f 41 immunogenic peptides in the Wells dataset had high BA scores (>0.5 representing 224 nM). In addition to the high PB and BA scores, we found that EP values were also high for these immunogenic peptides. In another experimentally validated dataset of immunogenic peptides from Puig-Saus et al. [21] (Fig. S6), we also observed high values of PB, BA, and EP, suggesting that these characteristics are important features of immunogenic peptides.

We then used a contingency table with the Wells dataset to investigate whether these three features conferred different effects on immunogenicity predictions, and whether they could be used in combination. By applying a PB threshold of 0.8, 32% of non-immunogenic peptides were effectively filtered out, while 97% of immunogenic peptides were retained (Fig. 4B, $P = 1.3*10^{-4}$). Introducing a BA threshold of >0.5 led to removal of 55% of the non-immunogenic peptides while retaining 91% of the immunogenic peptides (Fig. 4C, $P = 9.8*10^{-8}$). Further incorporating an EP threshold >10 TPM resulted in the elimination of 80% of non-immunogenic peptides but retention of 76% of immunogenic peptides (Fig. 4D, $P = 3.2*10^{-11}$).

Subsequent ROC analysis indicated that PB provided greater power than BA or EP in discriminate immunogenic from non-immunogenic peptides (Fig. 4E). To exploit these advantages, we adopted a composite scoring index of $(PB - 0.5) * BA$ to effectively amplify the influence of PB while retaining the discriminatory power of BA. AUROC analysis of this PB + BA index indicated that it could provide an accuracy of 0.841 in immunogenicity predictions. When incorporating EP into this model, we speculated that there is less variation in immunogenicity between highly expressed peptides due to saturation of antigen binding with HLA complexes in cell surface, whereas EP may confer a markedly greater effect on immunogenicity of weakly expressed peptides. We thus used an upper EP limit of 100 TPM and applied the log value of EP to calculate this joint probability/affinity/expression (i.e., PAE) score, as follows: $PAE = (PB - 0.5) * BA * \log(EP + 1)$ .We found that this PAE index could yield an AUROC of 0.858, which suggested that combining these features could substantially bolster the capacity to predict peptide immunogenicity.

## Higher PAE scores associated with better response to checkpoint inhibitors

Based on our above findings, we then tested the PAE index of neoantigen quality for assessing potential benefits of ICB in individual patients. For this analysis, we assembled data from whole exon sequencing and RNA-seq data from five cohorts of patients treated with ICB. The cohort data were processed through our pipeline to generate PAE scores for all patients (Fig. S7). We only considered neoantigens that bind with HLA-I molecules, since PAE scores rely on the available results of validation of HLA-I peptide immunogenicity, while the parameters for determining HLA-II peptide immunogenicity remains unknown. Upon integrating all neoantigens from a patient, we also incorporated VAF and adjusted the score for each neoantigen (see methods).

This analysis was conducted using data from three cohorts of melanoma patients (the Hugo [31], Riaz [16], and Liu cohorts [32]) treated with anti–PD-1/anti–PD-L1 therapies. In the Hugo and Liu cohorts, higher PAE scores were observed in patients with significantly longer survival times (Fig. 5A, $P = 8.9*10^{-3}$; Fig. 5C, $P = 6.9*10^{-5}$). In addition, higher PAE was associated with better treatment response in the Hugo and Liu cohorts (Fig. S9A, $P = 0.07$; $P = 0.008$). By contrast, higher PAE scores were not significantly correlated with either extended survival or improved treatment response in the Riaz cohort. However, positive responders (either complete or partial responses, CR + PR) had higher PAE scores than non-responders (stable disease and progressive disease, SD + PD) (Fig. 5B, Fig. S9A). Since Riaz and colleagues specifically selected patients unresponsive to CTLA-4 therapy for PD-1 treatment, predicting survival time is more challenging in these patients. To test non-small cell lung cancer (NSCLC) data, we examined patients treated with anti–PD-1/anti–PD-L1 regimens in a study by Ravi et al. [33] In this cohort, PAE shared a significant positive association with survival time and treatment response (Fig. 5D, $P = 0.0031$; Fig. S9A, $P = 0.002$). In a ccRCC cohort reported by Miao et al. [34], we observed that ccRCC patients treated with anti–PD-1/anti–PD-L1 strategies who had higher PAE scores also had significantly longer survival time and response trended better, albeit without reaching statistical significance (Fig. 5E, $P = 0.014$; Fig. S9A).

To compare the performance of PAE scores with other metrics used for neoantigen and mutation-related markers, we also examined the predictive power of DeepNeo [59] immunogenicity neoantigen burden, CSiN [60] score, NetMHCpan [9] neoantigen burden and TMB (Fig. S8A-D, Fig. S9B-E) in the same cohorts and using the same statistical tests. To ensure unbiased comparison, only HLA-I-binding neoantigens were taken into consideration when assessing neoantigen-related markers. We found that all four models had lower predictive accuracy of treatment

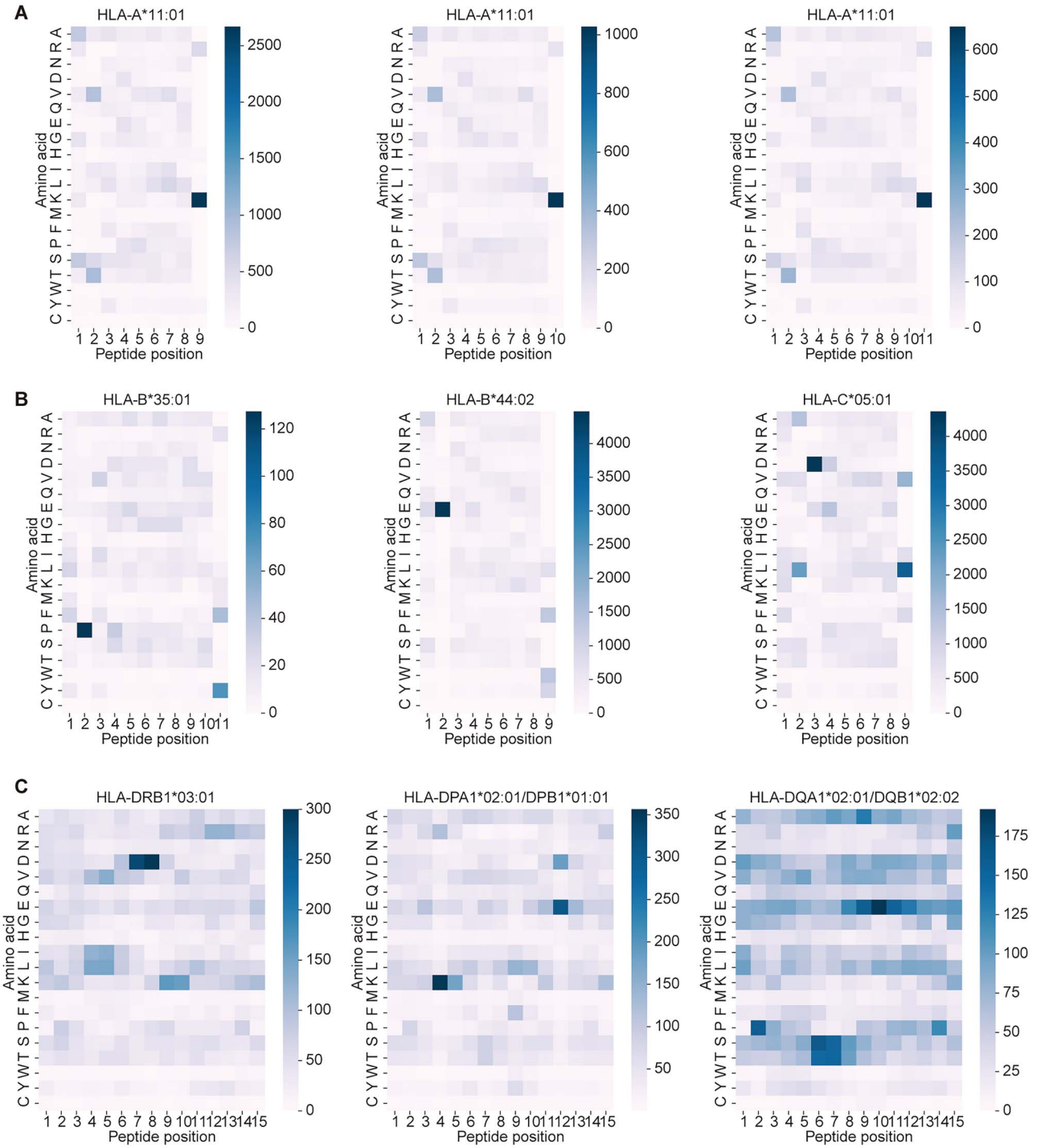


Figure 3. Heatmap visualization of cumulative attention scores for amino acids at each peptide position. Cumulative attention scores of amino acid binding at each peptide position for several well-characterized HLA alleles. (A) 9-mer, 10-mer, and 11-mer peptides binding with HLA-A*11:01. (B) 11-mer peptides binding with HLA-B*35:01, 9-mer peptides binding with HLA-B*44:02 and HLA-C*05:01. (C) 15-mer peptides binding with HLA-DRB1*03:01, HLA-PA1*02:01/DPB1*01:01, and HLADQA1* 02:01/DQB1*02:02.

response compared to PAE score. In addition, PAE score also provided the most robust prognostic capability across four cohorts, with the exception of the Riaz melanoma cohort, in which CSiN had the highest accuracy. Subsequent bootstrap analysis to evaluate the statistical significance of PAE improvement over the other four approaches, as conducted in other model comparisons [61], showed that CapHLA significantly outperformed than other methods in four of the five cohorts evaluated, except CSiN in the Riaz melanoma cohort Fig. 5F).

Next, we evaluated whether PAE captures unique information not reflected by established immunotherapy biomarkers. Specifically, we calculated the T-cell inflamed gene expression profile (GEP) [62], innate anti-PD-1 resistance gene signature (IPRES) [63], immuno-predictive score (IMPRES) [63], and MSI score as described in previous studies [64]. Additionally, we used sample gene set enrichment analysis (ssGSEA) to quantify tumor immune microenvironment-related gene signatures, including Angio [65] and myeloid inflammation [66]. The Spearman

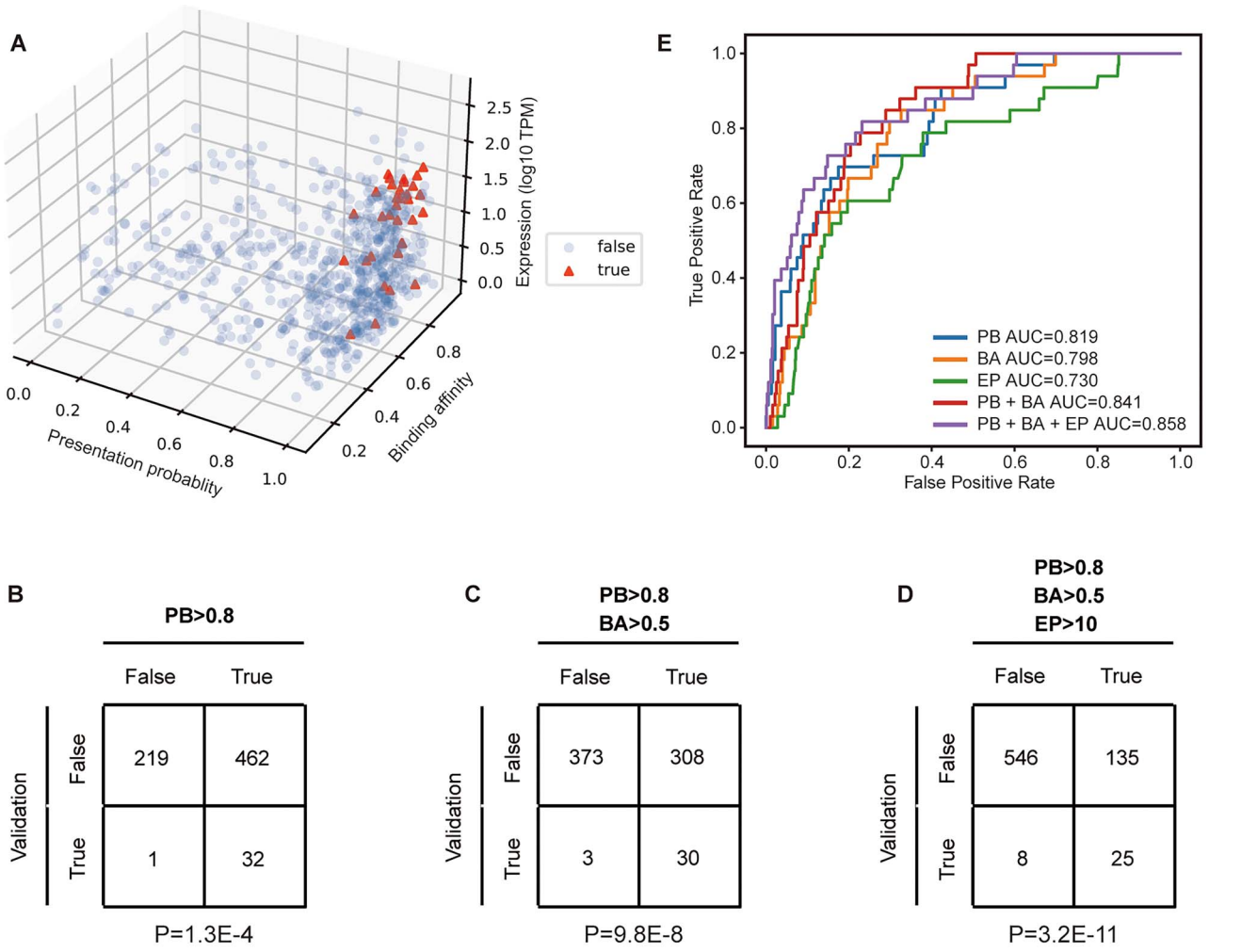


Figure 4. Features of immunogenic peptides. (A) a 3D plot of presentation probability, BA, and expression level of peptides from an immunogenic experiment. (B-D) contingency tables comparing validation status to PB status (B), PB and BA status (C), and PB, BA, and EP status (D) across all peptides. P values were calculated by Fisher's exact test. (E) Various feature combinations were assessed by ROC curves for their value in predicting peptide immunogenicity.

correlation between PAE score and these biomarker in each cohort showed that PAE functions as an independent biomarker for immunotherapy, with Spearman correlations below 0.5 for all comparisons (Fig. 5G).

## Higher PAE can predict more favorable prognosis in cancers

To investigate the relationship between neoantigen quality and long-term patient survival in various tumor types, we examined PAE association with patient prognosis in datasets from the Cancer Genome Atlas (TCGA). Employing Cox regression analysis, we calculated hazard ratios, along with the corresponding confidence intervals and significance level, to evaluate associations between scoring with the above methods and patient survival (Fig. 6A) showed that PAE was significantly negatively correlated with mortality risk (hazard ratio < 1, p value <0.05) in the UCEC (n = 417), KIRP (n = 238), STAD (n = 314), and BLCA (n = 384) cohorts. These results indicated that patients with higher PAE score had a reduced risk of mortality, whereas with the other methodologies, samples predicted with higher immunogenicity were not significantly associated with reduced mortality risk. In SKCM (n = 100), PAE score was negatively correlated with mortality risk but failed to reach significance (hazard ratio < 1, p value = 0.17), potentially due to the relatively small sample size in this cohort.

The patients with higher PAE score also survived significantly longer in UCEC, SKCM, KIRP, STAD, and BLCA cohorts (Fig. 6B-F), while no significant associations were identified through the same analysis for other models (Fig. S10). Bootstrap analysis evaluating the statistical significance of this comparison further highlighted the greater reliability of CapHLA over other methodologies in four of the five evaluated cohorts, while TMB had higher predictive power in the BLCA cohort Fig. 6G). These results implied that if patients with higher PAE scores tended to exhibit prolonged survival in a specific type of cancer, then it is likely that this cancer type may have characteristically greater T cell infiltration [67, 68]. Such tumor patients may have increased likelihood of positive response to ICB treatment.

## Conclusions

Peptide binding and presentation with HLA-I and HLA-II proteins is an essential step for initiating robust T-cell recognition. Therefore, delineating effective targets of immunotherapy, epitope screening, and vaccine design require accurate prediction of

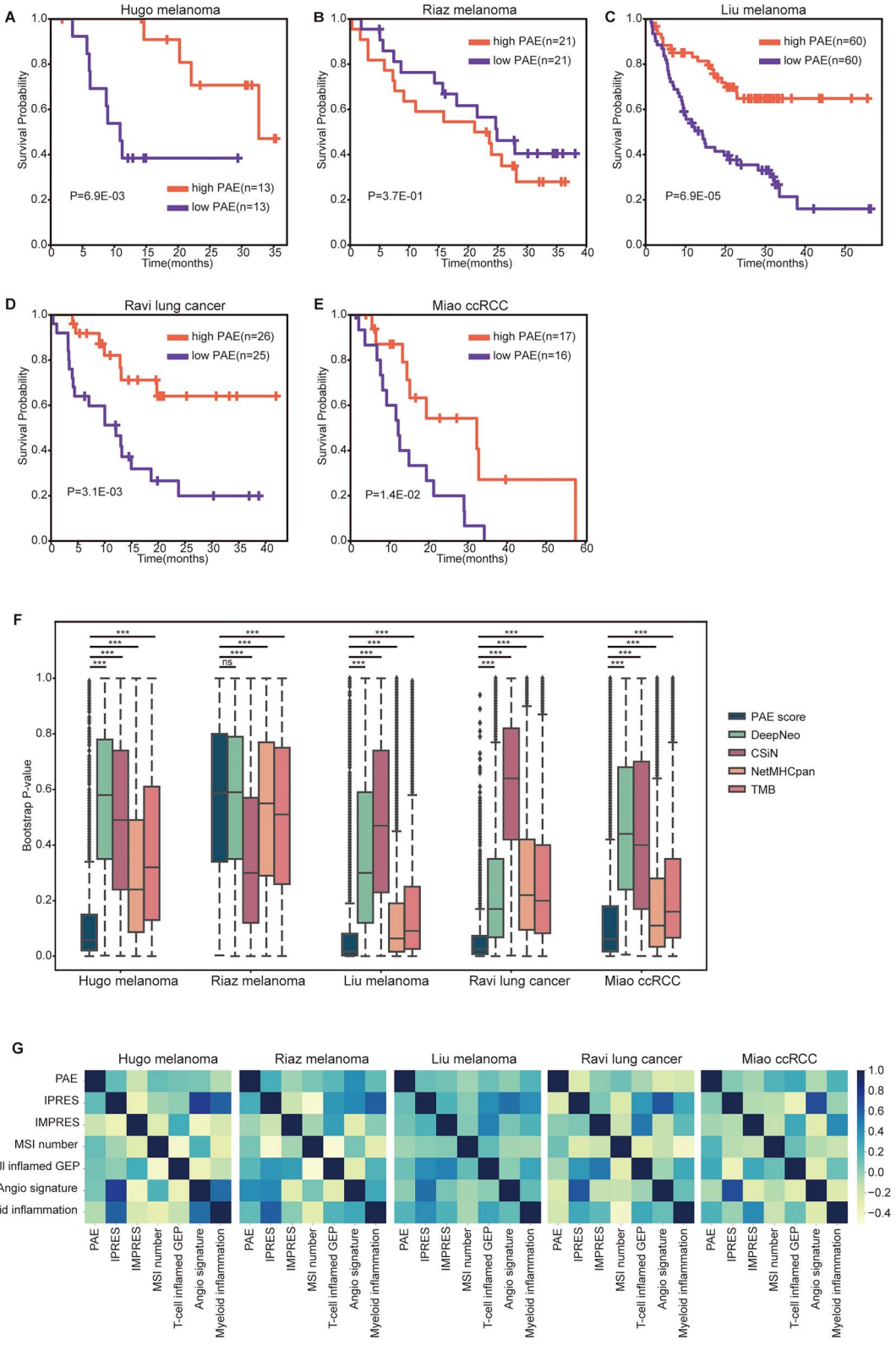


Figure 5. Association of PAE score with overall survival of ICB patients. (A-E) Kaplan–Meier survival curves were used to estimate overall survival stratified by PAE-high/−low status. P values for log-rank tests are shown. (A) the Hugo melanoma cohort. (B) the Riaz melanoma cohort. (C) the Liu melanoma cohort. (D) the Ravi NSCLC cohort. (E) the Miao clear cell renal cell carcinoma (ccRCC) cohort. (F) Boxplots of bootstrap P values evaluating the robustness of prognostic performance by PAE, DeepNeo, CSiN, NetMHCpan, and TMB, with each P value generated from a bootstrap resample of each cohort. Two-sided Wilcoxon signed-rank test was used to compare the bootstrap P values. $^{***}P < 0.001$. (G) Heatmaps of the pairwise spearman correlations of the CapHLA, IPRES, IMPRES, MSI number, IFN-$\gamma$signature, T cell inflamed GEP, Angio signature, and myeloid inflammation are shown for all cohorts.

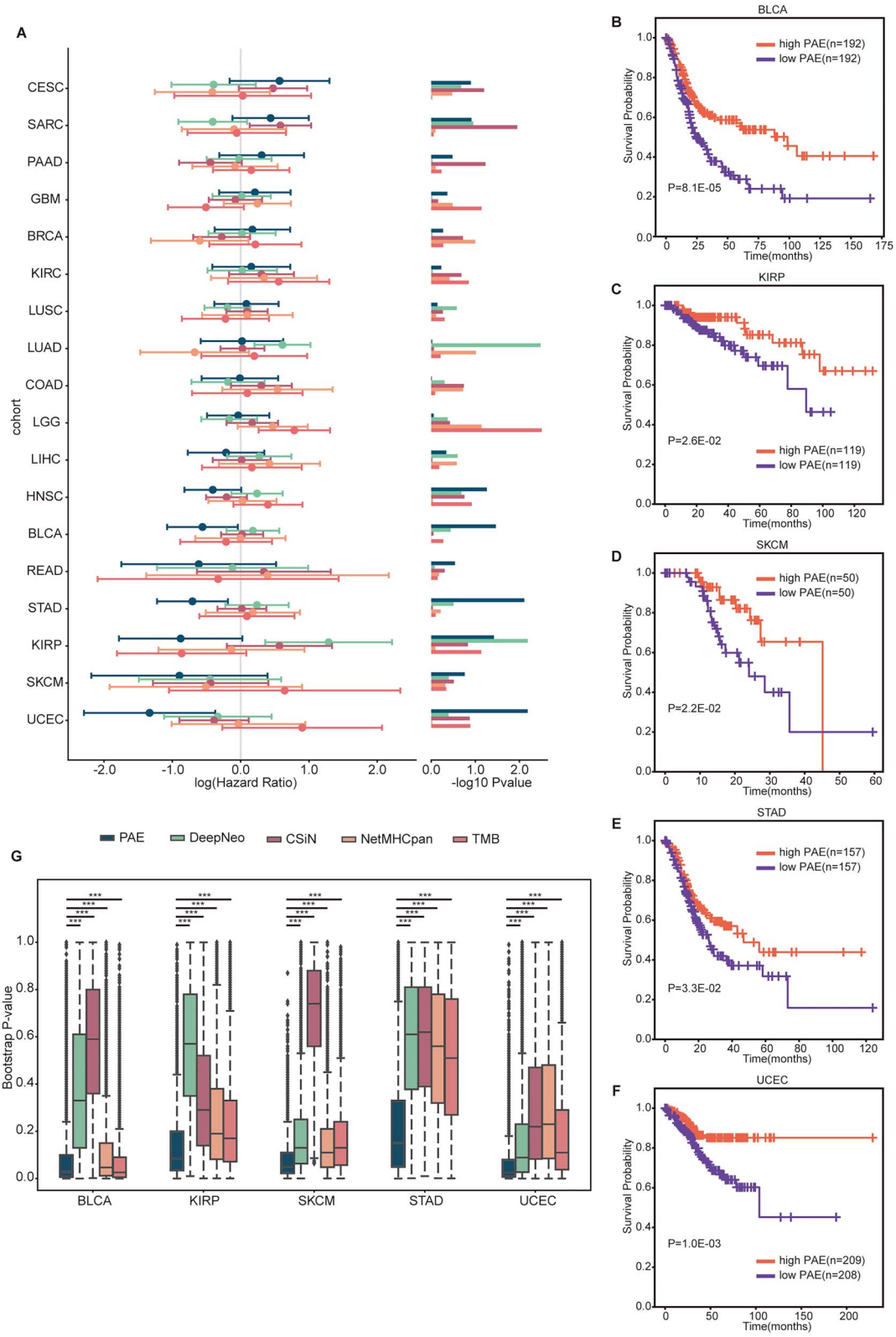


Figure 6. Association analysis of PAE scores with overall patient survival in TCGA data. (A) Forest plot of hazard ratio calculated by cox proportional hazards analysis of the TCGA cohort. Bar plot shows the -log10 p-value of the hazard ratio. (B-F) Kaplan–Meier survival curves were used to estimate overall survival stratified by PAE-high/—low status in the (B) the BLCA cohort; (C) the KIRP cohort; (D) the SKCM cohort; (E) the STAD cohort; (F) and the UCEC cohort. P values for log-rank tests are shown. (G) Boxplots of bootstrap P values evaluating the robustness of prognostic performance by PAE, DeepNeo, CSiN, NetMHCpan, and TMB, with each P value generated by bootstrap resampling of each cohort. Two-sided Wilcoxon signed-rank tests were used to compare the bootstrap P values. ***$P < 0.001$.

pHLA binding and likelihood of presentation. Here, in the current study, we developed CapHLA, a convolution and attention-based model, which is a generalized pan-specific method that is not restricted to a single HLA allele or peptide length. Due to the use of a convolutional and attention model that can effectively capture a wider spectrum of features in amino acid sequence features, CapHLA showed consistently higher accuracy, especially in predicting pHLA-II and non-classic length peptide binding, compared to other advanced tools, including the approach recommended by the IEDB. These results thus suggest that CapHLA can provide a comprehensive and higher accuracy solution for predicting pHLA binding and presentation. Moreover, prediction of peptide binding and surface presentation for both HLA-I and HLA-II using the same model provides greater convenience and comparability across studies to facilitate independent functional investigations.

In recent neoantigen prediction models, EL datasets are increasingly favored, while attention to BA data has waned. However, BA data provide highly relevant quantitative insights into BA that can complement the likelihood that a peptide will be presented in vivo. Through the analysis of recently published immunogenicity-validated neoantigens dataset, we underscore BA is a key parameter in immunogenic neoantigens. Based on these findings, we established PAE score as an informative quality index for screening neoantigens that incorporates peptide features of BA, presentation probability, and expression level to substantially bolster the predictive capacity for reliable prediction of peptide immunogenicity. Furthermore, by integrating all neoantigens within a patient sample, PAE score could potentially serve as a biomarker for predicting response to ICB treatment, as evidenced in our analysis of ICB cohorts and TCGA cohorts.

## Supplementary Data

Supplementary data are available at *Briefings in Bioinformatics* online.

## Acknowledgements

We thank all members of Dr. Ligang Wu's laboratory for their discussion and comments on this project. We appreciate the IT group of Center for Excellence in Molecular Cell Science for setting up the online web platform. We thank Qingfei Ren for her insightful suggestions in immunology and Zhen Zhao for his nice HLA protein cartoon picture.

## Author contributions

Y.C. and L.W. conceived the research and wrote the manuscript. Y.C. designed the model, collected data and wrote the code. L.W. supervised the work.

Conflict of interest: The authors declare no competing interests.

## Funding

This work was supported by the following funding: the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB0570000), the National Key R&D Program of China (2021YFA1100201 and 2022YFA1303301), and the National Natural Science Foundation of China (31970607 and 31470781) to Ligang Wu.

## Code availability

All code used in this study and the final trained models are provided in our public GitHub repository: https://github.com/changyunjian/CapHLA. Users could update CapHLA by themselves data using our training procedure. CapHLA is also accessible to the public via http://www.sibcb.ac.cn/hpc/portal/program.jsp.

## Data availability

All data, including model training and validation data, are provided in github (https://github.com/changyunjian/CapHLA). Raw sequencing data of Puig-Saus dataset a is available at dbGAP under accession number phs003153. Raw sequencing data of cohorts that received ICB therapy were obtained from each publication. More specifically, data were retrieved with the following accession numbers: Hugo melanoma [31] (SRP067938, SRP090294), Riaz melanoma [16] (SRP094781), Liu melanoma [32] (phs000980), Ravi lung cancer [33] (phs002822), Miao [34] ccRCC (phs001493). The Cancer Genome Atlas data were obtained from the GDC Data Portal (https://portal.gdc.cancer.gov/).

## References

1. Lang F, Schrörs B, Löwer M. *et al.* Identification of neoantigens for individualized therapeutic cancer vaccines. *Nat Rev Drug Discov* 2022;**21**:261–82. https://doi.org/10.1038/s41573-021-00387-y.
2. Huppa JB, Axmann M, Mörtelmaier MA. *et al.* TCR–peptide–MHC interactions in situ show accelerated kinetics and increased affinity. *Nature* 2010;**463**:963–7. https://doi.org/10.1038/nature08746.
3. Abelin JG, Keskin DB, Sarkizova S. *et al.* Mass spectrometry profiling of HLA-associated Peptidomes in mono-allelic cells enables more accurate epitope prediction. *Immunity* 2017;**46**:315–26. https://doi.org/10.1016/j.immuni.2017.02.007.
4. Abelin JG, Harjanto D, Malloy M. *et al.* Defining HLA-II ligand processing and binding rules with mass spectrometry enhances cancer epitope prediction. *Immunity* 2019;**51**:766–779.e717. https://doi.org/10.1016/j.immuni.2019.08.012.
5. Albert BA, Yang Y, Shao XM. *et al.* Deep neural networks predict class I major histocompatibility complex epitope presentation and transfer learn neoepitope immunogenicity. *Nature Machine Intelligence* 2023;**5**:861–72. https://doi.org/10.1038/s42256-023-00694-6.
6. Racle J, Guillaume P, Schmidt J. *et al.* Machine learning predictions of MHC-II specificities reveal alternative binding mode of class II epitopes. *Immunity* 2023;**56**:1359–1375.e1313. https://doi.org/10.1016/j.immuni.2023.03.009.
7. Gfeller D, Schmidt J, Croce G. *et al.* Improved predictions of antigen presentation and TCR recognition with MixMHCpred2.2 and PRIME2.0 reveal potent SARS-CoV-2 CD8(+) T-cell epitopes. *Cell Syst* 2023;**14**:72–83.e75. https://doi.org/10.1016/j.cels.2022.12.002.
8. Kim K, Kim HS, Kim JY. *et al.* Predicting clinical benefit of immunotherapy by antigenic or functional mutations affecting tumour immunogenicity. *Nat Commun* 2020;**11**:951. https://doi.org/10.1038/s41467-020-14562-z.
9. Reynisson B, Alvarez B, Paul S. *et al.* NetMHCpan-4.1 and NetMHCIIpan-4.0: Improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res* 2020;**48**:W449–54. https://doi.org/10.1093/nar/gkaa379.

10. Chu Y, Zhang Y, Wang Q. *et al.* A transformer-based model to predict peptide–HLA class I binding and optimize mutated peptides for vaccine design. *Nature Machine Intelligence* 2022;**4**: 300–11. https://doi.org/10.1038/s42256-022-00459-7.
11. Richters MM, Xia H, Campbell KM. *et al.* Best practices for bioinformatic characterization of neoantigens for clinical utility. *Genome Med* 2019;**11**:56. https://doi.org/10.1186/s13073-019-0666-2.
12. Wang M, Kurgan L, Li M. A comprehensive assessment and comparison of tools for HLA class I peptide-binding prediction. *Brief Bioinform* 2023;**24**:bbad150. https://doi.org/10.1093/bib/bbad150.
13. Ascierto PA, Marincola FM. 2015: The year of anti-PD-1/PD-L1s against melanoma and beyond. *EBioMedicine* 2015;**2**:92–3. https://doi.org/10.1016/j.ebiom.2015.01.011.
14. Anagnostou V, Smith KN, Forde PM. *et al.* Evolution of Neoantigen landscape during immune checkpoint blockade in non–small cell lung cancer. *Cancer Discov* 2017;**7**:264–76. https://doi.org/10.1158/2159-8290.Cd-16-0828.
15. Rizvi NA, Hellmann MD, Snyder A. *et al.* Mutational landscape determines sensitivity to PD-1 blockade in non–small cell lung cancer. *Science* 2015;**348**:124–8. https://doi.org/10.1126/science.aaa1348.
16. Riaz N, Havel JJ, Makarov V. *et al.* Tumor and microenvironment evolution during immunotherapy with Nivolumab. *Cell* 2017;**171**:934–949.e916. https://doi.org/10.1016/j.cell.2017.09.028.
17. Matsushita H, Sato Y, Karasaki T. *et al.* Neoantigen load, antigen presentation machinery, and immune signatures determine prognosis in clear cell renal cell carcinoma. *Cancer Immunol Res* 2016;**4**:463–71. https://doi.org/10.1158/2326-6066.CIR-15-0225.
18. Miller A, Asmann Y, Cattaneo L. *et al.* High somatic mutation and neoantigen burden are correlated with decreased progression-free survival in multiple myeloma. *Blood Cancer J* 2017;**7**:e612–2. https://doi.org/10.1038/bcj.2017.94.
19. Matsushita H, Hasegawa K, Oda K. *et al.* The frequency of neoantigens per somatic mutation rather than overall mutational load or number of predicted neoantigens per se is a prognostic factor in ovarian clear cell carcinoma. *Onco Targets Ther* 2017;**6**:e1338996. https://doi.org/10.1080/2162402x.2017.1338996.
20. Ott PA, Hu-Lieskovan S, Chmielowski B. *et al.* A phase Ib trial of personalized Neoantigen therapy plus anti-PD-1 in patients with advanced melanoma, non-small cell lung cancer, or bladder cancer. *Cell* 2020;**183**:347–362.e324. https://doi.org/10.1016/j.cell.2020.08.053.
21. Puig-Saus C, Sennino B, Peng S. *et al.* Neoantigen-targeted CD8+ T cell responses with PD-1 blockade therapy. *Nature* 2023;**615**: 697–704. https://doi.org/10.1038/s41586-023-05787-1.
22. Vita R, Mahajan S, Overton JA. *et al.* The immune epitope database (IEDB): 2018 update. *Nucleic Acids Res* 2019;**47**: D339–d343. https://doi.org/10.1093/nar/gky1006.
23. Mei S, Li F, Xiang D. *et al.* Anthem: A user customised tool for fast and accurate prediction of binding between peptides and HLA class I molecules. *Brief Bioinform* 2021;**22**:bbaa415. https://doi.org/10.1093/bib/bbaa415.
24. 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**:1007–17. https://doi.org/10.1110/ps.0239403.
25. Nielsen M, Lundegaard C, Blicher T. *et al.* NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. *PloS One* 2007;**2**:e796. https://doi.org/10.1371/journal.pone.0000796.
26. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* 1992;**89**:10915–9. https://doi.org/10.1073/pnas.89.22.10915.
27. Asgari E, Mofrad MR. Continuous distributed representation of biological sequences for deep proteomics and genomics. *PloS One* 2015;**10**:e0141287. https://doi.org/10.1371/journal.pone.0141287.
28. Vaswani A. *et al.* In: Luxburg U. *et al.* (eds.) *Proceedings of the 31st International Conference on Neural Information Processing Systems 6000–6010*. Long Beach, California, USA: Curran Associates Inc., 2017.
29. Wells DK, van Buuren MM, Dang KK. *et al.* Key parameters of tumor epitope immunogenicity revealed through a consortium approach improve Neoantigen prediction. *Cell* 2020;**183**:818–834.e813. https://doi.org/10.1016/j.cell.2020.09.015.
30. Liao Y, Smyth GK, Shi W. featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 2014;**30**:923–30. https://doi.org/10.1093/bioinformatics/btt656.
31. Hugo W, Zaretsky JM, Sun L. *et al.* Genomic and transcriptomic features of response to anti-PD-1 therapy in metastatic melanoma. *Cell* 2016;**165**:35–44. https://doi.org/doi:10.1016/j.cell.2016.02.065.
32. Liu D, Schilling B, Liu D. *et al.* Integrative molecular and clinical modeling of clinical outcomes to PD1 blockade in patients with metastatic melanoma. *Nat Med* 2019;**25**:1916–27. https://doi.org/10.1038/s41591-019-0654-5.
33. Ravi A, Hellmann MD, Arniella MB. *et al.* Genomic and transcriptomic analysis of checkpoint blockade response in advanced non-small cell lung cancer. *Nat Genet* 2023;**55**:807–19. https://doi.org/10.1038/s41588-023-01355-5.
34. Miao D, Margolis CA, Gao W. *et al.* Genomic correlates of response to immune checkpoint therapies in clear cell renal cell carcinoma. *Science* 2018;**359**:801–6. https://doi.org/10.1126/science.aan5951.
35. Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics* 2009;**25**:1754–60. https://doi.org/10.1093/bioinformatics/btp324.
36. McKenna A, Hanna M, Banks E. *et al.* The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;**20**:1297–303. https://doi.org/10.1101/gr.107524.110.
37. Cibulskis K, Lawrence MS, Carter SL. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* 2013;**31**:213–9. https://doi.org/10.1038/nbt.2514.
38. Kim S, Scheffler K, Halpern AL. *et al.* Strelka2: Fast and accurate calling of germline and somatic variants. *Nat Methods* 2018;**15**: 591–4. https://doi.org/10.1038/s41592-018-0051-x.
39. Koboldt DC, Zhang Q, Larson DE. *et al.* VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* 2012;**22**:568–76. https://doi.org/10.1101/gr.129684.111.
40. Lai Z, Markovets A, Ahdesmaki M. *et al.* VarDict: A novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res* 2016;**44**:e108. https://doi.org/10.1093/nar/gkw227.
41. Larson DE, Harris CC, Chen K. *et al.* SomaticSniper: Identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* 2012;**28**:311–7. https://doi.org/10.1093/bioinformatics/btr665.
42. Ramos AH, Lichtenstein L, Gupta M. *et al.* Oncotator: Cancer variant annotation tool. *Hum Mutat* 2015;**36**:E2423–9. https://doi.org/10.1002/humu.22771.

43. Kawaguchi S, Higasa K, Shimizu M. *et al.* HLA-HD: An accurate HLA typing algorithm for next-generation sequencing data. *Hum Mutat* 2017;**38**:788–97. https://doi.org/10.1002/humu.23230.
44. Chen S, Zhou Y, Chen Y. *et al.* Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 2018;**34**:i884–90. https://doi.org/10.1093/bioinformatics/bty560.
45. Dobin A, Davis CA, Schlesinger F. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;**29**:15–21. https://doi.org/10.1093/bioinformatics/bts635.
46. Pertea M, Pertea GM, Antonescu CM. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* 2015;**33**:290–5. https://doi.org/10.1038/nbt.3122.
47. Thorsson V. *et al.* The immune landscape of cancer. *Immunity* 2018;**48**:812–830.e814. https://doi.org/10.1016/j.immuni.2018.03.023.
48. Wang M, Lei C, Wang J. *et al.* TripHLApan: Predicting HLA molecules binding peptides based on triple coding matrix and transfer learning. *Brief Bioinform* 2024;**25**:bbae154. https://doi.org/10.1093/bib/bbae154.
49. 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 Systems* 2020;**11**:42–48.e47. https://doi.org/10.1016/j.cels.2020.06.010.
50. Nilsson JB, Kaabinejadian S, Yari H. *et al.* Accurate prediction of HLA class II antigen presentation across all loci using tailored data acquisition and refined machine learning. *Sci Adv* **9**:eadj6367. https://doi.org/10.1126/sciadv.adj6367.
51. Liu Z, Jin J, Cui Y. *et al.* DeepSeqPanII: An interpretable recurrent neural network model with attention mechanism for peptide-HLA class II binding prediction. *IEEE/ACM Trans Comput Biol Bioinform* 2022;**19**:2188–96. https://doi.org/10.1109/tcbb.2021.3074927.
52. Habel JR, Nguyen AT, Rowntree LC. *et al.* HLA-A∗11:01-restricted CD8+ T cell immunity against influenza a and influenza B viruses in indigenous and non-indigenous people. *PLoS Pathog* 2022;**18**:e1010337. https://doi.org/10.1371/journal.ppat.1010337.
53. Speir JA, Stevens J, Joly E. *et al.* Two different, highly exposed, bulged structures for an unusually long peptide bound to rat MHC class I RT1-aa. *Immunity* 2001;**14**:81–92. https://doi.org/10.1016/s1074-7613(01)00091-7.
54. Guo HC, Jardetzky TS, Garrettt TPJ. *et al.* Different length peptides bind to HLA-Aw68 similarly at their ends but bulge out in the middle. *Nature* 1992;**360**:364–6. https://doi.org/10.1038/360364a0.
55. Miles JJ, Borg NA, Brennan RM. *et al.* TCR alpha genes direct MHC restriction in the potent human T cell response to a class I-bound viral epitope. *J Immunol* 2006;**177**:6804–14. https://doi.org/10.4049/jimmunol.177.10.6804.
56. Macdonald WA, Purcell AW, Mifsud NA. *et al.* A naturally selected dimorphism within the HLA-B44 supertype alters class I structure, peptide repertoire, and T cell recognition. *J Exp Med* 2003;**198**:679–91. https://doi.org/10.1084/jem.20030066.
57. Kaur G, Gras S, Mobbs JI. *et al.* Structural and regulatory diversity shape HLA-C protein expression levels. *Nat Commun* 2017;**8**:15924. https://doi.org/10.1038/ncomms15924.
58. Jones EY, Fugger L, Strominger JL. *et al.* MHC class II proteins and disease: A structural perspective. *Nat Rev Immunol* 2006;**6**:271–82. https://doi.org/10.1038/nri1805.
59. Kim JY, Cha H, Kim K. *et al.* MHC II immunogenicity shapes the neoepitope landscape in human tumors. *Nat Genet* 2023;**55**:221–31. https://doi.org/10.1038/s41588-022-01273-y.
60. Lu T, Wang S, Xu L. *et al.* Tumor neoantigenicity assessment with CSiN score incorporates clonality and immunogenicity to predict immunotherapy outcomes. *Sci Immunol* 2020;**5**:aaz3199. https://doi.org/10.1126/sciimmunol.aaz3199.
61. Costello JC. *et al.* A community effort to assess and improve drug sensitivity prediction algorithms. *Nat Biotechnol* 2014;**32**:1202–12. https://doi.org/10.1038/nbt.2877.
62. Ayers M, Lunceford J, Nebozhyn M. *et al.* IFN-$\gamma$-related mRNA profile predicts clinical response to PD-1 blockade. *J Clin Invest* 2017;**127**:2930–40. https://doi.org/10.1172/jci91190.
63. Auslander N, Zhang G, Lee JS. *et al.* Robust prediction of response to immune checkpoint blockade therapy in metastatic melanoma. *Nat Med* 2018;**24**:1545–9. https://doi.org/10.1038/s41591-018-0157-9.
64. Dudley JC, Lin M-T, Le DT. *et al.* Microsatellite instability as a biomarker for PD-1 blockade. *Clin Cancer Res* 2016;**22**:813–20. https://doi.org/10.1158/1078-0432.CCR-15-1678.
65. Brauer MJ, Zhuang G, Schmidt M. *et al.* Identification and analysis of in vivo VEGF downstream markers link VEGF pathway activity with efficacy of anti-VEGF therapies. *Clin Cancer Res* 2013;**19**:3681–92. https://doi.org/10.1158/1078-0432.Ccr-12-3635.
66. Powles T, Nickles D, van Allen E. *et al.* Immune biomarkers associated with clinical benefit from atezolizumab (MPDL3280a; anti-PD-L1) in advanced urothelial bladder cancer (UBC). *J Immunother Cancer* 2015;**3**:P83. https://doi.org/10.1186/2051-1426-3-S2-P83.
67. Fridman WH, Zitvogel L, Sautès–Fridman C. *et al.* The immune contexture in cancer prognosis and treatment. *Nat Rev Clin Oncol* 2017;**14**:717–34. https://doi.org/10.1038/nrclinonc.2017.101.
68. Zheng L, Qin S, Si W. *et al.* Pan-cancer single-cell landscape of tumor-infiltrating T cells. *Science* 2021;**374**:abe6474. https://doi.org/10.1126/science.abe6474.



*Briefings in Bioinformatics*, 2025, **26(1)**, bbae595
https://doi.org/10.1093/bib/bbae595
Problem Solving Protocol