Supplementary Materials for

ACME: Pan-specific peptide-MHC class I binding prediction through attention-based deep neural networks

Yan Hu[1], Ziqiang Wang[2], Hailin Hu[3], Fangping Wan[4], Lin Chen[5], Yuanpeng Xiong[6, 7], Xiaoxia Wang[2], Dan Zhao[4], Weiren Huang[2, *] and Jianyang Zeng[4, 8, *]

1 Supplementary Notes

1.1 Supplementary details of the datasets used in the study

Several experimentally measured peptide-MHC binding affinity datasets were used in this study for model training, parameter tuning and model testing. The widely used IEDB MHC class I binding affinity dataset (`http://tools.immuneepitope.org/mhci/download/`) [1, 2] contains 186,684 binding affinity measurements between peptides and MHC class I molecules for different MHC alleles and species. Peptides sequences without corresponding MHC sequence information were excluded from our study. To focus on the prediction of MHC-peptide binding in humans, here we only kept the data of all the HLA-A and HLA-B alleles with at least 100 samples. The remaining dataset after these filtering steps contained 152,197 binding affinity measurements covering 61 alleles.

The dataset from [3] was combined with the aforementioned IEDB dataset to train a final version of our model. We used the same preprocessing criteria as described above to filter the combined dataset.

Many records of MHC sequences in the IPD-IMGT/HLA database (`ftp://ftp.ebi.ac.uk/pub/databases/ipd/imgt/hla/hla_prot.fasta`) [4] lack certain parts of the complete sequences, like the signal peptides, which makes it difficult to index each residue. Therefore, these MHC sequences were also excluded from our study to ensure that all the remaining MHC sequences were complete and can be aligned well with each other. Each of these MHC sequences contained a 24-residue signal peptide at the N-terminal, which was not used as part of the input to our model. In other words, the MHC sequences were indexed after removing the signal peptides.

1.2 Model structure

Our model takes both the encoded peptide and MHC sequence as input and predicts their binding affinity. The inputs to the model first pass through a convolutional layer for initial feature

[1] School of Life Sciences, Tsinghua University, Beijing, China.
[2] Department of Urology, Shenzhen Second People's Hospital, The First Affiliated Hospital of Shenzhen University, International Cancer Center, Shenzhen University School of Medicine, Shenzhen 518039, China.
[3] School of Medicine, Tsinghua University, Beijing, China.
[4] Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China.
[5] Turing AI Institute of Nanjing, Nanjing, China.
[6] Department of Computer Science and Technology, Tsinghua University, Beijing, China.
[7] Bioinfomatics Division, Beijing National Research Center for Information Science and Technology.
[8] MOE Key Laboratory of Bioinformatics, Tsinghua University, Beijing 100084, China.
[*] To whom correspondence should be addressed. Email: `zengjy321@tsinghua.edu.cn` and `pony8980@163.com`

extraction (Supplementary Figure 1). Then the extracted feature representations are forwarded into the convolutional and attention modules for further processing. The outputs of the two modules are then concatenated and then pass through a fully connected layer to produce the final output.

The peptide sequence is encoded into a 24×20 matrix while the MHC pseudo-sequence is encoded into a 34×20 matrix. Each column represents an encoded feature dimension, and each row stands for a position in the sequence. For initial feature extraction, each of these two matrices first goes through a one-dimensional convolutional layer (Supplementary Figure 2). For both peptide and MHC features, we use 128 convolutional filters of size 3×20 with stride size 1 to scan along the input sequence. ReLU is chosen as the activation function after each convolutional layer.

Then the extracted features are forwarded into both convolutional and attention modules separately. In the convolutional module (Supplementary Figure 3), the features extracted from the MHC matrix go through a round of max-pooling operation, followed by another round of convolution and max-pooling operations. For all the convolutional layers, 128 filters of width 3 are used. Note that the max-pooling operations used in our work are all one-dimensional max-pooling where down-sampling is performed along the sequence, and both window length and stride size are set to 2 for max-pooling. Next, we integrate these MHC features of different abstraction levels with peptide features. To this end, the MHC features after 0, 1 and 2 rounds of convolution and max-pooling operations are concatenated with three identical copies of peptide features after one round of convolution, respectively. Each of the three concatenated vectors goes through a separate fully connected layer, and then the corresponding 256-dimensional output vectors are concatenated and then forwarded to two consecutive fully connected layers (with 64 and 2 output dimensions, respectively) to generate the output of the convolutional layer (Supplementary Figure 3**c**).

In the attention module (Supplementary Figure 4 and Supplementary Figure 5), each column vector $\boldsymbol{F}_j$ from the initially extracted feature map first goes through an attention layer and is then converted into a real number w_j. Next, each w_j is converted into w'_j using softmax normalization. After that, the model computes the weighted average $\boldsymbol{F}_{avg}$ over all the column vectors, which is then used as the final output of the attention module.

In the end, the outputs of the convolution and attention modules are concatenated and forwarded into a fully connected layer with a sigmoid activation function, generating the final output of the model.

1.3 Model training

During model training and hyperparameter tuning, The performance was evaluated mainly using the Pearson correlation coefficient (PCC), while the area under the receiver operating characteristics curve (AUROC) was also used as an additional metric to evaluate the prediction results. For the calculation of AUROC, the experimentally determined IC_{50} values were converted into one or zero labels using a threshold of 500 nM. The hyper-parameters in ACME include the number and the size of convolutional filters, stride size for convolution, padding strategy (i.e., valid padding or same padding in convolution), pooling size and stride, choice of activation function, and the sizes of fully connected layers. Due to the large number of hyperparameter settings, it was difficult to perform grid search to fine tune all of them. Therefore, we only used grid search to optimize the number of convolutional filters as well as the size of each fully connected layer, while setting other hyperparameters to default values in Keras [5] or empirical values according to our previous experience on training similar networks. The search grid for the filter number was $\{32, 64, 128\}$, and the grids for the fully connected layers were: $\mathrm{Fc}_{1,1}$, $\mathrm{Fc}_{1,2}$, $\mathrm{Fc}_{1,3}$:$\{16, 64, 256\}$, Fc_2:$\{1, 16, 64\}$, and Fc_3:$\{1, 2, 4\}$).

Each model consists of an ensemble of n deep neural networks, where $n = 5$ in five-fold cross-validation and $n = 25$ in the final version of our model. The average prediction score over all the networks in the ensemble is used as the final output. We use the mean squared error as the loss

function. In addition, the Adam optimizer [6] is used to compute stochastic gradient descent in the training procedure.

1.4 Five-fold cross-validation

We randomly split all the binding affinity data of each MHC allele into five subsets of roughly equal sizes. Then we used one subset from each allele as a validation set and merged all the other data into the training set, which was then used to train an ensemble of five deep neural networks. Afterward, the ensemble of the five networks was tested on the validation set. The average result over the five folds was reported as the final output. We compared the Pearson correlation coefficient (PCC) and the area under the receiver operating characteristics curve (AUROC) values of our model with those of NetMHCpan 3.0 [7] for different alleles and peptides of different lengths in this five-fold cross-validation procedure. Those alleles with less than 10 samples for validation were excluded from the study because the performance on such small datasets can be easily affected by random noise in the data. In sporadic cases, for an allele with scarce data, we could generate a test set with no positive or negative sample, leading to an invalid result when we calculate AUROC. Therefore, we also excluded those alleles with fewer than three valid results to make sure that all the performances reported in this article were reliable.

In addition, we also carried out a series of additional ablation tests to study the contribution of the convolutional and attention modules in affinity prediction. Five different structures were tested: (1) the original ACME model; (2) a model with only the convolutional module (Supplementary Figure 6a); (3) a model with only the attention module (Supplementary Figure 6b); (4) a model with only the attention module, whose hyperparameters were optimized (the number of filters was set to 256); and (5) a model with only the attention module and an additional fully connected layer (64 nodes) (Supplementary Figure 6c). The structures of these models are visualized in Supplementary Figure 6, and the results are summarized in Supplementary Table 1.

The results demonstrated that the performance of the convolution-only model was comparable to ACME, while adding an attention module slightly improved the performance on 11-mers. This indicated that the convolutional module plays a primary role in prediction while the attention module plays a secondary role. Besides, the attention-only model, with original or optimized hyperparameters, had lower predictive power, but adding an extra fully connected layer greatly improved its performance. This meant that the output of the attention module possesses sufficient information for peptide-MHC binding prediction, as long as it is integrated into a suitable network architecture. After testing four different ways to integrate attention (all the above except the convolution-only structure), ACME still had the best performance. There are two possible explanations for this. First, the prediction performance may have already reached saturation on our current dataset, or alternatively, there could exist a better to integrate attention into the framework that is yet to be discovered. It needs to be emphasized that prediction performance is not the only objective that we pursue, and we also care about the contribution of the attention module in model interpretability. However, how to improve its contribution to affinity prediction still needs further exploration, and future studies are needed to find better ways to integrate the attention module into the model.

1.5 Testing the generalizability of ACME

To prove the generalization ability of our model, we also chose the most recent IEDB weekly datasets as external test sets to further evaluate our model. Before this test, we augmented the original IEDB training data by also incorporating another dataset from [3] and applied a bootstrapping-like strategy to train a more reliable version of our model. More specifically, the whole combined dataset was randomly divided into five subsets of approximately the same size, and then every four subsets were used to train a separate deep neural network, which resulted

in five trained models in total. We performed such a process with random partitions of the data five times, and overall we obtained an ensemble of 25 trained deep neural networks, whose average prediction result was reported as the final output. We then evaluated this final version of our model on 30 IEDB weekly benchmark datasets of human MHC molecules, which are encoded by human leukocyte antigen type A and B genes (i.e., HLA-A and HLA-B). The prediction performances of other state-of-the-art methods on these datasets, evaluated using the Spearman rank correlation coefficient (SRCC), are available on the IEDB weekly benchmarking website (`http://tools.immuneepitope.org/auto_bench/mhci/weekly/`) [8]. To compare our prediction results to those of other state-of-the-art methods, we also used SRCC to evaluate the performance of ACME. The results are shown in Figure 3a. In addition, to better demonstrate the performance difference between ACME and previous methods, as well as the differences among previous methods, we also show the results in Supplementary Figure 7, where the y-axis ranges from 0.40 to 0.60.

We further tested whether our model can generalize what it has learned to make accurate predictions for those alleles without any training data. Here, for each allele of interest, we removed all the associated binding data from the training set and then used the remaining data to train our model. In particular, data of all other alleles were split into five subsets of approximately equal sizes. We then trained one deep neural network on every four subsets, resulting in an ensemble of five networks whose average prediction result was used as the final output. After that, we used this ensemble to make binding predictions for the allele of interest.

1.6 Making predictions for MHC alleles without training data

As introduced in the previous section, we tested the ability of our model to make predictions for those MHC alleles without training data. Our test results suggested that in most cases, ACME can make accurate predictions for an allele without any training data. More specifically, for HLA-A alleles, ACME achieved an average PCC of 0.77 and an average AUROC of 0.90. For HLA-B, ACME achieved an average PCC of 0.68 and an average AUROC of 0.88.

Next, we set out to investigate why the model achieved different performances on individual alleles. Although removing the data associated with individual alleles of interest resulted in training sets of different sizes, there was only a weak correlation between the number of samples removed and the prediction performance (Spearman rank correlation coefficient $\rho = 0.23$), which did not explain the performance differences since the correlation coefficient was positive. Thus, most likely it was not the total number, but the effective number of remaining training samples that was correlated to the performance. Therefore, we defined the following reference information (RI) term to measure how much information ACME can gain from the training dataset when making predictions for a new allele,

$$RI_i = \sum_{j=1, j\neq i}^{m} \frac{n_j}{D(seq_j, seq_i)^2 + 1},$$

where RI_i stands for the reference information that the model can learn from all other alleles when making prediction for the i^{th} allele, $D(seq_j, seq_i)$ stands for the editing distance between the encoded sequences of the j^{th} and i^{th} MHC alleles, m stands for the total number of alleles, and n_j stands for the total number of training samples of the j^{th} allele. We observed a significant positive correlation between RI and the prediction performance (Supplementary Figure 8a, Spearman rank correlation coefficient $\rho = 0.53$). We suspected that the performance difference on HLA-A and HLA-B alleles might be associated with the RI difference between these two types of alleles. Hence, we further compared the RI levels of HLA-A and HLA-B alleles, and the results are shown in Supplementary Figure 8b. We found that the RI values of HLA-A alleles were significantly higher than those of HLA-B alleles ($p = 2.5 \times 10^{-9}$, one-sided Mann-Whitney U test), which suggested

that the lower performance of ACME on HLA-B was probably associated with the corresponding lower RI levels.

Given that RI and prediction performance are positively correlated, when we want to make prediction for a new allele, we can use RI to help decide whether or not the prediction results are reliable. Calculating the RI distribution over the space of MHC sequence can also help detect the most information-deficient regions. Collecting more experimental data for these regions will be beneficial for improving the prediction accuracy for those data-deficient alleles.

1.7 Experimental validation

We experimentally validated the ability of ACME to identify neo-antigens derived from somatic mutations in tumors. Here, HLA-A*02:01 and HLA-B*27:05 were chosen as representative alleles with high and relatively low population frequencies, respectively. We selected the most frequent 5000 somatic missense mutations in human cancer from the COSMIC database [9] and then generated the sequences of all the 9-mer peptides that may potentially present these mutations (the human proteome was obtained from the UniProt website [10]). For each allele, we used ACME to predict the binding affinities of these peptides and selected the top 25 peptides with the highest predicted binding affinities and bottom 25 peptides with the lowest predicted binding affinities for the downstream experimental validation.

We next experimentally measured the actual binding affinities for the above chosen peptides. For HLA-A*02:01, T2 cells (5×10^5 mL^{-1}) were incubated with β_2m ($3\,\mu g\,mL^{-1}$, Sigma) and each peptide ($10\,\mu g\,mL^{-1}$, Sangon Biotech) at 4 °C for 4 h. The cells were stained with mouse anti-HLA-A2 antibody (abcam, ab79523) to measure the expression of HLA-A2 by flow cytometry. The mean fluorescence intensity (MFI) of HLA-A2 was used to indicate the binding affinity of each peptide with the HLA-A2 molecule. T2 cells with β_2m only were used as a negative control. T2 cells with β_2m and peptide OVAL235, which had been previously reported to bind strongly to the HLA-A2 molecule [11], were used as a positive control.

For HLA-B*27:05, we used the ProImmune REVEAL peptide-MHC binding assay to determine the binding affinities of the chosen peptides. This assay evaluates the binding affinity of a peptide to an MHC molecule by assessing its ability to stabilize the MHC complex. When the complex is bound and stabilized by the peptide, the MHC protein and β_2m form a native conformation that can be recognized by a labeled antibody and thus generate a positive signal. The percentage of this signal relative to the reference signal generated by a known T cell epitope (positive control) is defined as the ProImmune REVEAL binding score and used to measure the binding affinity of this peptide.

In addition, we also adopted an alternative approach to study cases where different models predict very differently. In particular, we first downloaded the experimental binding dataset by Khan et al. [12], which contained the peptide sequences derived from the HSV-1 genome and the corresponding experimentally measured binding affinities with HLA-A*02:01. We then compared the corresponding binding prediction results made by ACME and NetMHCpan 3.0. Among all the 434 9-mers in the dataset, 171 showed more than 5 percentage points difference in the affinities predicted by ACME and NetMHCpan 3.0. On 149 out of the 171 peptides, ACME made more accurate predictions compared to NetMHCpan 3.0. These 149 peptides had relatively lower measured binding affinities ($p = 0.021$, one-sided Mann-Whitney U test) compared to the other 22 peptides that NetMHCpan 3.0 predicted more accurately, as shown in Supplementary Figure 9. The results indicated that ACME achieved better performance than NetMHCpan 3.0 in discriminating peptides with medium or low binding affinities (at least for HLA-A*02:01). In addition, it is worth noting that even for peptides with relatively high measured binding affinities (≥ 0.8), ACME was still more accurate in most cases (24 out of 30), suggesting that ACME was more accurate within a wide range of binding affinities. Moreover, there were 47 peptides that showed more than 10

percentage point difference in the affinities predicted by ACME and NetMHCpan 3.0, and ACME made more accurate predictions on 45 of them. The information of these 45 peptides is listed in Supplementary Table 2. To briefly summarize these validation results, at least for the most common human MHC allele, HLA-A*02:01, ACME generally achieved more accurate predictions, when its prediction scores were significantly different from those of NetMHCpan 3.0.

1.8 Generating the binding motifs of different MHC alleles

To better illustrate the underlying biological principles captured by our model, we also used ACME to generate binding motifs for different alleles. In particular, we hypothesized that if a certain amino acid type at a specific site contributes significantly to binding, it should appear more often among strong binders and also be associated with a higher attention score. We use a matrix $\boldsymbol{A}$ to represent such an enrichment of attention, where each element in the i^{th} row and j^{th} column corresponds to the summed attention score of the i^{th} amino acid type at the j^{th} position in the selected peptides, that is,

$$A_{ij} = \sum_{k=1}^{n} w_{kj}\delta(r_{kj}, i),$$

where n represents the total number of strong binders and w'_{kj} stands for the attention score assigned to the j^{th} position of the k^{th} peptide (which is also defined in Section 2.3.3). Here, r_{kj} stands for the amino acid type at the j^{th} position of the k^{th} peptide. For example, r_{kj} is 0 for the first amino acid type (alanine) and 19 for the last amino acid type (valine). Besides, $\delta(x, y)$ stands for a binary indicator function defined by

$$\delta(x, y) = \begin{cases} 1, & x = y; \\ 0, & x \neq y. \end{cases}$$

For each allele, we selected the top 1000 peptides with the highest predicted binding affinities ($n = 1000$) from 10,000 random ones and then generated the corresponding matrix $\boldsymbol{A}$ to identify important residues that significantly contribute to peptide-MHC binding. To further investigate the residues with the potential to destabilize peptide-MHC complexes, we also selected the bottom 1000 peptides with the lowest predicted binding affinities from 10,000 random ones and then generated their corresponding motifs for downstream analyses.

1.9 Validating the binding motifs of different MHC alleles

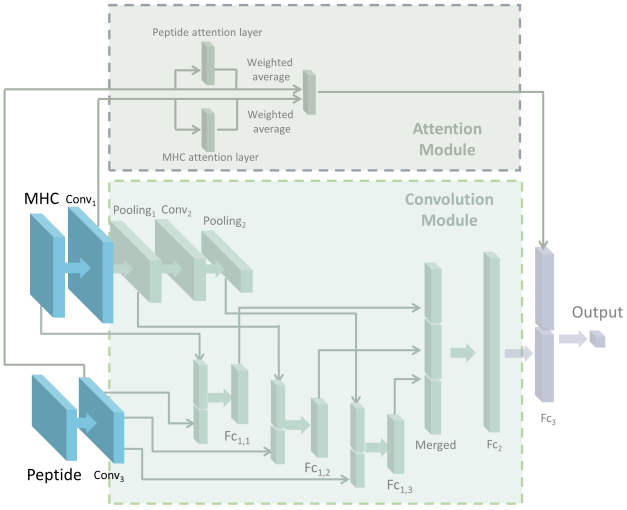
Before we combine the attention module with the convolutional neural network to extract explainable binding patterns of different MHC alleles, it is necessary to validate the ability of our attention module to discriminate important positions in the peptides from the background. We hypothesized that if the attention module can indeed accurately evaluate the importance of each residue, then masking the information at those positions of higher attention scores should affect the predicted binding affinities more dramatically.

To validate this hypothesis, we first randomly sampled 10,000 peptides from the human proteome [10] for each MHC allele. We then predicted the binding affinities of these random peptides and selected the top 1000 peptides with the highest predicted binding affinities. Suppose that the original predicted binding affinity of a peptide is denoted by A^{ori}. We then used zero to mask the residue of the highest attention score in the peptide and obtained the predicted binding affinity of the modified peptide, denoted by A^{hm}. Likewise, we also masked the position of the lowest attention score and obtained another predicted binding affinity, denoted by A^{lm}. Therefore, $D_1 = |A^{ori} - A^{hm}|$ and $D_2 = |A^{ori} - A^{lm}|$ reflect the influence of the highest and the lowest-attention positions on peptide-MHC binding, respectively (Supplementary Figure 10**a**). If the D_1 values are

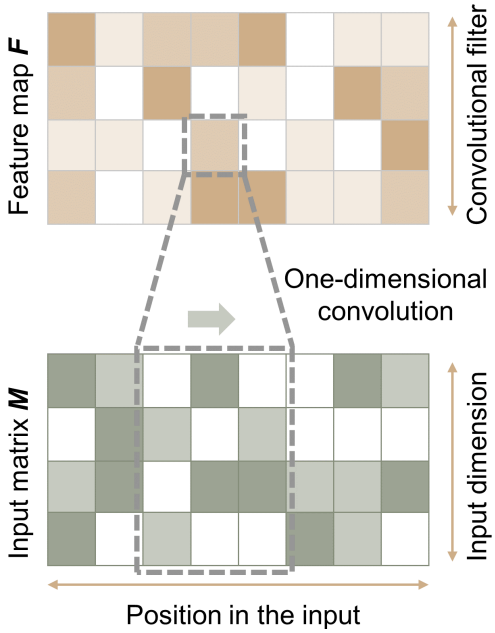
significantly higher than the D_2 values, it means that the attention module employed in ACME can indeed detect the important residues.

Our comparison showed that masking the positions with high attention scores led to a significantly larger change in the predicted binding affinities than the other case which masked the low-attention positions ($p < 10^{-300}$, one-sided Mann-Whitney U test, Supplementary Figure 10**b**). This result supported our hypothesis that our employed attention module can discriminate those important residues from the background.

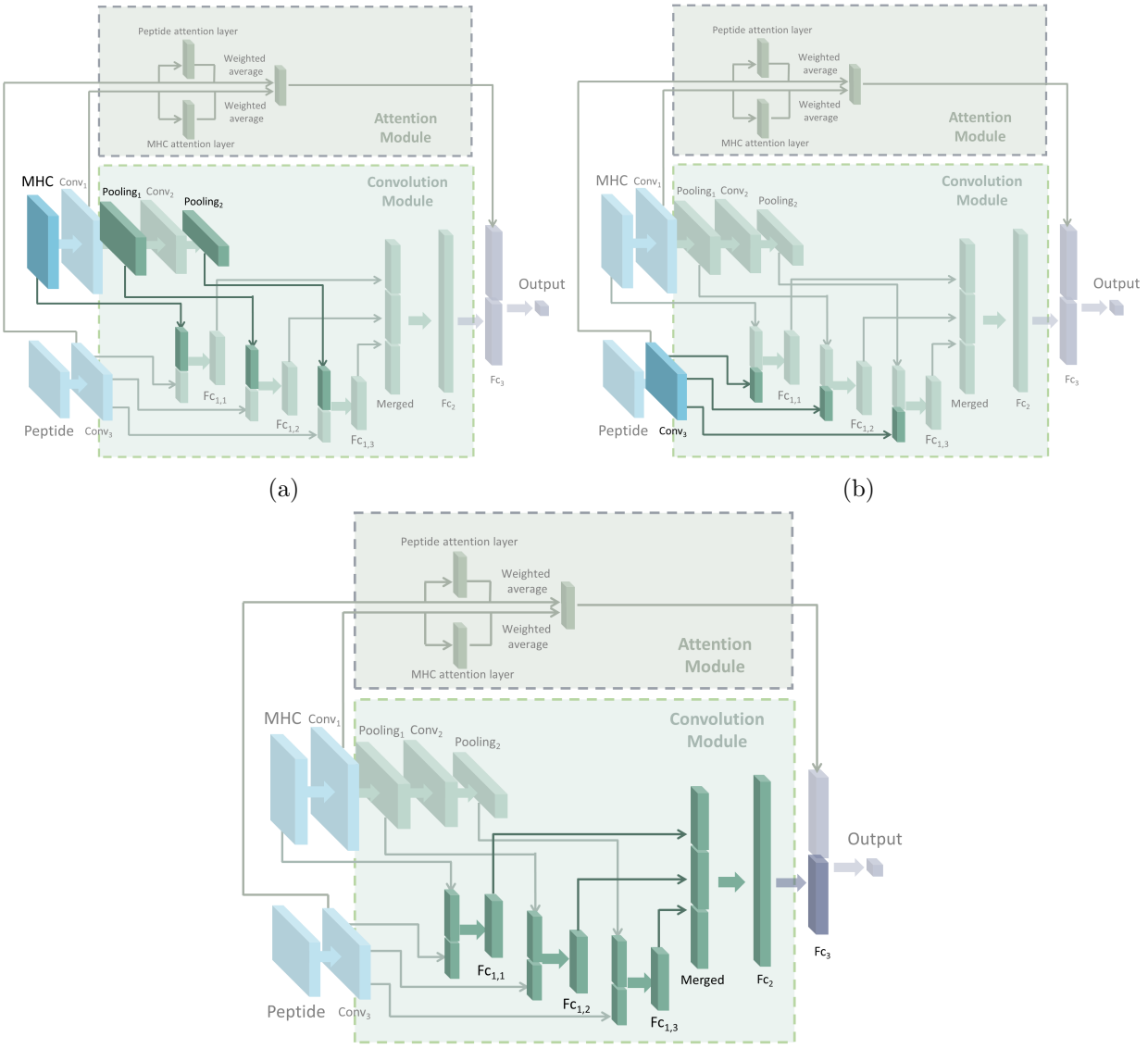
2 Supplementary Figures



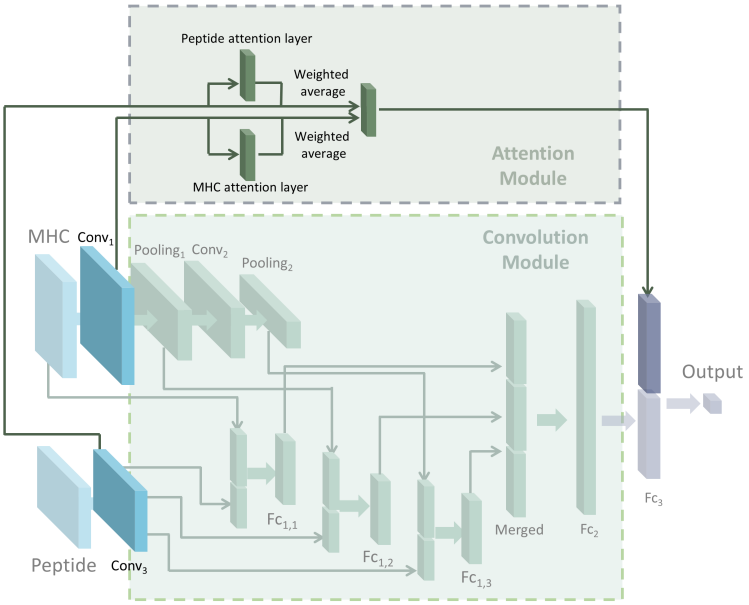
Supplementary Figure 1: Initial feature extraction in the ACME framework.



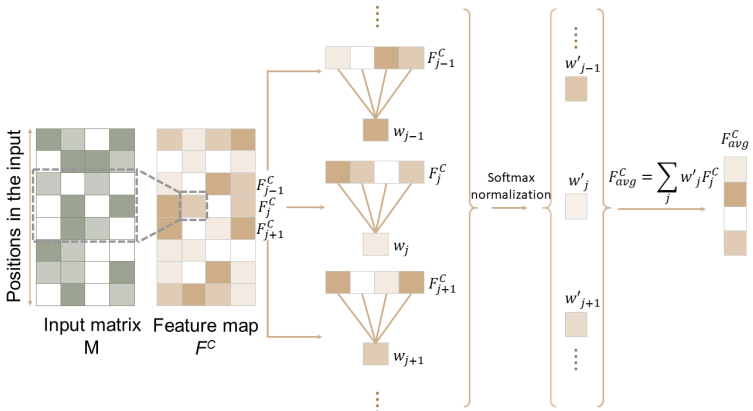
Supplementary Figure 2: One-dimensional convolution operation in the ACME framework.



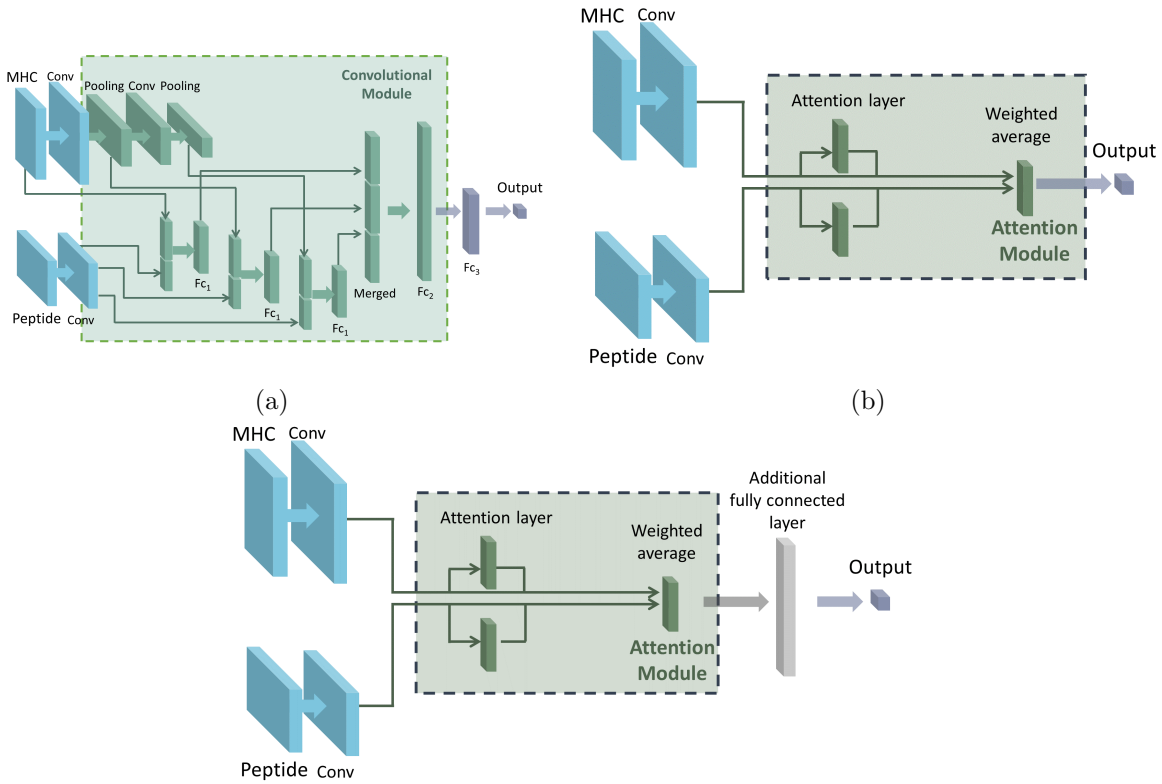
Supplementary Figure 3: The convolutional module in the ACME framework. **a** The MHC features of different abstraction levels. **b** The three identical copies of the features extracted from the input peptide sequence. **c** Integrating the MHC features of different abstraction levels with the features extracted from the peptide sequence.



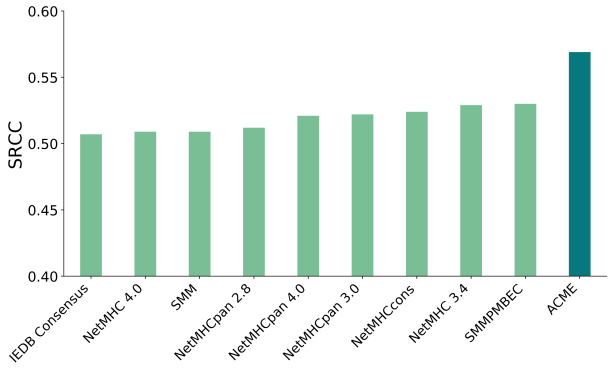
Supplementary Figure 4: The attention module in the ACME framework. This module computes the weighted average of the feature vectors corresponding to each input position. It learns to assign higher weights to those positions that are more crucial in peptide-MHC binding to make more accurate predictions.



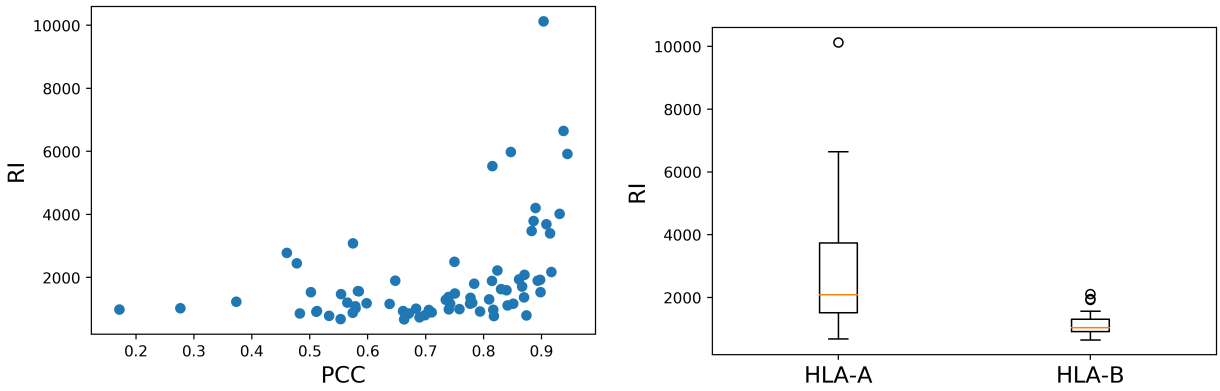
Supplementary Figure 5: Schematic illustration of the attention module. The first convolutional layer extracts a feature map $\boldsymbol{F}$ from the input matrix, in which each row vector $\boldsymbol{F}_j$ corresponds to a position in the input sequence. $\boldsymbol{F}_j$ is then used as the input to the attention layer to generate its positional weight w_j, which becomes w'_j after softmax normalization. $\boldsymbol{F}_{avg}$ stands for the final weighted average over all row vectors in $\boldsymbol{F}$. More details of the attention module can be found in Section 2.3.



Supplementary Figure 6: Model structures in the ablation test. (a) The model with only the convolutional module. (b) The model with only the attention module (either with original or optimized hyperparameters). (c) The model with only the attention module and an extra fully connected layer.

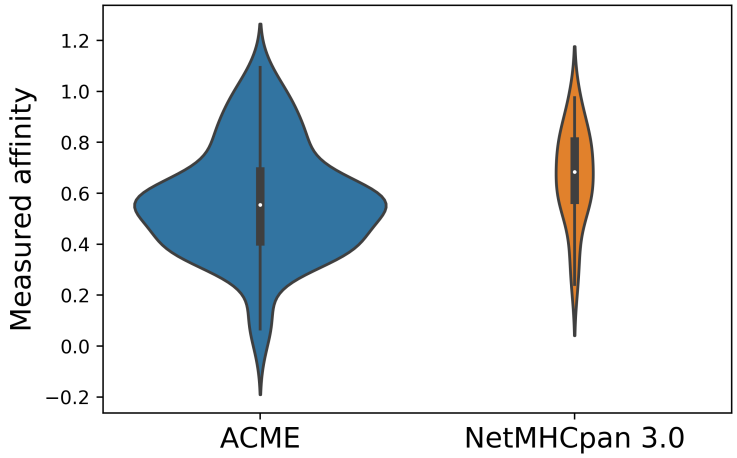


Supplementary Figure 7: The performances of different prediction algorithms on the most recent IEDB benchmark datasets, measured in terms of the Spearman rank correlation coefficient (SRCC) between measured and predicted affinities. The SRCC scores of other prediction methods were obtained from [8]. Each bar represents the averaged performance of a specific prediction algorithm over all the datasets tested.

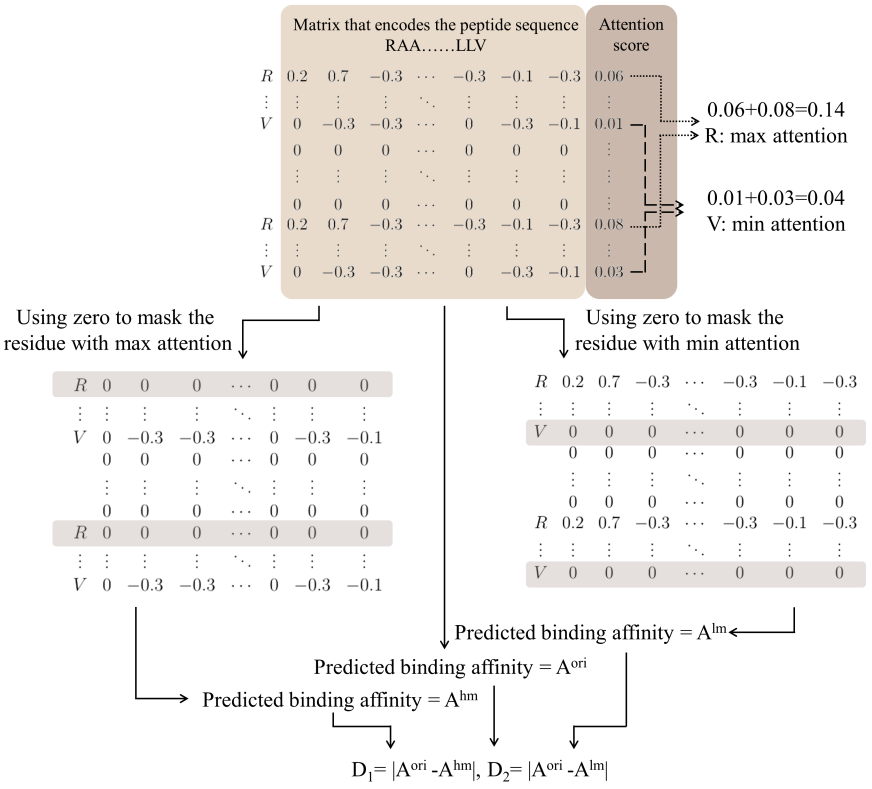


(a) (b)

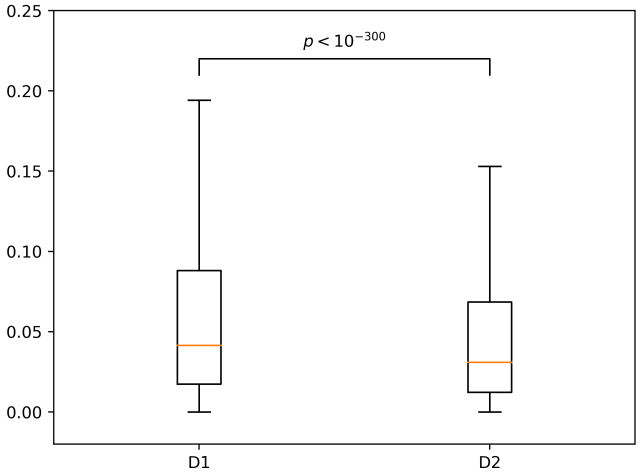
Supplementary Figure 8: Investigating the factors that affect the prediction performance of ACME for alleles without training data. (a) There existed a significant positive correlation between RI and model performance on novel alleles that were not seen in training data (Spearman rank correlation coefficient $\rho = 0.53$). (b) The RI values of HLA-A alleles were significantly higher than those of HLA-B alleles ($p = 2.5 \times 10^{-9}$, one-sided Mann-Whitney U test).



Supplementary Figure 9: The distribution of the experimentally measured binding affinities of the peptides with distinct predicted affinities by ACME and NetMHCpan 3.0. All the peptides with > 5 percentage point difference were selected. The peptide sequences were derived from the HSV-1 genome and the corresponding experimentally measured binding affinities were obtained from [12]. There are in total 171 peptides with > 5 percentage point prediction differences. ACME made more accurate predictions for 149 of them, and their experimentally measured affinities are shown on the left. NetMHCpan 3.0 was more accurate for 22 peptides, whose experimentally measured affinities are shown on the right. The widths of the violins represent the numbers of corresponding peptides. The former group (i.e., where ACME achieved more accurate predictions) of peptides had relatively lower measured binding affinities ($p = 0.021$, one-sided Mann-Whitney U test). Even for the peptides with high measured binding affinities (≥ 0.8), ACME was still more accurate in most cases (24 out of 30), suggesting that ACME achieved more accurate predictions within a wide range of binding affinities.

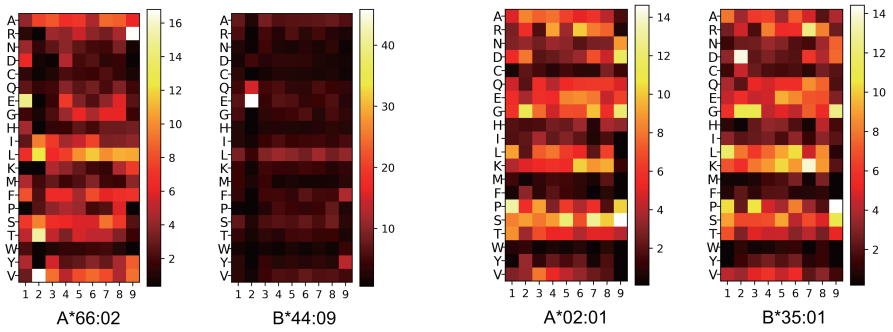


(a)

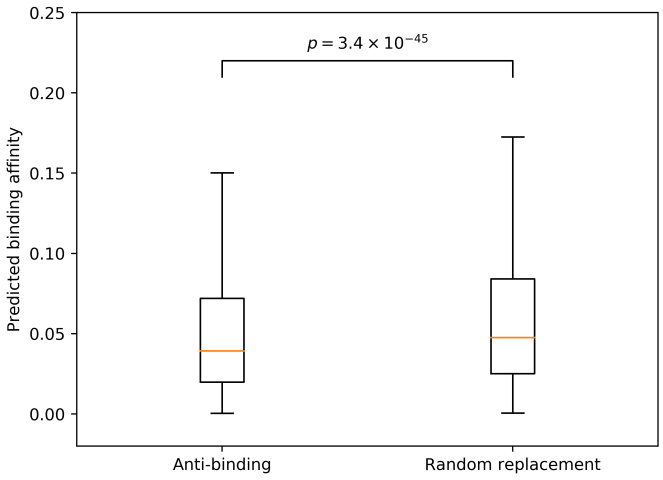


(b)

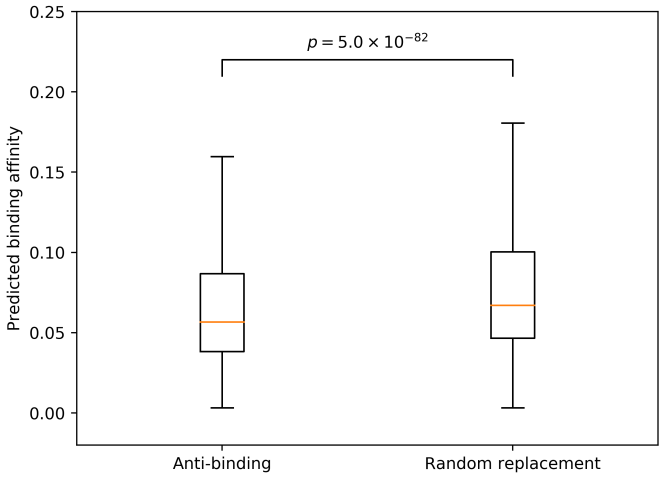
Supplementary Figure 10: A masking test for validating the quality of the attention values assigned to individual peptide positions. **a** Schematic illustration of the masking test to evaluate the ability of our attention module to detect important residues for MHC-peptide binding. The residues with the highest and the lowest attention values were masked, respectively, and then the corresponding changes of the predicted binding affinities (denoted by D_1 and D_2, respectively) were calculated. **b** Results of the masking test. D_1 is significantly higher than D_2. The p value was calculated using the one-sided Mann-Whitney U test.



Supplementary Figure 11: Examples of sequence motifs for the peptides bound to different MHC alleles. **a** The heatmaps of two sequence motifs generated by ACME for the rare MHC alleles not included in training data. **b** The heatmaps of two sequence motifs for the peptides with low predicted binding affinities.



(a)



(b)

Supplementary Figure 12: Replacing the original residue with an anti-binding residue (see Section 3.4 for more details) resulted in significantly lower predicted binding affinities compared to the replacement by a random residue. **a** Results for HLA-A*02:01. **b** Results for HLA-B*35:01, the p values were calculated using the one-sided Mann-Whitney U test.

Supplementary Table 1: Five-fold cross-validation performance of different model structures in the ablation tests. Six models were tested, including (1) the original ACME model, (2) a model with only the convolutional module, (3) a model with only the attention module, (4) a model with only the attention module, whose hyperparameters were optimized (the number of filters was set to 256), and (5) a model with only the attention module and an additional fully connected layer (64 nodes). Model performance was measured in Pearson Correlation Coefficient (PCC).

Model structure	9-mer	10-mer	11-mer
ACME	0.844	0.809	0.802
Convolution only	0.845	0.808	0.787
Attention only	0.273	0.345	0.282
Attention only (optimized)	0.344	0.399	0.283
Attention with fc	0.809	0.782	0.793

Supplementary Table 2: Cases with more than 10 percentage point difference in the predicted affinities by ACME and NetMHCpan 3.0. The experimentally measured binding affinities were reported in [12]. ACME made more accurate predictions for 45 out of the 47 peptides.

Peptide sequence	ACME prediction score	NetMHCpan3.0 prediction score	Measured affinity
ALDHYDCLI	0.541	0.658	0.520
ALDQACFRI	0.647	0.546	0.590
ALMRGRPGL	0.399	0.534	0.314
ALQTDNYTL	0.700	0.595	0.689
AVGELLAPV	0.678	0.801	0.590
AVLADFSLV	0.555	0.686	0.547
AVVPIIPFL	0.532	0.640	0.359
DLAEWVPRV	0.739	0.586	0.865
FADTVVACV	0.683	0.544	0.623
FAFRYVNRL	0.518	0.398	0.556
FLHLYLFLT	0.500	0.630	0.414
FLLKQFHAA	0.702	0.857	0.633
FLLSGTAIA	0.878	0.763	0.936
FLRSCHWVL	0.497	0.645	0.412
FLSRPINTI	0.562	0.697	0.421
FLYLAFVAL	0.496	0.652	0.299
FMAAKAAHL	0.650	0.800	0.481
FVADVQHAA	0.425	0.612	0.172
FVGVIILGV	0.747	0.622	0.836
FVYTPSPYV	0.677	0.801	0.551
GMHPRGVHA	0.118	0.236	-0.031
HMFCDPMCA	0.444	0.568	0.383
IAPNASLGV	0.391	0.208	0.566
IILTLIPRL	0.673	0.572	0.552
ILVVSLLLV	0.525	0.649	0.508
KTWFLVPLI	0.757	0.627	0.951
LITNYLPSV	0.836	0.724	0.848
LLLALRHPA	0.398	0.526	0.409
LLLISMYAL	0.818	0.707	0.857
LLLRQWLHV	0.564	0.692	0.430
LVVTAIVYV	0.507	0.648	0.517
MLLATREYV	0.670	0.785	0.646
MLWTTDKHV	0.469	0.632	0.338
RLSPFPALV	0.708	0.813	0.622
RLYRWQPDL	0.458	0.588	0.370
RMGELTAEI	0.572	0.682	0.583
RVYNIIQLV	0.476	0.641	0.086
SLQQELAHM	0.528	0.405	0.606
SMDDDTYVA	0.488	0.616	0.417
SVYALGFGV	0.589	0.765	0.338
VIACLLVAV	0.588	0.738	0.508
VLAAGVLVV	0.596	0.699	0.496
VLDCVVTGA	0.419	0.532	0.364
VLGCDAALV	0.489	0.617	0.422
VLWSLLWLG	0.765	0.444	0.969
WLETELVFV	0.630	0.735	0.589
YLSRTQRLA	0.429	0.285	0.566

3 Supplementary Data

Supplementary Data 1. Supplementary results on the performance comparison between ACME and NetMHCpan 3.0 through a five-fold cross-validation.

Supplementary Data 2. Detailed prediction results of ACME and other previous methods on the IEDB weekly benchmark datasets.

Supplementary Data 3. Detailed prediction results of ACME on those alleles that were not seen in training data.

Supplementary Data 4. Detailed information of the peptides used in our experimental validation, including peptide sequences, predicted and experimentally measured binding affinities, and the corresponding mutated proteins that generated the peptides.

Supplementary File 1. The binding sequence motifs generated by ACME for all alleles in the training dataset.

The above files can also be found in `https://github.com/HYsxe/ACME`.

References

1. Vita, R. *et al.* The immune epitope database (IEDB) 3.0. *Nucleic Acids Research* **43,** D405–D412 (2015).
2. Kim, Y. *et al.* Dataset size and composition impact the reliability of performance benchmarks for peptide-MHC binding predictions. *BMC Bioinformatics* **15,** 241–241 (2014).
3. Pearson, H. *et al.* MHC class I-associated peptides derive from selective regions of the human genome. *Journal of Clinical Investigation* **126,** 4690–4701 (2016).
4. Robinson, J. *et al.* The IPD and IMGT/HLA database: allele variant databases. **43** (2014).
5. Chollet, F. *et al. Keras* `https://keras.io`. 2015.
6. Kingma, D. & Ba, J. Adam: A Method for Stochastic Optimization (2014).
7. Nielsen, M. & Andreatta, M. NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. *Genome Medicine* **8,** 33 (2016).
8. Trolle, T. *et al.* Automated benchmarking of peptide-MHC class I binding predictions. *Bioinformatics* **31,** 2174–2181 (2015).
9. Forbes, S. A. *et al.* COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Research* **45,** D777–D783 (2017).
10. Apweiler, R. *et al.* UniProt: the Universal Protein knowledgebase. *Nucleic Acids Research* **32,** D115–D119 (2004).
11. Jin, S. *et al.* Humoral immune responses against tumor-associated antigen OVA66 originally defined by serological analysis of recombinant cDNA expression libraries and its potentiality in cellular immunity. *Cancer Science* **99,** 1670–1678.
12. Khan, A. A. *et al.* Bolstering the Number and Function of HSV-1–Specific CD8+ Effector Memory T Cells and Tissue-Resident Memory T Cells in Latently Infected Trigeminal Ganglia Reduces Recurrent Ocular Herpes Infection and Disease. *The Journal of Immunology* **199,** 186–203. ISSN: 0022-1767 (2017).