Briefings in Bioinformatics, 22(5), 2021, 1–16

https://doi.org/10.1s093/bib/bbaa415
Problem Solving Protocol

Anthem: a user customised tool for fast and accurate prediction of binding between peptides and HLA class I molecules

Shutao Mei[†], Fuyi Li[†], Dongxu Xiang, Rochelle Ayala, Pouya Faridi, Geoffrey I. Webb, Patricia T. Illing, Jamie Rossjohn, Tatsuya Akutsu, Nathan P. Croft, Anthony W. Purcell and Jiangning Song

Corresponding authors: Nathan P. Croft, Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Melbourne, VIC 3800, Australia. Tel: +61-3-9902-0473; E-mail: Nathan.Croft@monash.edu; Anthony W. Purcell, Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Melbourne, VIC 3800, Australia. Tel: +61-3-9902-9265; E-mail: Anthony.Purcell@monash.edu; Jiangning Song, Biomedicine Discovery Institute, Department of Biochemistry and Molecular Biology, and Monash Centre for Data Science, Monash University, Melbourne, VIC 3800, Australia. Tel: +61-3-9902-9304; E-mail: Jiangning.Song@monash.edu

[†]These authors contributed equally to this work.

Abstract

Neopeptide-based immunotherapy has been recognised as a promising approach for the treatment of cancers. For neopeptides to be recognised by $CD8^+$ T cells and induce an immune response, their binding to human leukocyte antigen

Shutao Mei is a PhD candidate in the Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Australia. His research interests are bioinformatics, machine learning and immunopeptidomics.

Fuyi Li is currently a Research Fellow in the Department of Microbiology and Immunology, The Peter Doherty Institute for Infection and Immunity, The University of Melbourne, Australia. His research interests are bioinformatics, computational biology, machine learning and data mining.

Dongxu Xiang is a Research Assistant in the Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Australia. His research interests are bioinformatics, computational biology, biochemistry and machine learning.

Rochelle Ayala is a Senior Research Assistant in the Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Australia.

Pouya Faridi is a Research Fellow in the Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Australia. His research interests are peptidomics, cellular immunology and clinical trials.

Geoffrey I. Webb received his PhD degree in 1987 from La Trobe University, Australia. He is Director of the Monash Centre for Data Science and Professor in Faculty of Information Technology at Monash University, Australia. His research interests include machine learning, data mining, computational biology and user modelling.

Patricia T. Illing is a Research Fellow in the Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Australia. Her research focusses on understanding peptide and small molecule presentation by MHC molecules in the context of infection and adverse drug reactions.

Jamie Rossjohn is a Professor and Group Leader in the Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Australia. He is currently an ARC Australian Laureate Fellow. His research is focused on an understanding the structural and biophysical basis of MHC-restriction, TCR engagement and the structural correlates of T-cell signalling.

Tatsuya Akutsu received his DEng degree in Information Engineering in 1989 from the University of Tokyo, Japan. Since 2001, he has been a Professor in the Bioinformatics Center, Institute for Chemical Research, Kyoto University, Japan. His research interests include bioinformatics and discrete algorithms.

Nathan P. Croft is a Senior Lecturer and Group Leader in the Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Australia. He focuses on identifying and quantifying peptides displayed by MHC molecules on the surface of infected cells for scrutiny by $CD8^+$ T cells.

Anthony W. Purcell is a Professor and Group Leader in the Monash Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Australia. He is currently an NHMRC Principal Research Fellow. He specialises in targeted and global quantitative proteomics of complex biological samples, with a specific focus on identifying targets of the immune response and host-pathogen interactions.

Jiangning Song is an Associate Professor and Group Leader in the Monash Biomedicine Discovery Institute and Biochemistry and Molecular Biology, Monash University, Australia. He is a member of the Monash Centre for Data Science, Monash University. His research interests include artificial intelligence, bioinformatics, machine learning, big data analytics and pattern recognition.

Submitted: 30 September 2020; **Received (in revised form):** 29 November 2020

class I (HLA-I) molecules is a necessary first step. Most epitope prediction tools thus rely on the prediction of such binding. With the use of mass spectrometry, the scale of naturally presented HLA ligands that could be used to develop such predictors has been expanded. However, there are rarely efforts that focus on the integration of these experimental data with computational algorithms to efficiently develop up-to-date predictors. Here, we present Anthem for accurate HLA-I binding prediction. In particular, we have developed a user-friendly framework to support the development of customisable HLA-I binding prediction models to meet challenges associated with the rapidly increasing availability of large amounts of immunopeptidomic data. Our extensive evaluation, using both independent and experimental datasets shows that Anthem achieves an overall similar or higher area under curve value compared with other contemporary tools. It is anticipated that Anthem will provide a unique opportunity for the non-expert user to analyse and interpret their own in-house or publicly deposited datasets.

Key words: HLA-I peptide binding prediction; machine learning; scoring function; web server; model customisation

Introduction

The binding between peptides liberated from protein antigens during intracellular proteolysis, and human leukocyte antigen class I (HLA-I) molecules is a prerequisite for T cell recognition and the initiation of an immune response. Therefore, predicting whether a peptide can bind to a given HLA-I molecule can fast track epitope identification and thereby improve vaccine design efficiency [1]. With the broadening employment of mass spectrometry to generate large immunopeptidomics datasets, a considerable number of experimentally validated peptide binders are being identified for different HLA-I allotypes [2, 3]. Consequently, the availability of large-scale immunopeptidome data provides growth opportunities to develop data-driven prediction models for peptide-HLA-I binding and achieve accurate prediction results.

In recent years, considerable efforts have been focused on developing methods and tools for the prediction of HLA-I peptide binding. To date, several benchmarking studies have summarised and evaluated the performance of these tools [1, 4, 5]. In general, these HLA-I peptide binding prediction tools can be classified into two major categories based on their prediction algorithms: those that are scoring function-based or machine learning-based. For scoring function-based tools such as MixMHCpred [6], Pickpocket [7] and PSSMHCpan [8], peptides are encoded according to a specific scoring function, such as position weight matrix (PWM) or blocks substitution matrix 62 (BLOSUM62) (blocks substitution matrix built from BLOCK database [9] with less than 62% similarity). Then, sequence-based features are extracted and used to generate a scoring matrix for a specific HLA-I allotype. However, all these tools rely on only one or two scoring function methods and thus cannot comprehensively encode the amino acids of a peptide sequence. Machine learning-based tools, such as NetMHCpan [10], NetMHC [11] and MHCflurry [12], on the other hand, utilise machine learning algorithms, neural networks (NNs) or deep learning techniques, to train the model based on the peptide sequence information. These tools use feature representation algorithms to generate a higher dimension of features from the peptide sequence. Their prediction performance is often better than some of the scoring function-based tools [5]. However, considerable time and effort are required to train the model, particularly for tools that are developed based on deep learning algorithms [13, 14]. In addition, the use of higher-dimensional features for model training can lead to data overfitting issues [15].

In parallel with these algorithmic developments, the capacity to generate large amounts of immunopeptidomic data is increasing, with recent studies improving the depth of coverage and quality of peptide identifications [16–21]. As a consequence, it is not efficient or practical to frequently train new predictive models based on such rapidly accumulated data and update online servers for the wider community. In addition, some laboratories prefer only to use prediction models trained by their in-house data, especially in cases considering data privacy and intellectual property.

In this study, we introduce Anthem, which facilitates model training using user input data with potential improvements in the performance of HLA-I peptide binding prediction. Anthem utilises five different scoring functions that are commonly used in HLA peptide binding prediction as well as sequence-based prediction tools [7, 8, 21–25], allowing comprehensive encoding of the amino acid sequence of each peptide. Specifically, for each binder length (length range 8–14 amino acids) of each HLA-I allotype trained, we employed the wrapper feature selection method to select the best scoring function combination set, which can achieve optimal prediction performance in terms of area under curve (AUC) value [26, 27]. Then, we built an aggregating one-dependence estimators (AODE) model trained by the integration of the outcomes of each scoring function from the set. Evaluation on independent datasets including peptide binders from various HLA-I molecules illustrates that Anthem could achieve overall similar or higher AUC values compared with seven other current HLA-I peptide binding prediction tools [10, 13, 14, 21, 24, 25, 28, 29]. Moreover, we further evaluated Anthem in comparison with seven other prediction tools using in-house generated immunopeptidomics datasets that delineated peptide binders for five HLA-I allotypes (HLA-A*01:01, -A*02:01, -A*24:02, -B*08:01 and -B*18:01). In this evaluation, Anthem achieved highest AUC on three of five HLA-I allotypes in their primary binder length (HLA-A*01:01, -A*02:01 and -B*18:01); and second highest on the remaining two (HLA-A*24:02 and -B*08:01). In addition, with the function of customised model training, we implemented a novel function within Anthem to allow end-users to easily and efficiently train a new model based on their own input dataset and apply it to make testable predictions. The web server of Anthem has been implemented and is available at http://anthem.erc.monash.edu/.

Materials and methods

Dataset collection

To obtain a comprehensive dataset containing peptide binders across different HLA-I allotypes, we collected data from three sources: (i) HLA-I binders from databases including IEDB [30], EPIMHC [31], MHCBN [32] and SYFPEITHI [33]. Of note, EPIMHC, MHCBN and SYFPEITHI only store binary data (i.e. positive or negative) to distinguish whether a peptide has been experimentally verified to be a binder or not, whereas for some peptides in the IEDB database, quantitative measurements (e.g. binding affinity) have been recorded in addition to the binary result. To construct the positive dataset (contains experimentally verified peptide binders) for model training and independent dataset, we collected all positive peptides from all the above four public databases; (ii) Allotype-specific HLA-I ligands identified by mass spectrometry in previously published studies [16, 34–48] and (iii) Peptide binders in the training datasets from other HLA-I binding prediction tools [8, 12–14, 25, 30, 49–56].

For peptide binders collected from the above three sources, we further refined the entries by the following criteria: (i) Peptide binders with non-specific HLA-I allele name, such as 'HLA-A30' or 'HLA class I' were excluded; (ii) Peptide binders that contained unusual or ambiguous amino acids, namely B, J, O, U, X and Z were excluded; (iii) For each HLA-I allotype, only the peptide binders with the length range of 8–14 amino acids were selected, as this is the typical length range for HLA-I peptide binding [57] and (iv) For each peptide binder (with the length of 8–14 amino acids), in order to get a model with robust performance, only a length range that contains no less than 50 peptide binders to be used for model training was selected for model construction for the corresponding HLA-I allotype [23].

In total, the final curated dataset used for training and evaluating the model consisted of 40 HLA-A, 56 HLA-B and 16 HLA-C allotypes. The number of ligands of each peptide length per HLA-I allotype ranged from 63 to 20 106. For peptide binders in each available length for each HLA-I allotype, we selected 80% to train a model for the corresponding length of that HLA-I allotype. The remaining 20% of peptide binders were used to construct an independent test dataset. In order to generate the negative data for both training and independent test datasets for each HLA-I allotype, we randomly selected sequence segments from the source proteins of IEDB HLA-I immunopeptidomes according to previous studies [10, 23] as well as the three following rules: (i) Matched length of sequence segment to corresponding HLA-I ligands,(ii) The number of sequence segments was the same as the training and independent test datasets and (iii) For a given sequence length of an HLA-I allotype, the two positive datasets (i.e. training and independent test) and two corresponding negative datasets had no identical peptides with each other.

A summary of the peptide binders used in the training and independent test datasets are shown in Supplementary Table S1. The training and independent test samples are provided in the Supplementary Excel File. It should be noted that there was not always enough data for some ligand lengths e.g. 13 or 14-mer for some allotypes. This precluded building models for ligands of this length for these HLA allotypes.

In-house experimental data

The methods used to generate in-house datasets are described as previously [2, 58]. Briefly, The Epstein Barr virus-transformed B-lymphoblastoid cell line C1R, which expresses very low levels of endogenous class I HLA molecules [59, 60], was stably transfected with either HLA-B*08:01 or HLA-B*18:01 by electroporation [61]. Transfected cells were grown in RPMI-1640 (Invitrogen) supplemented with 50 IU/ml penicillin, 50 μg/ml streptomycin, 7.5 mM HEPES (Sigma), 2 mM L-glutamine (MP Biomedical), 75 μM ß-mercaptoethanolamine (Sigma), 0.1 mM non-essential amino acids (Invitrogen) and 10% foetal calf serum (RF-10). A total of 0.3 mg/ml hygromycin (Invitrogen) was added to select for stable expression of the transfected HLA-I allotypes, which was confirmed by flow cytometry. Cells were harvested by centrifugation (3724 × g at 4°C for 15 min), washed with PBS, pelleted by centrifugation (524 × g at 4°C for 10 min) and snap-frozen in liquid nitrogen. Pellets were stored at −80°C until processing. The isolation HLA-I bound peptides and peptide sequence identification were performed as in our previous work [58]. In order to get peptides that are restricted by the transfected allele of interest, endogenous peptides derived from the low-level expression of HLA-B*35:03 and HLA-C*04:01 as well as redundant peptides overlapped in the training and independent datasets were removed. Finally, the remaining peptides from HLA-B*08:01 and -B*18:01 as well as from HLA-A*01:01, -A*02:01, -A*24:02 of our previous experimental work [58] were used as the experimental datasets to further evaluate the HLA-I peptide binding prediction tools in this study.

The method for generating negative data for the experimental dataset of each HLA-I allotype is the same as described before for the independent dataset, with an exception of the rules that have been changed to: (i) Matched length of sequence segment to the corresponding HLA-I ligands; (ii) The number of sequence segments was the same for the experimental dataset and (iii) For a given sequence length of an HLA-I allotype, the positive datasets (i.e. the training, independent and experimentally identified peptides) and the three corresponding negative datasets should have no identical peptides with each other.

The datasets used for the case study of training a custom Anthem model are derived from the PRIDE repository [62] (PRIDE accession: PXD015398). The data are originated from our previous study using CIR transfected with HLA-B*57:01, thus raw data were searched and processed as per in-house experimental data and the C1R endogenous peptides has been removed [63]. Data were searched against the reviewed human proteome (UniProt Swiss-Prot, accessed October 2018), using the following parameters: parent mass error tolerance 25 ppm, fragment mass error tolerance 0.1 Da, enzyme none, variable modifications methionine oxidation and glutamine and asparagine deamidation, up to three variable modifications per peptide.

Sequence scoring functions in Anthem

Anthem employs a two-layer prediction system. The first layer utilises five widely used sequence scoring functions that calculate a binding probability score from sequence information. In the second layer, AODE algorithm trains a model based on these peptide sequence dependent features for predicting peptides of 8–14 amino acids in length that bind to a given HLA-I allotype. A description of each of the scoring functions used by Anthem is provided below.

Amino acid frequency

The amino acid frequency (AAF) score of a peptide is a widely used sequence scoring function [64–66], which can be calculated as:

$$\mathrm{Score}_{\mathrm{Frequency}} = \sum_{\mathrm{i}=1,\mathrm{len}} NF_{\mathrm{i}}, \tag{1}$$

where NF_{i} is the normalised relative AAF, whereas len denotes the length of the peptide. NF_{i} is defined as:

$$\mathrm{NF}_{\mathrm{i}} = \frac{f_{\mathrm{i}}}{f_{\mathrm{i,max}}}, \tag{2}$$

where $f_i = \frac{n_i}{N}$ represents the frequency value of the amino acid at the position i, whereas $f_{i,\max}$ denotes the frequency value of the most common amino acid at the same position.

WebLogo-based sequence conservation

WebLogo [67] is a widely used sequence logo generation tool based on the calculation of the sequence conservation score (W) [64, 66, 68]. Here, we applied the conservation score generated by WebLogo to rank all the peptide binders. The conservation score of a peptide can be calculated as:

$$\mathrm{Score}_{\mathrm{WebLogo}} = \sum_{i=1,len} \mathrm{W}_i, \tag{3}$$

where W_i denotes the conservation score of the amino acid at the position i.

Position-specific scoring matrix

Position-specific scoring matrix (PSSM)-based methods can generate a matrix based on peptide binders with the same length of an HLA-I allotype and infer the motif information of that HLA-I allotype. Several HLA-I binding prediction tools for predicting potential binders are developed based on the scoring matrix generated by PSSMs [7, 8, 21]. In this study, we adopted two PSSM methods to construct the three corresponding scoring matrices. We used the PSSM method described in the PSSMHCpan tool [8] to generate a scoring matrix based on peptide binders firstly, which is defined as:

$$P_{ai} = \log \frac{F_{ai}}{BG_a}, \tag{4}$$

where P_{ai} *is* the element in the matrix, F_{ai} denotes the frequency of amino acid a at the position i in the training dataset. BG_a denotes the background frequency of amino acid a from the UniProt database [69].

The other scoring matrix was generated by converting the peptide binders with the same length of an HLA-I allotype to the PWM, which was used as a feature in MixMHCpred 2.0.2 [21, 28]. Briefly, the element P_{ai} in PWM was calculated by dividing the frequency of amino acid a at the position i with the total number of amino acids at the position i, Then, it was further divided by 0.05 and log-transformed to get the position weight of amino acid a at the position i.

Substitution matrix index

The substitution matrix index (SMI) has been widely used in generating sequence-based features. Here, we used BLOSUM62 to extract the feature. BLOSUM62 is a typical scoring matrix used to evaluate the similarity between divergent sequences. It has been used by several existing HLA-I binding prediction tools [10, 23–25]. Generally, for a given test peptide (test), the SMI score can be calculated as:

$$\mathrm{Score}_{\mathrm{SMI}} = \sum_{\mathrm{i=1,len}} \{\mathrm{Matrix}\}\left(\mathrm{Test}\,[\mathrm{i}]\right)[\mathrm{i}] \tag{5}$$

$\sum_{\mathrm{i=1}}^{\mathrm{len}}\{\mathrm{Matrix}\}(\mathrm{Test}[\mathrm{i}])[\mathrm{i}]$ sums up the substitution score of the test peptide against the matrix generated by corresponding training dataset based on the BLOSUM62 matrix.

Aggregating one-dependence estimators

Anthem utilises the AODE computational framework [70], which is a variant of Naïve Bayes (NB) for building a prediction model and has been successfully used in many bioinformatics fields [71–73]. AODE is designed to avoid any model selection by enumerating all possible one-dependence classifiers in each of which there is an attribute as the parent of all others. In order to improve accuracy, when classifying an object $x = \langle x_1, \cdots, x_n \rangle$, the models are excluded, where the parent attribute appears in training data fewer than m times (m is set to 1 in present study). Its low training time complexity is computationally desirable when learning from large data sets [70]. In this study, given that each peptide can be represented by the vector $\mathrm{x} = \langle \mathrm{x}_1, \cdots, \mathrm{x}_n \rangle$, where x_i is a scoring function derived value; an AODE model can be trained to assign the peptide binder label y based on its posterior probability:

$$\mathrm{P}\left(\mathrm{y}|\mathrm{x}\right) = \frac{\mathrm{P}\left(\mathrm{y},\mathrm{x}\right)}{\mathrm{P}(\mathrm{x})} \propto \mathrm{P}\left(\mathrm{y},\mathrm{x}\right). \tag{6}$$

By aggregating all possible one-dependence classifiers, $P(y, x)$ can be formulated as:

$$P\left(y,x\right) = \frac{\sum_{1\leq i\leq n\wedge F(x_i)\geq m} P\left(y,x_i\right) P\left(x|y,x_i\right)}{|\left\{1 \leq i \leq n \wedge F\left(x_i\right) \geq m\right\}|}, \tag{7}$$

where $F(x_i)$ denotes the number of the training examples having attribute value x_i and is used to enforce the limit m on parent attributes.

Therefore, the label assignment (i.e. the peptide binder label y) can be derived as follows:

$$\underset{\mathrm{y}}{\mathrm{argmax}} \left(\sum_{i:1\leq i\leq n wedge F(x_i)\geq m} \hat{\mathrm{P}}\left(y,x_i\right) \prod_{j=1}^{n} \hat{\mathrm{P}}\left(x_j|y,x_i\right) \right), \tag{8}$$

where $\hat{\mathrm{P}}$ denotes the probability estimate.

Performance comparison on the averaged 5-fold cross-validation test was performed to compare the performance of AODE with that of other commonly used algorithms including eXtreme gradient boosting [74], logistic regression [66], NN [10], support vector machine [75], decision tree [76], random forest [77] and NB [78]. The results demonstrate that AODE achieved the best performance in terms the AUC value on the majority of HLA-I allotypes. The distribution of the AUC values and the detailed results of other evaluation metrics [i.e. sensitivity, specificity, accuracy and Matthew's correlation coefficient (MCC)] among different machine-learning algorithms are provided in Supplementary Figure S1 and Supplementary Table S2, respectively.

Performance evaluation

We used several standard performance measures, including sensitivity, specificity, accuracy and MCC [MCC value is between −1 and 1. A high MCC value (close to 1) means that the model can predict well the true labels for the peptides] [79] to comprehensively evaluate and compare the predictive performance of different machine algorithms or tools. These measures can be derived from the following four scalar quantities: TP (true positive: number of the correctly predicted peptides that bind to HLA class I molecules), TN (true negative: number of the correctly predicted peptides as non-binders of HLA class I molecules), FP (false positive: number of the incorrectly predicted peptides that bind to HLA class I molecules) and FN (false negative: number of the non-correctly predicted peptides as non-binders of HLA class I molecules). The above four measures are calculated as follows:

$$\begin{cases} \text{Sensitivity} = \frac{\text{TP}}{(\text{TP}+\text{FN})} \times 100 & 0 \le \text{Sensitivity} \le 1 \\ \text{Specificity} = \frac{\text{TN}}{(\text{TN}+\text{FP})} \times 100 & 0 \le \text{Specificity} \le 1 \\ \text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} & 0 \le \text{Accuracy} \le 1 \\ \text{MCC} = \frac{(\text{TP}\times\text{TN})-(\text{FN}\times\text{FP})}{\sqrt{(\text{TP}+\text{FN})\times(\text{TN}+\text{FP})\times(\text{TP}+\text{FP})\times(\text{TN}+\text{FN})}} & -1 \le \text{MCC} \le 1 \end{cases} \quad (9)$$

With the set of formulations in Eq. 9, the meanings of sensitivity, specificity, accuracy and MCC are easy to understand, as discussed in a series of similar HLA peptide binding studies and related works [7, 80, 81].

Moreover, we also plotted the receiver-operating characteristic (ROC) curves and calculated the AUC values to evaluate the predictive performance of Anthem and compare it with other existing algorithms or methods.

Results

Overview of Anthem

The framework of Anthem is provided in Figure 1. Figure 1A shows the construction and assessment processes of Anthem. There are four major steps including data collection and sequence scoring, feature selection, model training and performance evaluation. At the data collection step, the training and independent datasets were collected from both HLA-I peptide databases and HLA-I repertoires published by the research community as described in the section on dataset collection. In addition, in-house HLA-I peptides identified by mass spectrometry from PRIDE repository [62] (PRIDE accession: PXD014754, PXD021157, including HLA-A*01:01, -A*02:01, -A*24:02, -B*08:01 and -B*18:01) were also collected as an experimental test dataset to perform a case study for performance validation of Anthem. Next, selected scoring functions (see Materials and methods) were individually utilised to build a matrix based on peptides in training dataset for each binder length of each HLA-I allotype. Then, wrapper feature selection was implemented to determine the optimal combination of scoring functions that can achieve the best prediction performance in terms of AUC value [26, 27]. Finally, the scores derived from the selected scoring functions based on the training dataset were used as the input features to train the AODE model. The Anthem web server (http://anthem.erc.monash.edu/) was then implemented based on these models with a user-friendly interface. In terms of performance evaluation, the scores derived from the independent dataset and experimental dataset were used as two individual sources to evaluate the performance of Anthem relative to other HLA-I peptide binding prediction tools.

Figure 1B shows three function modes implemented and provided by Anthem. In general, in 'Prediction' mode, protein or peptide sequences entered by the user are first subjected to five scoring functions to generate the corresponding input features. Then, Anthem will utilise the corresponding models with the best feature set to make HLA-binding predictions. The outcomes generated by Anthem include prediction result and visualisation plots. In 'Train your model' mode, the input training dataset provided by the user is first encoded into the feature values based on five scoring functions. Then, Anthem will train a new model based on the selected features after the feature selection process. The outcomes include model files, model performance, the results of evaluation and prediction on the test and prediction files; users can also download these files locally for further use. In 'Use your model' mode, similar to the 'Prediction' mode, protein or peptide sequences are encoded into the input features by five scoring functions. Then, Anthem will apply the user-uploaded model generated under the 'Train your model' mode to perform the prediction based on the correspondingly derived feature combination. The outcomes generated are same as the 'Prediction' mode.

Performance comparison between Anthem and other prediction tools using independent datasets

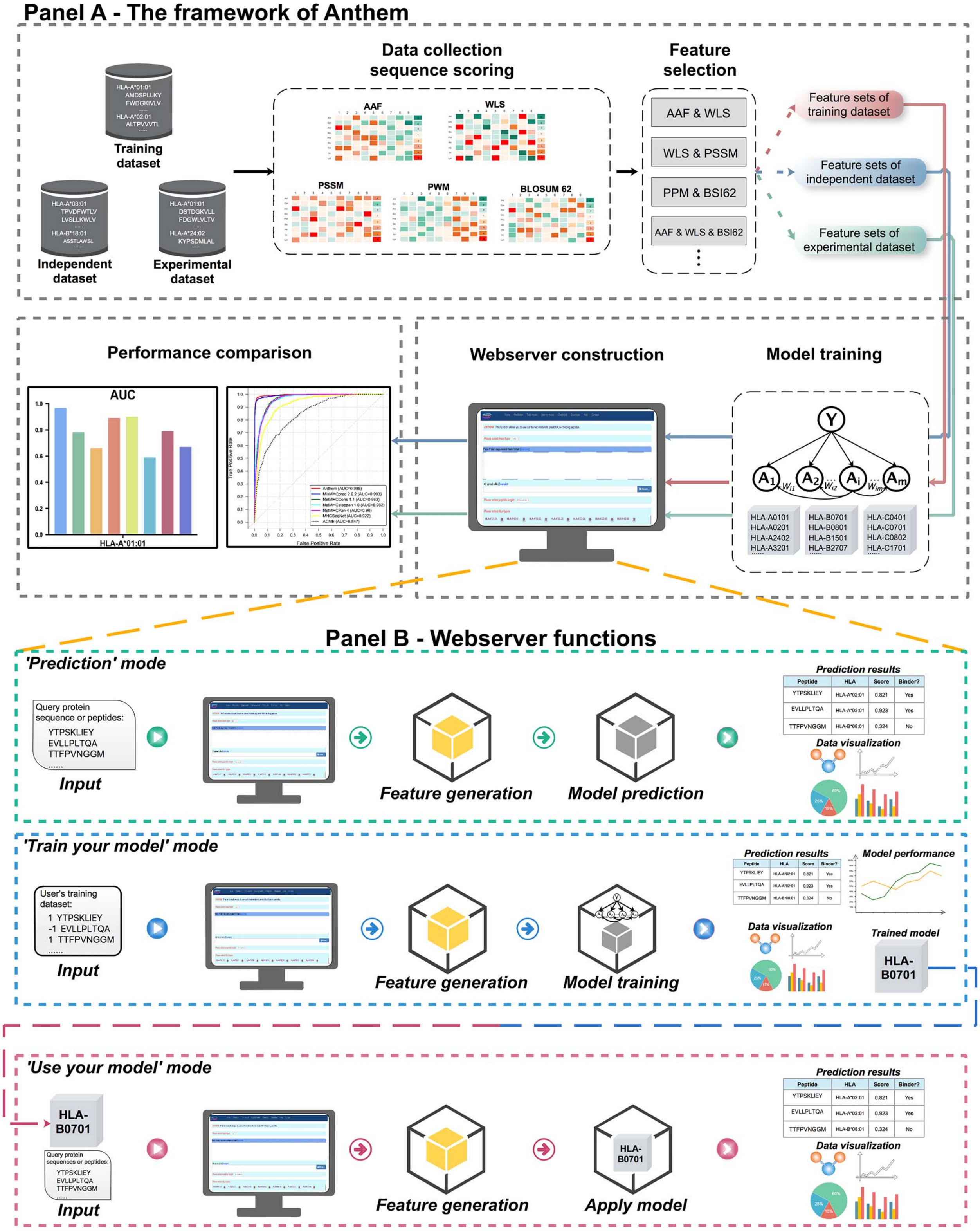
In this section, we compared the predictive performance of Anthem and seven existing tools on independent test datasets, which comprise 87 035 peptides binders from 112 HLA-I allotypes as well as randomly selected negative peptides as illustrated in the section of dataset collection. Specifically, Anthem was benchmarked against MixMHCpred 2.0.2 [21, 28], NetMHCpan 4.1 [10, 23], NetMHCcons 1.1 [29], NetMHCstabpan 1.0 [24], ACME [25], MHCSeqNet [14] and DeepSeqPan [13]. Some of these tools are reviewed as top-performing HLA-I prediction tools [5], whereas others are newly developed tools that have not been rigorously assessed yet. Figures 2 and 3 show the overall distribution of performance in terms of AUC value and other metrics (sensitivity, specificity, accuracy and MCC), respectively. As shown in Figure 2, the distribution of AUC value is centred on the range from 0.8 to 1.0 among different tools when the peptide length is from 8 to 11. However, the distribution of AUC values becomes wider when peptide length is from 12 to 14. Overall, it can be seen that Anthem achieved a distribution of AUC which is centred on higher values among all peptide lengths compared with other tools.

According to Figure 3, it can be seen that the value distributions of the other four metrics of different tools are similarly impacted by peptide length (except for MHCNetSeq, which displayed a wider value distribution on most metrics at different lengths). The details of evaluation performance can be found in Supplementary Table S3.

Overall, the evaluation of performance on independent datasets shows that no tool achieved the best performance across all HLA-I allotypes. Anthem has achieved the best performance in terms of AUC value on 12 HLA-I allotypes for all available peptide lengths, whereas for nearly 76 allotypes, the performance of Anthem is moderate among the tools on most of peptide lengths. Figure 4 shows an example of ROC curves on HLA-B*27:05, representative of Anthem achieving the highest AUC values among all length ranges compared to other tools (8–13-mers; for 14-mers, refer to the Supplementary Figure S2), whereas Figure 5 shows ROC curves on HLA-A*11:01, where Anthem achieves moderate AUC values compared with other tools.

Performance comparison between Anthem and other prediction tools using experimental datasets

In addition to evaluating the performance of Anthem using publicly available datasets, we were interested to further compare the prediction performance of Anthem with other tools on naturally presented HLA-I bound peptide data experimentally generated in our laboratory. Therefore, we performed a case



Prediction results

Peptide	HLA	Score	Binder?
YTPSKLIEY	HLA-A*02:01	0.821	Yes
EVLLPLTQA	HLA-A*02:01	0.923	Yes
TTFPVNGGM	HLA-B*08:01	0.324	No

Prediction results

Peptide	HLA	Score	Binder?
YTPSKLIEY	HLA-A*02:01	0.821	Yes
EVLLPLTQA	HLA-A*02:01	0.923	Yes
TTFPVNGGM	HLA-B*08:01	0.324	No

Prediction results

Peptide	HLA	Score	Binder?
YTPSKLIEY	HLA-A*02:01	0.821	Yes
EVLLPLTQA	HLA-A*02:01	0.923	Yes
TTFPVNGGM	HLA-B*08:01	0.324	No

Figure 1. The framework of Anthem. (**A**) graphical illustration of construction and evaluation steps of Anthem, which includes data collection, sequence scoring, feature selection, model training and evaluation and web server construction; (**B**) graphical illustration of three function modes for using Anthem. The top function is 'Prediction', the middle is 'Train your model', whereas the bottom is 'Use your model'.

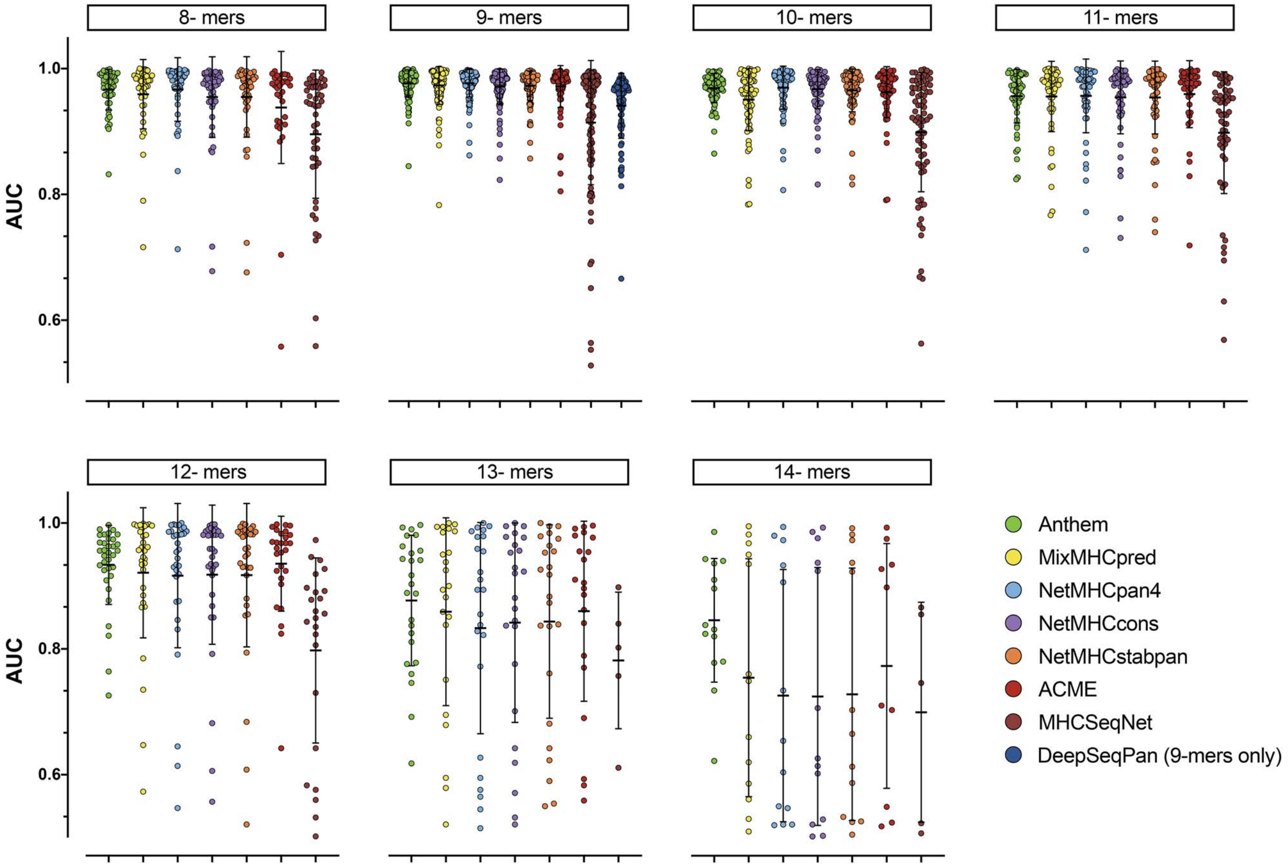


Figure 2. The distribution of AUC value of different tools on peptide length from 8 to 14 (for DeepSeqPan, only 9-mers), the line and error bar in each scatter dot plot represents mean with SD.

study based on in-house generated data from five HLA-I alleles (4912 peptides for HLA-A*01:01, 4673 peptides for HLA-A*02:01, 4419 peptides for HLA-A*24:02, 21 331 peptides for HLA-B*08:01 and 21 593 peptides for HLA-B*18:01). The experimental methods were as previously described [45, 58] and as briefly described in the methods and materials section. We used these peptide repertoires as the experimental datasets to evaluate the relative predictive performance of Anthem. The performance comparison results between Anthem and state-of-the-art methods are summarised in Supplementary Table S4. Figure 6 shows the details of performance of different tools on peptides with the predominant length (9-mer) from five HLA-I allotypes. No tool was able to achieve the best performance across all five metrics (AUC, sensitivity, specificity, accuracy and MCC). However, Anthem achieved the best AUC value on 3/5 allotypes and the best MCC value among all HLA-I allotypes.

It should be noted that in terms of sensitivity (TP rate), Anthem did not achieve the best performance on any of five HLA-I allotypes, whereas in terms of specificity (TN rate), Anthem has achieved the best performance for all HLA-I allotypes. These data indicate that the prediction results from Anthem are more stringent towards having less FPs (non-binders that are predicted as binders) compared to the default binder threshold of other tools. However, given that this may cause Anthem to lose some TPs (binders that are predicted as binders) we have deployed a function within Anthem that allows end users to adjust the binder threshold from the recommended default.

The Anthem web server

As an implementation of the methodology of Anthem, we have created an online web server using the PHP programming language, which is freely available at http://anthem.erc.monash.edu/. Anthem has three functions: the first one is for the prediction of peptides binding to HLA class I molecules using existing models, which is called 'Prediction' mode; the second function of Anthem is designed for users who want to train new models based on their own datasets, which is termed as 'Train your model' mode. The last function of Anthem is for users to upload their own trained model to make predictions, which is called 'Use your model' mode (Figure 7A).

In 'Prediction' mode, the primary function is to allow users to perform HLA-I peptide binding predictions based on their query peptide list or protein sequence. The input should be the sequences of interest in either FASTA format or peptide sequence format (Figure 7B-1). In 'Train your model' and 'Use your model' mode, users can train a new model based on their in-house data and use the trained model for making new HLA-I peptide binding predictions (Figure 7B-2). In 'Train your model' mode, users need to provide a training dataset file containing peptides to train and build the model. In addition, users can optionally provide a test dataset file to test the trained model. Sequences of interest can be also optionally provided to make predictions by using the trained model. In 'Use your model' mode, users need to provide the trained model file generated from the mode of 'Train your model', sequences of interest in either FASTA format or peptide sequence format. We tested the

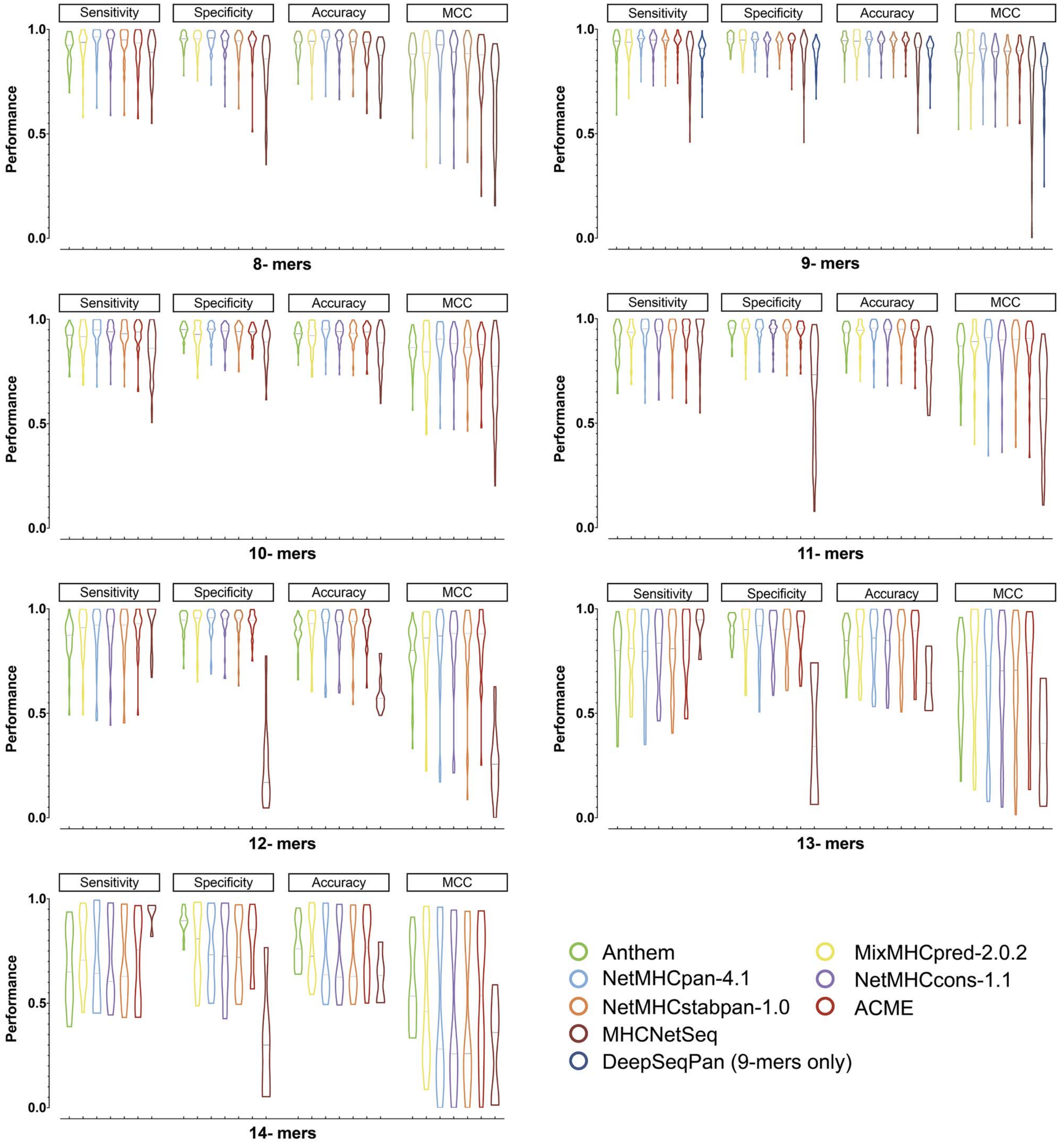


Figure 3. The distribution value of four metrics (sensitivity, specificity, accuracy and MCC) of different tools on peptide length from 8 to 14 (for DeepSeqPan, only 9-mers); the dashed line in each violin plot represents the median.

computational efficiency of the Anthem web server operating on both 'Prediction' and 'Train your model' modes. The averaged computational time for processing 1000 peptide sequences (9-mer) as well as a fasta sequence containing 20 000 amino acids (this is the maximum number of amino acids in a fasta sequence for Anthem binding prediction; here, the 20 000 amino acids were generated by translating from the vaccinia Western Reserve strain genome to represent a relevant biological example [82], with the genome downloaded from Genbank [83]) from five randomly selected HLA-I allotypes was 5.4 and 22.6 sec, respectively. In addition, the averaged computational time for training a model based on 20 000 peptide sequences (9-mer, based on five randomisations) was 29.6 sec.

Once submitted, a unique Job ID will be generated to refer to the job summary page (Figure 7B-3) during the job execution process. Users can use this ID to track their job execution progress, and access or download their prediction results once completed (Figure 7B-4 and B-5). The outputs of 'Prediction' mode are the prediction results and visualisation figures, including predicted binders and binder position (if provided with FASTA format).

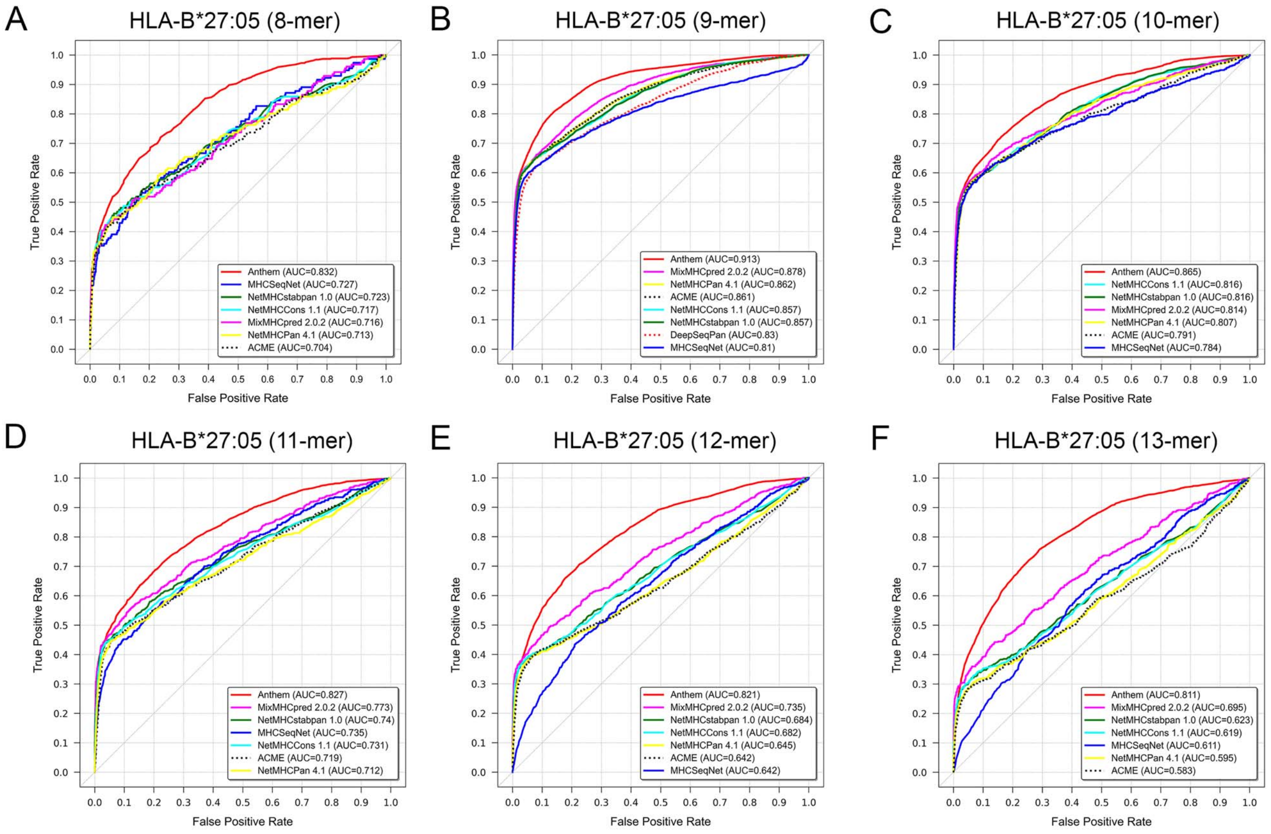


Figure 4. ROC curves and the corresponding AUC values of predictors for HLA-I peptide binding prediction on the independent data dataset specific to (**A**) HLA-B*27:05 (8-mer), (**B**) HLA-B*27:05 (9-mer), (**C**) HLA-B*27:05 (10-mer), (**D**) HLA-B*27:05 (11-mer), (**E**) HLA-B*27:05 (12-mer) and (**F**) HLA-B*27:05 (13-mer).

The outputs of 'Train your model' mode include the trained model file, the performance file, the ROC curve of the model, the ROC curve of the test results (if provided with test file), visualisation figures including predicted binders (if provided with prediction file) and binder position (if provided with prediction file in FASTA format). The outputs of 'Use your model' mode are the same as the 'Prediction' mode, which include the prediction results and visualisation figures, including predicted binders and binder position (if provided with FASTA format) (Figure 7B-4 and B-5).

Case study of model customisation

To illustrate the utility of Anthem's model customisation function, we performed a case study on the modified HLA-B*57:01 peptide repertoire induced by abacavir treatment [63, 84], comparing the results to NetMHCpan 4.1. This is an ideal test dataset for Anthem because abacavir alters the amino acids favoured as the C-terminal binding residue of HLA-B*57:01 peptide ligands, and whilst the molecular mechanism of this process is understood to be due to occupation of the antigen binding cleft of HLA-B*57:01 by abacavir, there is no binding prediction tool trained to accommodate such a drug-induced repertoire shift. First, we used the existing HLA-B*57:01 model of Anthem to predict the unmodified peptide repertoire (i.e. peptides from untreated abacavir datasets) and compared the prediction results of Anthem with those of NetMHCpan4.1 (Figure 8A). Next, we trained a new Anthem model (via the 'Train your model' function) using 80% of the peptides that were both unique to abacavir treatment and lacking C-terminal tryptophan (which is disfavoured by binding of abacavir within the antigen-binding cleft [84]). Then, we tested the new model and NetMHCpan4.1 on the remaining 20% of the peptides meeting these criteria (Figure 8B). As show in Figure 8A, when predicting the unmodified repertoire, Anthem performed similarly to NetMHCpan 4.1 in terms of the percentage of predicted binders. When predicting peptides unique to abacavir treatment, Anthem achieved an overall better performance in terms of the percentage of predicted binders compared with NetMHCpan 4.1 (Figure 8B). The case study results showed that the model customisation function in Anthem provides a promising approach for the prediction of specific peptides that can cater for researchers' specific needs.

Comparison with other web servers

Several HLA-I peptide binding prediction tools have been deployed publicly and can be accessed as web servers, including the tools evaluated herein, such as NetMHCpan 4.1, NetMHCcons 1.1, NetMHCstabpan 1.0, as well as other tools such as SYFPEITHI [33], RANKPEP [85], PickPocket [7], SMMPMBEC [86, 87], NetMHC 4.0 [11], ConvMHC [54] and IEDB-AR-Consensus [86]. We provide a detailed comparison of the main characteristics of Anthem with these web servers according to eight different aspects, reported in Table 1. In particular, Anthem has three significant advantages: (i) Anthem employs a feature selection step to select the best feature combination to optimise the model performance in terms of the AUC value; (ii) Anthem employs a robust Bayesian ensemble algorithm, AODE, to integrate diverse sequence scoring functions and (iii) Anthem provides its own model training function that allows users to build

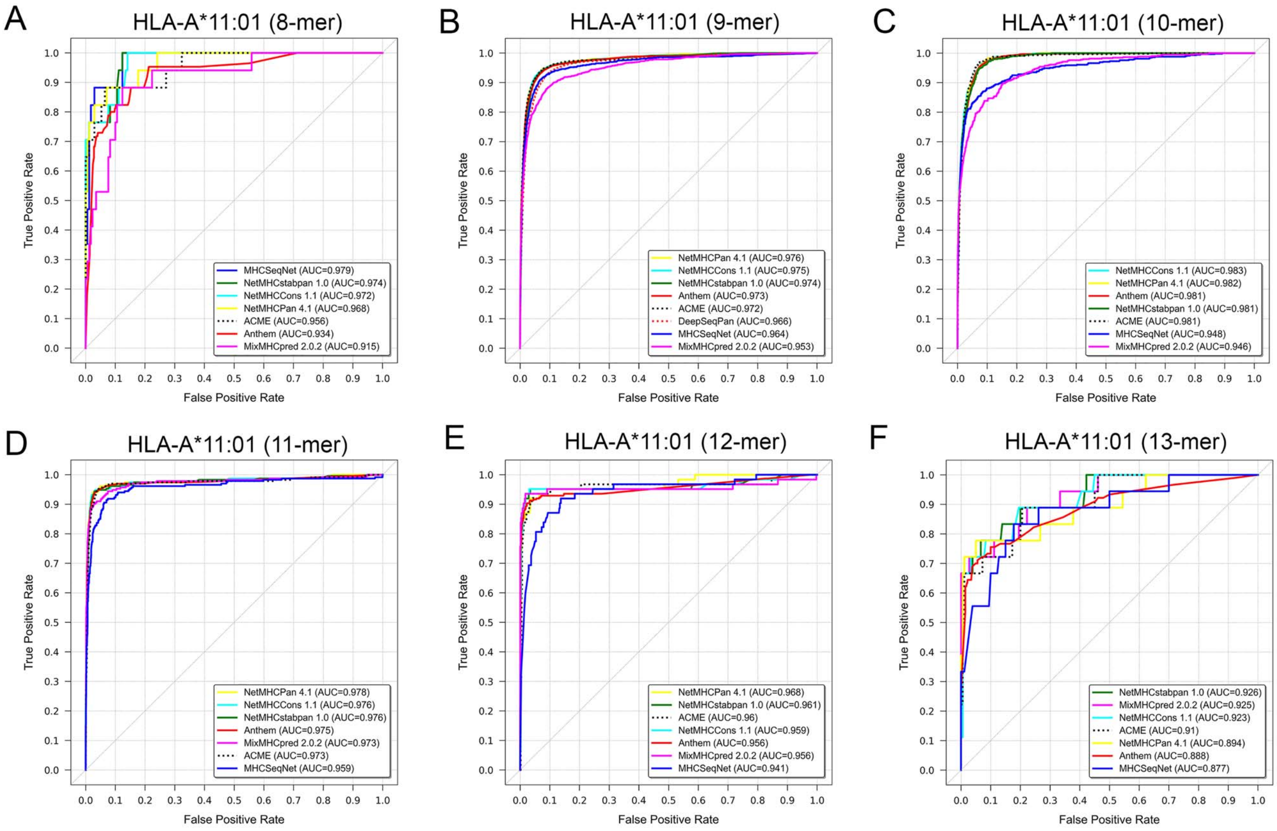


Figure 5. ROC curves and the corresponding AUC values of predictors for HLA-I peptide binding prediction on the independent data dataset specific to (**A**) HLA-A*11:01 (8-mer), (**B**) HLA-A*11:01 (9-mer), (**C**) HLA-A*11:01 (10-mer), (**D**) HLA-A*11:01 (11-mer), (**E**) HLA-A*11:01 (12-mer) and (**F**) HLA-A*11:01 (13-mer).

their own prediction models based on in-house data easily and conveniently.

Discussion

In this study, we present Anthem, a two-layer and user-friendly computational method for user customised HLA-I peptide binding prediction. Anthem combines the advantages of both a scoring function-based method and machine learning-based method. The evaluation results from both independent and experimental datasets demonstrate that Anthem achieves either an overall level of performance comparable to other existing tools or, for specific HLA-I allotypes, improved performance. For instance, Anthem has achieved the highest AUC value on all available peptide lengths in 12 HLA-I allotypes using independent datasets. This might be contributed by the integration of both scoring function and machine learning methods. It is noteworthy that although MixMHCpred 2.0.1 utilised the PWM as the only scoring function for the prediction of peptide binding, it achieved a similar performance compared to Anthem and NetMHCpan 4.1. This indicates that simple scoring functions such as PWM and PSSM may play an important role in representing and extracting both the conserved and variable information in peptide sequences. Accordingly, the features derived from such scoring functions are crucial for improving the prediction performance of HLA-I peptide binding. In addition, Anthem is trained by the datasets collected from the most recent updates of public sources which may explain other discrepancies between tools. Therefore, it can be expected that with more and more immunopeptidome data being accessible from the public, tools that are regularly updated to include new data for training models should have improved performance on specific HLA-I allotypes. However, a potential drawback is that users may have to wait a long time for developers of the tool to update the training data, which is not convenient and efficient. Therefore, we empower Anthem to support customised training by users.

The implemented function of Anthem makes it a tool that end users can easily use to train customised models based on their own data. Such a function is expected to be useful for several reasons. Firstly, in the context of less well-studied HLA-I allotypes as well as non-human analogues, users would not be able to test predictions if any such alleles are absent from existing tools. Anthem can overcome these issues by users providing the corresponding training data to train a customised model for further prediction. Second, immunopeptidome data from cell lines or patient samples are usually derived from mixtures of HLA allotypes, and in some cases the precise allotypes may be unknown. Such scenarios present challenges with existing tools, as specific alleles must be specified to generate prediction scores. Anthem overcomes this issue by again allowing users to train a model based on their own input data, which can be derived from mixed HLA samples. Third, current tools are also challenged by performing prediction on some specific datasets, for example, spliced peptides [88] or the peptide repertoires that have been changed by treatments such as IFN-γ [89] or small-molecule drugs like abacavir [84], as these datasets are not

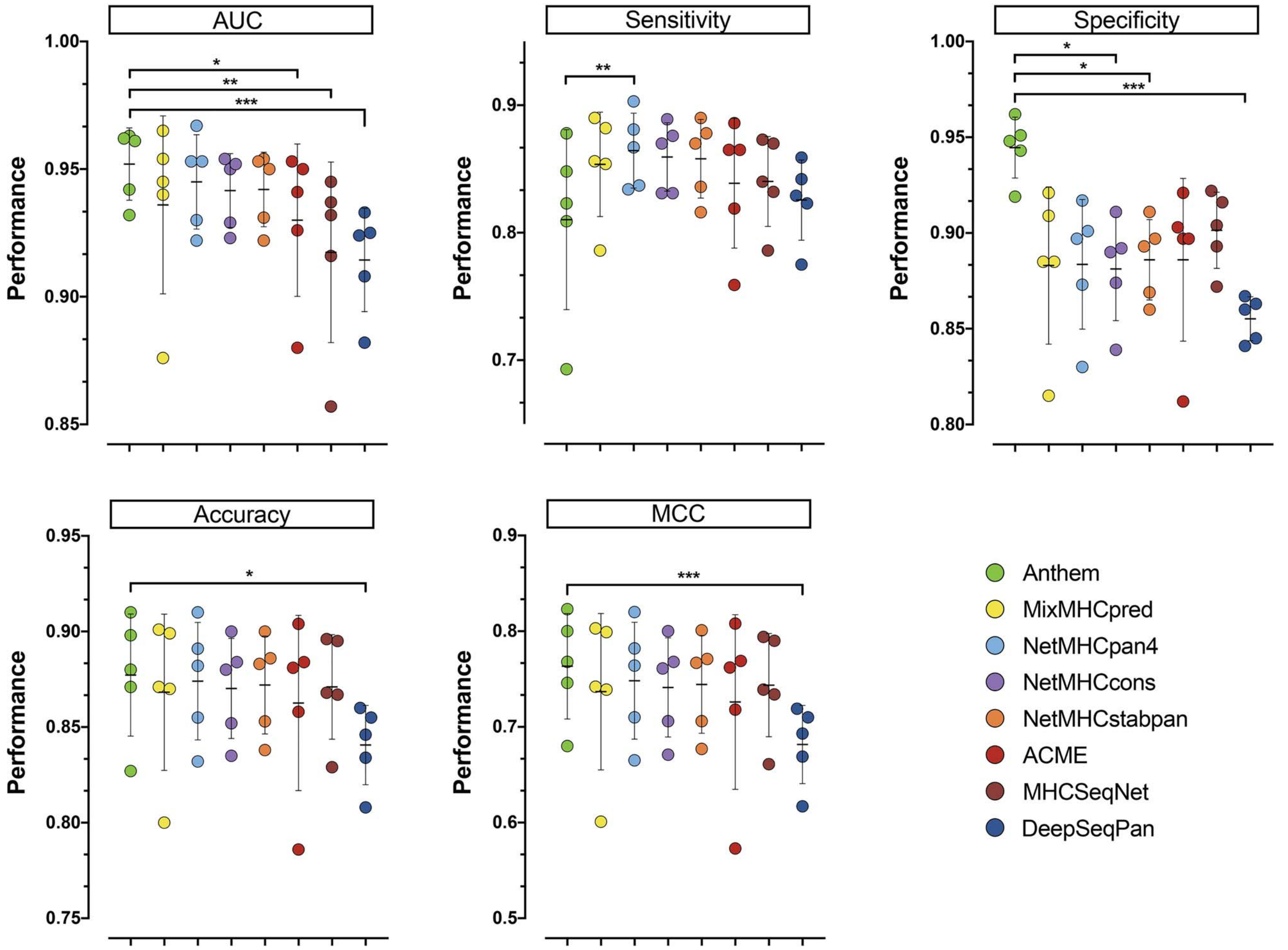


Figure 6. The distribution value of five metrics (AUC, sensitivity, specificity, accuracy and MCC) of different tools on the predominant length (9-mer) of five HLA-I allotypes from experimental datasets, the line and error bar in each scatter dot plot represents mean with SD. The significance difference level was determined by the one-way analysis of variance performed using GraphPad Prism 7 ($P < 0.05$ is statistically significant and is labelled as '*' i.e. *$P = 0.0332$, **$P = 0.0021$ and ***$P = 0.0002$).

included in the training datasets of current tools. Anthem can easily overcome this issue by the function of model customisation. Finally, for a patient sample, a model trained by the patient's immunopeptidome data can be used to assess the impact of HLA haplotypes and background genetics on peptide selection and binding in a more physiological context.

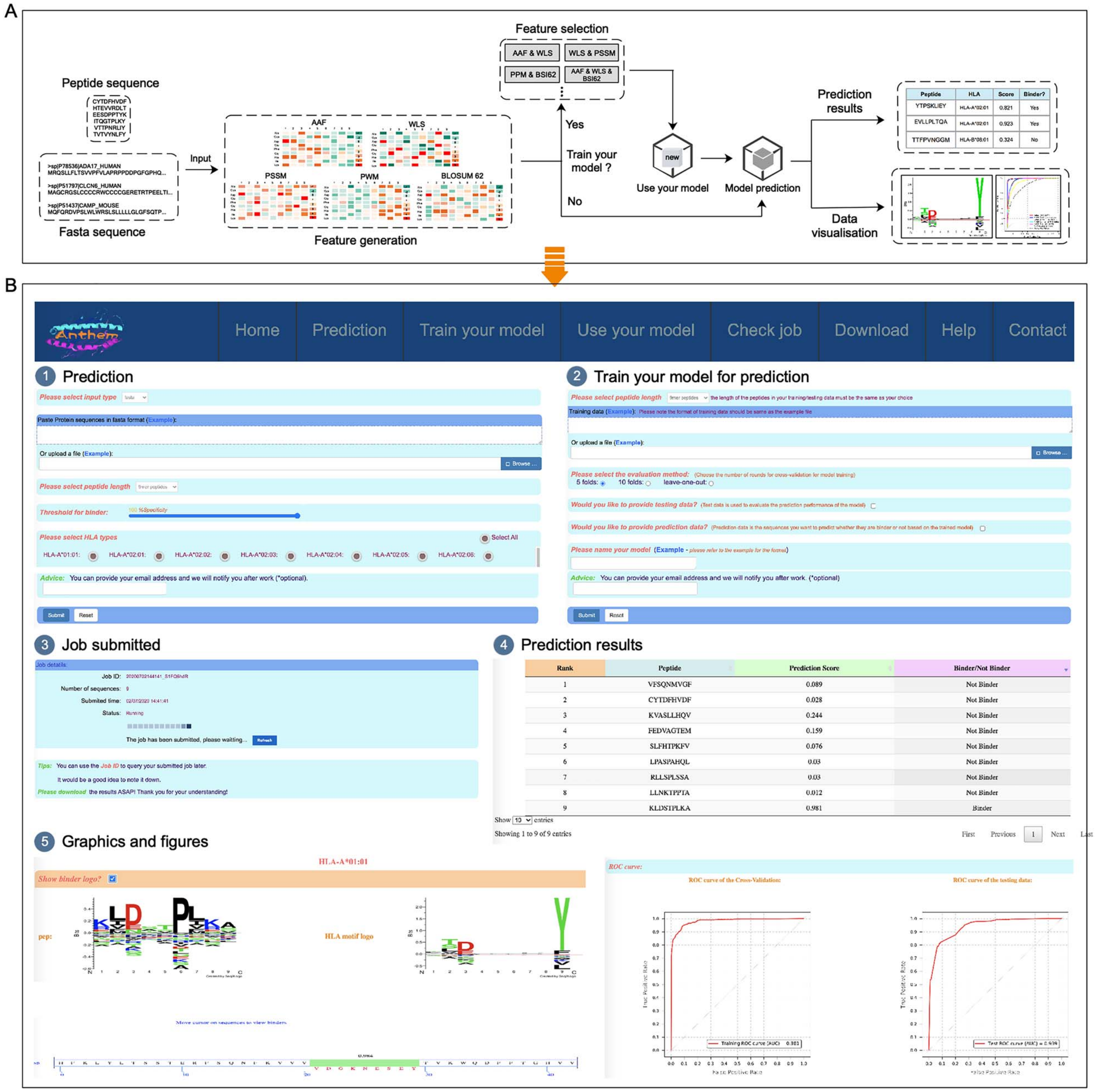
It should be noted that Anthem still has some limitations. For example, Anthem achieved the lowest AUC values when predicting on certain HLA-I allotypes compared with other tools on the independent test, such as 9-mer peptides of HLA-A*02:17, -A*02:50, -A*24:06. This suggests the AODE algorithm employed by Anthem may not be optimal for performing binding prediction on such allotypes. Therefore, we plan to examine the possibility of including other novel and promising algorithms such as deep learning [13, 14, 90, 91], ensemble/hybrid machine learning [92, 93], one-class classification algorithms, or positive-unlabelled learning algorithms in upgraded versions of Anthem in the future. The algorithm that achieves the best performance will be selected to construct the prediction model for the corresponding HLA molecule. We anticipate Anthem will be a valuable addition to current efforts in developing next-generation bioinformatics tools for HLA-I peptide binding prediction, and finally toward the development of improved T cell immunotherapy design.

Data availability

The Anthem package is freely available in the Github repository (https://github.com/17shutao/Anthem). ROC curves of each evaluated HLA-I allotype from both independent dataset and experimental dataset can be accessed in the Github repository (https://github.com/17shutao/Anthem). The mass spectrometry proteomics data have been deposited to the ProteomeXchange consortium via the PRIDE [62] partner repository with the dataset identifier PXD021157 (username: reviewer61482@ebi.ac.uk, password: yE8BElvC), including the raw .wiff files and peptide CSV files exported by PEAKS Studio 8.5, as well as mzIdentML files.

Key Points

- We developed a new tool for HLA-peptide binding prediction, *Anthem,* which combines multiple scoring functions and machine learning with a feature selection approach to train models for accurate HLA-I peptide binding prediction.
- Evaluation based on both independent and experimental datasets showed Anthem achieved equivalent, and in some cases improved prediction performance compared with contemporary tools.



Peptide	HLA	Score	Binder?
YTPSKLIEY	HLA-A*02:01	0.821	Yes
EVLLPLTQA	HLA-A*02:01	0.923	Yes
TTFPVNGGM	HLA-B*08:01	0.324	No

1 Prediction

Please select input type: fasta

Paste Protein sequences in fasta format (Example):

Or upload a file (Example): Browse ...

Please select peptide length: 9mer peptides

Threshold for binder: 100 %Specificity

Please select HLA types ● Select All

HLA-A*01:01: ● HLA-A*02:01: ● HLA-A*02:02: ● HLA-A*02:03: ● HLA-A*02:04: ● HLA-A*02:05: ● HLA-A*02:06: ●

Advice: You can provide your email address and we will notify you after work (*optional).

Submit Reset

2 Train your model for prediction

Please select peptide length: 9mer peptides the length of the peptides in your training/testing data must be the same as your choice

Training data (Example): Please note the format of training data should be same as the example file

Or upload a file (Example): Browse ...

Please select the evaluation method: (Choose the number of rounds for cross-validation for model training)
5 folds: ● 10 folds: ○ leave-one-out: ○

Would you like to provide testing data? (Test data is used to evaluate the prediction performance of the model) □

Would you like to provide prediction data? (Prediction data is the sequences you want to predict whether they are binder or not based on the trained model) □

Please name your model (Example - please refer to the example for the format)

Advice: You can provide your email address and we will notify you after work. (*optional)

Submit Reset

3 Job submitted

Job details:

Job ID: 20200702144141_S1FQ6hdR
Number of sequences: 9
Submited time: 02/07/2020 14:41:41
Status: Running

The job has been submitted, please waitting... Refresh

Tips: You can use the Job ID to query your submitted job later.
It would be a good idea to note it down.
Please download the results ASAP! Thank you for your understanding!

4 Prediction results

Rank	Peptide	Prediction Score	Binder/Not Binder
1	VFSQNMVGF	0.089	Not Binder
2	CYTDFHVDF	0.028	Not Binder
3	KVASLLHQV	0.244	Not Binder
4	FEDVAGTEM	0.159	Not Binder
5	SLFHTPKFV	0.076	Not Binder
6	LPASPAHQL	0.03	Not Binder
7	RLLSPLSSA	0.03	Not Binder
8	LLNKTPPTA	0.012	Not Binder
9	KLDSTPLKA	0.981	Binder

Show 10 entries
Showing 1 to 9 of 9 entries
First Previous 1 Next Last

5 Graphics and figures

Figure 7. Demonstration of the Anthem server. (A) The intuitive graphical workflow. (B) An example to showcase the usage of the Anthem server from (1 and 2) the input page for 'Prediction' and 'Train your model'/'Use your model', respectively, (3) the job summary page, (4) the prediction results and (5) data visualisations.

- Critically, Anthem enables users to train and customise their specific models using their own data sets for performing self-defined HLA-binding prediction. This is a novel feature of Anthem making such customised models accessible to the broader community.
- A user-friendly web server (http://anthem.erc.monash.edu/) and standalone package (https://github.com/17shutao/Anthem) have been developed and made publicly available.

Supplementary Data

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

Acknowledgement

The authors acknowledge the provision of instrumentation, training and technical support by the Monash Biomedical Proteomics Facility. Computational resources were supported by the R@CMon/Monash Node of the NeCTAR

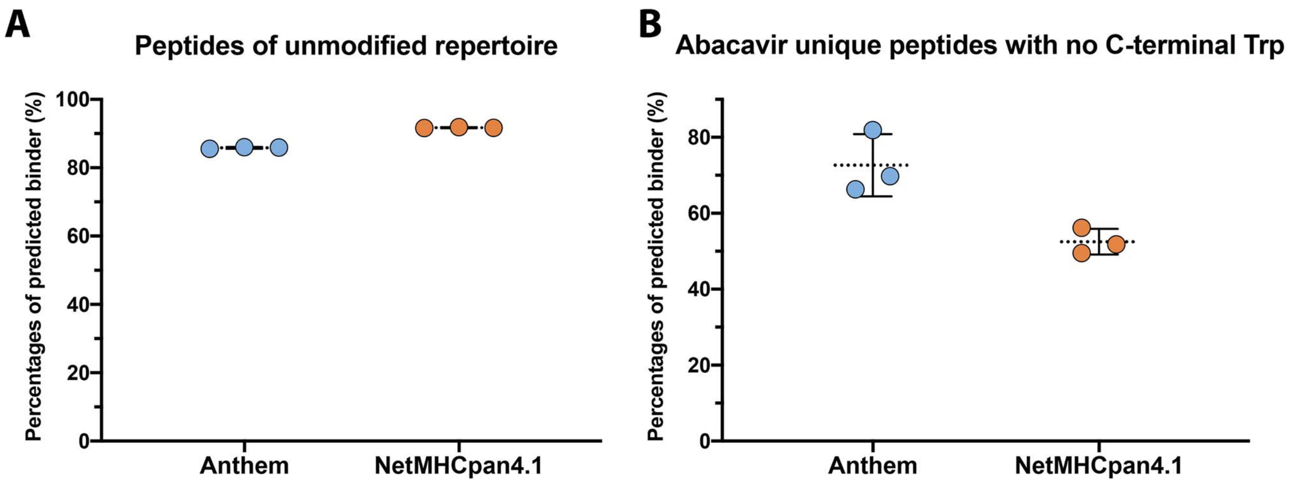


Figure 8. Prediction performance of Anthem and NetMHCpan 4.1 for predicting normal HLA-B*57:01 peptide repertoire (**A**) and peptides unique to abacavir treatment with no C-terminal Trp (**B**). For each subfigure, the comparison was performed based on three datasets. The dash line/error bar represents the mean with SD.

Table 1. Comparison of the Anthem web server with other web servers for HLA-I peptide binding prediction

Web server[a]	Feature[b]	Feature selection	Integrated method[b]	Software availability[c]	Max data upload[b]	Step-by-step instruction[c]	Email of result[c]	Train model by users[c]
SYFPEITHI		No	Matrix	No	Single sequence	No	No	No
RANKPEP		No	Matrix	No	NM	Yes	No	No
PickPocket 1.1		No	Matrix	Yes	≤5000 sequences and ≤20,000 amino acids	Yes	Yes	No
SMMPMBEC		No	Matrix	Yes	≤200 sequences or ≤10 MB	Yes	Yes	No
IEDB-AR-Consensus		No	Matrix and NN	Yes	≤200 sequences or ≤10 MB	Yes	Yes	No
NetMHCcons 1.1		No	Matrix and NN	Yes	≤5000 sequences and ≤20,000 amino acids	Yes	Yes	No
NetMHC 4.0	Sequence-based features	No	NN	Yes	≤5000 sequences and ≤20,000 amino acids	Yes	Yes	No
NetMHCstabpan 1.0	Physicochemical features	No	NN	Yes	≤5000 sequences and ≤20,000 amino acids	Yes	Yes	No
NetMHCpan 4.1	Sequence-based features and binary features	No	NN	Yes	≤5000 sequences and ≤20,000 amino acids	Yes	Yes	No
ConvMHC	Sequence-based features and physicochemical features	No	NN	No	NM	No	No	No
Anthem	AAF, WLS, PSSM, PWM, BLOSUM 62	Yes	AODE	Yes	≤5000 sequences and ≤20,000 amino acids	Yes	Yes	Yes

[a]The URL addresses for accessing the listed prediction tools are as follows: SYFPEITHI, http://www.syfpeithi.de/index.html/; RANKPEP, http://imed.med.ucm.es/Tools/rankpep.html/; PickPocket 1.1, http://www.cbs.dtu.dk/services/PickPocket/; SMMPMBEC, https://github.com/ykimbiology/smmpmbec/; IEDB-AR-Consensus, http://tools.iedb.org/mhci/; NetMHCcons-1.1, http://www.cbs.dtu.dk/services/NetMHCcons/; NetMHC4.0, http://www.cbs.dtu.dk/services/NetMHC/; NetMHCstabpan 1.0, http://www.cbs.dtu.dk/services/NetMHCstabpan/; NetMHCPan-4.1, http://www.cbs.dtu.dk/services/NetMHCpan/; ConvMHC, http://jumong.kaist.ac.kr:8080/convmhc/ and Anthem, http://anthem.erc.monash.edu/.
[b]Abbreviations: WLS, WebLogo-based sequence conservation; NM, not mentioned.
[c]Yes: The publication has provided the mentioned function; No: The publication has not provided the mentioned function.

Research Cloud, an initiative of the Australian Government's Super Science Scheme and the Education Investment Fund.

Funding

This work was supported by grants from the National Health and Medical Research Council of Australia (NHMRC) (1165490, 1144652 and 1127948), a Major Inter-Disciplinary Research project awarded by Monash University, a strategic research award from the Juvenile Diabetes Research Foundation (1-SRA-2019-806-S-B) and the Collaborative Research Program of Institute for Chemical Research, Kyoto University (2019-32 and 2018-28). A.W.P. is supported by an NHMRC Principal Research Fellowship (1137739).

References

1. Lundegaard C, Lund O, Buus S, *et al.* Major histocompatibility complex class I binding predictions as a tool in epitope discovery. *Immunology* 2010;**130**:309–18.

2. Purcell AW, Ramarathinam SH, Ternette N. Mass spectrometry–based identification of MHC-bound peptides for immunopeptidomics. *Nat Protoc* 2019;**14**:1687.
3. Ramarathinam SH, Croft NP, Illing PT, *et al*. Employing proteomics in the study of antigen presentation: an update. *Expert Rev Proteomics* 2018;**15**:637–45.
4. Zhang L, Udaka K, Mamitsuka H, *et al*. Toward more accurate pan-specific MHC-peptide binding prediction: a review of current methods and tools. *Brief Bioinform* 2011;**13**:350–64.
5. Mei S, Li F, Leier A, *et al*. A comprehensive review and performance evaluation of bioinformatics tools for HLA class I peptide-binding prediction. *Brief Bioinform* 2020;**21**:1119–35.
6. Gfeller D, Bassani-Sternberg M. Predicting antigen presentation–what could we learn from a million peptides? *Front Immunol* 2018;**9**:1716.
7. Zhang H, Lund O, Nielsen M. The PickPocket method for predicting binding specificities for receptors based on receptor pocket similarities: application to MHC-peptide binding. *Bioinformatics* 2009;**25**:1293–9.
8. Liu G, Li D, Li Z, *et al*. PSSMHCpan: a novel PSSM-based software for predicting class I peptide-HLA binding affinity. *GigaScience* 2017;**6**:1–11.
9. Pietrokovski S, Henikoff JG, Henikoff S. The blocks database—a system for protein classification. *Nucleic Acids Res* 1996;**24**:197–200.
10. 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**:449–54.
11. Andreatta M, Nielsen M. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics* 2015;**32**:511–7.
12. O'Donnell TJ, Rubinsteyn A, Bonsack M, *et al*. MHCflurry: open-source class I MHC binding affinity prediction. *Cell systems* 2018;**7**:129–132. e124.
13. Liu Z, Cui Y, Xiong Z, *et al*. DeepSeqPan, a novel deep convolutional neural network model for pan-specific class I HLA-peptide binding affinity prediction. *Sci Rep* 2019;**9**:794.
14. Phloyphisut P, Pornputtapong N, Sriswasdi S, *et al*. MHCSeqNet: a deep neural network model for universal MHC binding prediction. *BMC Bioinformatics* 2019;**20**:270.
15. Poernomo A, Kang D-K. Biased dropout and crossmap dropout: learning towards effective dropout regularization in convolutional neural network. *Neural Netw* 2018;**104**:60–7.
16. Mommen GP, Frese CK, Meiring HD, *et al*. Expanding the detectable HLA peptide repertoire using electron-transfer/higher-energy collision dissociation (EThcD). *Proc Natl Acad Sci* 2014;**111**:4507–12.
17. Liepe J, Marino F, Sidney J, *et al*. A large fraction of HLA class I ligands are proteasome-generated spliced peptides. *Science* 2016;**354**:354–8.
18. Caron E, Kowalewski DJ, Koh CC, *et al*. Analysis of major histocompatibility complex (MHC) immunopeptidomes using mass spectrometry. *Mol Cell Proteomics* 2015;**14**:3105–17.
19. Yu Q, Wang B, Chen Z, *et al*. Electron-transfer/higher-energy collision dissociation (EThcD)-enabled intact glycopeptide/glycoproteome characterization. *J Am Soc Mass Spectrom* 2017;**28**:1751–64.
20. Chong C, Marino F, Pak H, *et al*. High-throughput and sensitive immunopeptidomics platform reveals profound interferonγ-mediated remodeling of the human leukocyte antigen (HLA) ligandome. *Mol Cell Proteomics* 2018;**17**:533–48.
21. Bassani-Sternberg M, Chong C, Guillaume P, *et al*. Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allostery regulating HLA specificity. *PLoS Comput Biol* 2017;**13**:e1005725.
22. Zhang Y-H, Xing Z, Liu C, *et al*. Identification of the core regulators of the HLA I-peptide binding process. *Sci Rep* 2017;**7**:42768–78.
23. Jurtz V, Paul S, Andreatta M, *et al*. NetMHCpan-4.0: improved peptide–MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. *J Immunol* 2017;**199**:3360–8.
24. Rasmussen M, Fenoy E, Harndahl M, *et al*. Pan-specific prediction of peptide–MHC class I complex stability, a correlate of T cell immunogenicity. *J Immunol* 2016;**197**:1517–24.
25. Hu Y, Wang Z, Hu H, *et al*. ACME: pan-specific peptide–MHC class I binding prediction through attention-based deep neural networks. *Bioinformatics* 2019. doi: 10.1093/bioinformatics/btz427.
26. Ling CX, Huang J, Zhang H. AUC: a better measure than accuracy in comparing learning algorithms. In: *Conference of the canadian society for computational studies of intelligence*. Springer, 2003, 329–41.
27. Kohavi R, John GH. Wrappers for feature subset selection. *Artificial Intelligence* 1997;**97**:273–324.
28. Gfeller D, Guillaume P, Michaux J, *et al*. The length distribution and multiple specificity of naturally presented HLA-I ligands. *J Immunol* 2018;**201**:3705–16.
29. Karosiene E, Lundegaard C, Lund O, *et al*. NetMHCcons: a consensus method for the major histocompatibility complex class I predictions. *Immunogenetics* 2012;**64**:177–86.
30. Dhanda SK, Mahajan S, Paul S, *et al*. IEDB-AR: immune epitope database—analysis resource in 2019. *Nucleic Acids Res* 2019;**47**:502–6.
31. Reche PA, Zhang H, Glutting J-P, *et al*. EPIMHC: a curated database of MHC-binding peptides for customized computational vaccinology. *Bioinformatics* 2005;**21**:2140–1.
32. Lata S, Bhasin M, Raghava GP. MHCBN 4.0: a database of MHC/TAP binding peptides and T-cell epitopes. *BMC Res Notes* 2009;**2**:61.
33. Rammensee H-G, Bachmann J, Emmerich NPN, *et al*. SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics* 1999;**50**:213–9.
34. Hassan C, Chabrol E, Jahn L, *et al*. Naturally processed non-canonical HLA-A* 02: 01 presented peptides. *J Biol Chem* 2015;**290**:2593–603.
35. Marcilla M, Alpízar A, Lombardía M, *et al*. Increased diversity of the HLA-B40 ligandome by the presentation of peptides phosphorylated at their main anchor residue. *Mol Cell Proteomics* 2014;**13**:462–74.
36. Mobbs JI, Illing PT, Dudek NL, *et al*. The molecular basis for peptide repertoire selection in the human leukocyte antigen (HLA) C* 06: 02 molecule. *J Biol Chem* 2017;**292**:17203–15.
37. Yair-Sabag S, Tedeschi V, Vitulano C, *et al*. The peptide repertoire of HLA-B27 may include ligands with lysine at P2 anchor position. *Proteomics* 2018;**18**:1700249.
38. Müller M, Gfeller D, Coukos G, *et al*. 'Hotspots' of antigen presentation revealed by human leukocyte antigen ligandomics for neoantigen prioritization. *Front Immunol* 2017;**8**: 1367.
39. 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.

40. Kalaora S, Barnea E, Merhavi-Shoham E, *et al.* Use of HLA peptidomics and whole exome sequencing to identify human immunogenic neo-antigens. *Oncotarget* 2016;**7**:5110.
41. Ternette N, Purcell AW. Immunopeptidomics special issue. *Proteomics* 2018;**18**:1–4.
42. Schellens IM, Hoof I, Meiring HD, *et al.* Comprehensive analysis of the naturally processed peptide repertoire: differences between HLA-A and B in the immunopeptidome. *PloS One* 2015;**10**:e0136417.
43. 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.
44. Schittenhelm RB, Sian TCLK, Wilmann PG, *et al.* Revisiting the arthritogenic peptide theory: quantitative not qualitative changes in the peptide repertoire of HLA–B27 allotypes. *Arthritis & rheumatology* 2015;**67**:702–13.
45. Illing PT, Pymm P, Croft NP, *et al.* HLA-B57 micropolymorphism defines the sequence and conformational breadth of the immunopeptidome. *Nat Commun* 2018;**9**:4693.
46. Marcilla M, Alvarez IA, Ramos-Fernández A, *et al.* Comparative analysis of the endogenous peptidomes displayed by HLA-B* 27 and Mamu-B* 08: two MHC class I alleles associated with elite control of HIV/SIV infection. *J Proteome Res* 2016;**15**:1059–69.
47. Hillen N, Mester G, Lemmel C, *et al.* Essential differences in ligand presentation and T cell epitope recognition among HLA molecules of the HLA-B44 supertype. *Eur J Immunol* 2008;**38**:2993–3003.
48. Kaur G, Gras S, Mobbs JI, *et al.* Structural and regulatory diversity shape HLA-C protein expression levels. *Nat Commun* 2017;**8**:1–12.
49. Boehm KM, Bhinder B, Raja VJ, *et al.* Predicting peptide presentation by major histocompatibility complex class I: an improved machine learning approach to the immunopeptidome. *BMC Bioinformatics* 2019;**20**:1–11.
50. Alvarez B, Reynisson B, Barra C, *et al.* NNAlign_MA; MHC Peptidome Deconvolution for accurate MHC binding motif characterization and improved T-cell epitope predictions. *Mol Cell Proteomics* 2019;**18**:2459–77.
51. Stranzl T, Larsen MV, Lundegaard C, *et al.* NetCTLpan: pan-specific MHC class I pathway epitope predictions. *Immunogenetics* 2010;**62**:357–68.
52. Vang YS, Xie X. HLA class I binding prediction via convolutional neural networks. *Bioinformatics* 2017;**33**:2658–65.
53. Nielsen M, Andreatta M. NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. *Genome Med* 2016;**8**:33.
54. Han Y, Kim D. Deep convolutional neural networks for pan-specific peptide-MHC class I binding prediction. *BMC Bioinformatics* 2017;**18**:585.
55. Singh H, Raghava G. ProPred1: prediction of promiscuous MHC class-I binding sites. *Bioinformatics* 2003;**19**:1009–14.
56. Shao XM, Bhattacharya R, Huang J, *et al.* High-throughput prediction of MHC class I and class II neoantigens with MHCnuggets. *Cancer Immunol Res* 2020;**8**:396–408.
57. Neefjes J, Jongsma ML, Paul P, *et al.* Towards a systems understanding of MHC class I and MHC class II antigen presentation. *Nat Rev Immunol* 2011;**11**:823–36.
58. Mei S, Ayala R, Ramarathinam SH, *et al.* Immunopeptidomic analysis reveals that deamidated HLA-bound peptides arise predominantly from deglycosylated precursors. *Mol Cell Proteomics* 2020;**19**:1236–47.
59. Storkus W, Howell D, Salter R, *et al.* NK susceptibility varies inversely with target cell class I HLA antigen expression. *J Immunol* 1987;**138**:1657–9.
60. Zemmour J, Little A, Schendel D, *et al.* The HLA-A, B" negative" mutant cell line C1R expresses a novel HLA-B35 allele, which also has a point mutation in the translation initiation codon. *J Immunol* 1992;**148**:1941–8.
61. Giam K, Ayala-Perez R, Illing P, *et al.* A comprehensive analysis of peptides presented by HLA-A1. *Tissue Antigens* 2015;**85**:492–6.
62. Perez-Riverol Y, Csordas A, Bai J, *et al.* The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res* 2018;**47**: 442–50.
63. Thomson PJ, Illing PT, Farrell J, *et al.* Modification of the cyclopropyl moiety of abacavir provides insight into the structure activity relationship between HLA-B* 57: 01 binding and T-cell activation. *Allergy* 2020;**75**:636–47.
64. Song J, Li F, Leier A, *et al.* PROSPERous: high-throughput prediction of substrate cleavage sites for 90 proteases with improved accuracy. *Bioinformatics* 2017;**34**:684–7.
65. Chen Z, Zhao P, Li F, *et al.* iFeature: a python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics* 2018;**34**:2499–502.
66. Li F, Li C, Marquez-Lago TT, *et al.* Quokka: a comprehensive tool for rapid and accurate prediction of kinase family-specific phosphorylation sites in the human proteome. *Bioinformatics* 2018;**34**:4223–31.
67. Crooks GE, Hon G, Chandonia J-M, *et al.* WebLogo: a sequence logo generator. *Genome Res* 2004;**14**:1188–90.
68. Jiang N, Cui J, Meng J, *et al.* A tomato nucleotide binding sites– leucine-rich repeat gene is positively involved in plant resistance to phytophthora infestans. *Phytopathology* 2018;**108**:980–7.
69. Consortium TU. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 2017;**45**:158–69.
70. Webb GI, Boughton JR, Wang Z. Not so naive Bayes: aggregating one-dependence estimators. *Machine learning* 2005;**58**:5–24.
71. Li F, Song J, Li C, *et al.* PAnDE: averaged n-dependence estimators for positive unlabeled learning. *ICIC Express Letters, Part B: Applications* 2017;**8**:1287–97.
72. Li F, Zhang Y, Purcell AW, *et al.* Positive-unlabelled learning of glycosylation sites in the human proteome. *BMC Bioinformatics* 2019;**20**:112.
73. Wong KC, Chen J, Zhang J, *et al.* Early cancer detection from multianalyte blood test results. *iScience* 2019;**15**:332–41.
74. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, 785–94.
75. Bhasin M, Raghava G. SVM based method for predicting HLA-DRB1* 0401 binding peptides in an antigen sequence. *Bioinformatics* 2004;**20**:421–3.
76. Safavian SR, Landgrebe D. A survey of decision tree classifier methodology. *IEEE Trans Syst Man Cybern* 1991;**21**:660–74.
77. Marino SR, Lin S, Maiers M, *et al.* Identification by random forest method of HLA class I amino acid substitutions associated with lower survival at day 100 in unrelated donor hematopoietic cell transplantation. *Bone Marrow Transplant* 2012;**47**:217–26.
78. Huang L, Karpenko O, Murugan N, *et al.* A meta-predictor for MHC class II binding peptides based on naive Bayesian approach. In: *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2006, 5322–5.

79. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure* 1975;**405**:442–51.
80. Aranha MP, Spooner C, Demerdash O, *et al*. Prediction of peptide binding to MHC using machine learning with sequence and structure-based feature sets. *Biochimica et Biophysica Acta (BBA)-General Subjects* 2020;129535.
81. Bonsack M, Hoppe S, Winter J, *et al*. Performance evaluation of MHC class-I binding prediction tools based on an experimentally validated MHC-peptide binding dataset. *Cancer Immunol Res* 2019;**7**:719–36.
82. Prazsák I, Tombácz D, Szűcs A, *et al*. Full genome sequence of the western reserve strain of vaccinia virus determined by third-generation sequencing. *Genome Announc* 2018;**6**.
83. Benson DA, Cavanaugh M, Clark K, *et al*. GenBank. *Nucleic Acids Res* 2012;**41**:D36–42.
84. Illing PT, Vivian JP, Dudek NL, *et al*. Immune self-reactivity triggered by drug-modified HLA-peptide repertoire. *Nature* 2012;**486**:554–8.
85. Reche PA, Glutting J-P, Reinherz EL. Prediction of MHC class I binding peptides using profile motifs. *Hum Immunol* 2002;**63**:701–9.
86. Kim Y, Ponomarenko J, Zhu Z, *et al*. Immune epitope database analysis resource. *Nucleic Acids Res* 2012;**40**:525–30.
87. Kim Y, Sidney J, Pinilla C, *et al*. Derivation of an amino acid similarity matrix for peptide: MHC binding and its application as a Bayesian prior. *BMC Bioinformatics* 2009;**10**: 394–404.
88. Faridi P, Li C, Ramarathinam SH, *et al*. A subset of HLA-I peptides are not genomically templated: evidence for cis-and trans-spliced peptide ligands. *Science Immunology* 2018;**3**:3947–58.
89. Faridi P, Woods K, Ostrouska S, *et al*. Spliced peptides and cytokine driven changes in the immunopeptidome of melanoma. *Cancer Immunol Res* 2020;**8**:1322–34.
90. Li F, Chen J, Leier A, *et al*. DeepCleave: a deep learning predictor for caspase and matrix metalloprotease substrates and cleavage sites. *Bioinformatics* 2020;**36**:1057–65.
91. Liu Q, Chen J, Wang Y, *et al*. DeepTorrent: a deep learning-based approach for predicting DNA N4-methylcytosine sites. *Brief Bioinform* 2020. doi: 10.1093/bib/bbaa124.
92. Jia C, Bi Y, Chen J, *et al*. PASSION: an ensemble neural network approach for identifying the binding sites of RBPs on circRNAs. *Bioinformatics* 2020;**36**:4276–82.
93. Li F, Chen J, Ge Z, *et al*. Computational prediction and interpretation of both general and specific types of promoters in Escherichia coli by exploiting a stacked ensemble-learning framework. *Brief Bioinform* 2020. doi: 10.1093/bib/bbaa049.