OXFORD

Briefings in Bioinformatics, 2023, 1–9
https://doi.org/10.1093/bib/bbad150
Review

A comprehensive assessment and comparison of tools for HLA class I peptide-binding prediction

Meng Wang, Lukasz Kurgan and Min Li

Corresponding author. Lukasz Kurgan, Department of Computer Science, Virginia Commonwealth University, Richmond, VA 23284, USA. E-mail: lkurgan@vcu.edu; Min Li, School of Computer Science and Engineering, Central South University, Changsha 410083, China. E-mail: limin@mail.csu.edu.cn

Abstract

Human leukocyte antigen class I (HLA-I) molecules bind intracellular peptides produced by protein hydrolysis and present them to the T cells for immune recognition and response. Prediction of peptides that bind HLA-I molecules is very important in immunotherapy. A growing number of computational predictors have been developed in recent years. We survey a comprehensive collection of 27 tools focusing on their input and output data characteristics, key aspects of the underlying predictive models and their availability. Moreover, we evaluate predictive performance for eight representative predictors. We consider a wide spectrum of relevant aspects including allele-specific analysis, influence of negative to positive data ratios and runtime. We also curate high-quality benchmark datasets based on analysis of the consistency of the data labels. Results reveal that each considered method provides accurate results, which can be explained by our analysis that finds that their predictive models capture meaningful binding motifs. Although some methods are overall more accurate than others, we find that none of them is universally superior. We provide a comprehensive comparison of the convenience as well as the accuracy of the methods under specific prediction scenarios, such as for specific alleles, metrics of predictive performance and constraints on runtime. Our systematic and broad analysis provides informative clues to the users to identify the most suitable tools for a given prediction scenario and for the developers to design future methods.

Keywords: HLA-peptide, binding prediction, tools comparison

INTRODUCTION

Major histocompatibility complex (MHC) is an important gene group that drives the vertebrate immune system. These genes facilitate the detection and recognition of foreign threats, resulting in a series of immune responses. In recent years immunotherapy has become a promising approach to cancer treatment, especially since its adverse effects are far lower than those of chemotherapy or radiotherapy [1, 2]. This approach exploits the fact that cancer cells produce unique neoepitopes that are recognized by MHC [3, 4]. This is one of many factors that motivate research toward improving our understanding of mechanisms of peptide presentation in immune processes.

MHC molecules are divided into three subtypes: class I, class II and class III. We focus on the class I MHC molecules (MHC-I), which present endogenous peptides to CD8 + T cells [5]. MHC-I are heterodimers consisting of a heavy α chain and a light β chain where the interaction with the peptide occurs in the α chain. The human MHC is called the human leukocyte antigen (HLA) and research shows that HLA-I molecules primarily bind peptides that are 8–12 amino acids in length [6].

Given the high promiscuity and importance of these interactions to the immune response, vaccinology and immunotherapy research benefits from predictions of peptide-HLA-I binding [7, 8]. The efforts that produce these predictive tools benefit from the availability of experimentally verified peptide ligand databases, such as the Immune Epitope Database (IEDB) [9]. The recent rapid increase in the amount of these data, in part due to the use the mass spectrometry-based methods [10, 11], resulted in the development of many new computational predictors. These methods primarily target the prediction of the HLA-I peptides since peptides interacting with HLA-II are longer and more complex, making their prediction more challenging [12]. In particular, recent years have witnessed the application of modern machine learning methods, including deep neural networks, to develop even more accurate predictors of HLA-I peptides [13–16]. At a coarse-grained level, these tools are categorized into two groups: structure-based methods that rely on the protein and peptide structures [17–19] and sequence-based methods that make predictions solely from the peptide sequence [13–16, 20, 21]. We focus on the latter type of method since the structure-based approaches are more computationally demanding and limited to the peptides/proteins with known structures.

Several surveys of the sequence-based predictors were published in recent years [22–26]. However, they miss the most recent tools, some provide a rather superficial description of the predictors, and they present an empirical comparative assessment that

Meng Wang received his BS and MS degrees in safety engineering from the Central South University, Changsha, China, in 2020. Currently, he is working toward a PhD degree in computer science and technology at the Central South University, Changsha, China. His current research interests include bioinformatics and protein–peptide interactions.

Lukasz Kurgan received his MSc degree (with honors) in Automation and Robotics from the AGH University of Science and Technology, Poland, in 1999 and a PhD degree in Computer Science from the University of Colorado at Boulder, USA, in 2003. He is currently an Endowed Professor of Computer Science at the Virginia Commonwealth University, USA. He is a Fellow of AIMBE and AAIA, and a member of the European Academy of Sciences and Arts. His main research interests include high-throughput structural bioinformatics of proteins and small RNAs. More details are on the website of his lab at http://biomine.cs.vcu.edu/

Min Li received her BS degree in Communication Engineering and MS and PhD degrees in Computer Science from the Central South University, Changsha, China, in 2001, 2004 and 2008, respectively. She is currently a Professor at the School of Computer Science and Engineering at the Central South University. Her main research interests include bioinformatics and system biology.

Received: December 28, 2022. **Revised:** March 27, 2023. **Accepted:** March 29, 2023

© The Author(s) 2023. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbad150/7126340 by Virginia Commonwealth University Libraries user on 21 April 2023

is limited in scope. For instance, the most recent and broadest to-date survey covers 15 predictors and provides a relatively rudimentary evaluation of predictive performance [22]. We cover a more comprehensive collection of 27 predictors, which includes a new and significantly improved version of the popular NetMHCpan, NetMHCpan-4.1 [27] and several other recently released methods, such as MHCflurry 2.0 [28], MHCSeqNet [29] and ACME [30]. We discuss their inputs, predictive models, outputs and availability to comprehensively cover their key characteristics, limitations and advantages. Moreover, we present arguably the most well-rounded empirical assessment for a collection of eight representative predictors. This assessment (i) analyses consistency between two types of ligand labels: qualitative and quantitative; (ii) reports a broad set of metrics of predictive performance (overall performance/performance limited by allele type and peptide length): the area under the receiver operating characteristic curve (AUC), the area under the curve of precision-recall (PR) curve (AUPR) and Spearman's rank correlation coefficient (SRCC); (iii) provides a multifaceted analysis of key aspects that influence the comparison of performance under different positive and negative sample ratios; (iv) analyzes the ability of different predictors on the capability of binding motifs and (v) evaluates the sensitivity of data volume and running speed. The unparallel breadth and depth of this study provide useful insights for both the users of the current methods and the developers of the future predictors.

SURVEY OF CURRENT PREDICTORS

We cover a comprehensive collection of 27 sequence-based predictors of the HLA-I peptides (Table S1). These methods rely on models derived from training data, primarily using machine learning algorithms for that purpose. We discuss three key aspects of the design and implementation of these models including the collection and annotation of the training datasets, approaches to encode the raw inputs into feature-based representations that can be processed by the machine learning models, and algorithms that are used to produce these models (Table S1). We also summarize several other important factors related to their availability and formulation and restrictions on the inputs that they use and outputs that they produce.

Training datasets

The majority of the predictors depend on the training data extracted from IEDB [9, 31]. IEDB is the largest public resource for HLA ligands and T cell epitopes and primarily relies on the epitope data harvested from the PubMed database.

One of the consequential decisions that designers of these predictors make is the selection of the predicted values. At present, two common choices are the quantitative affinity score and the qualitative labels; see the 'Outputs' column in Table S1. The quantitative score denotes the interaction affinity between epitopes and alleles. The binding affinity is real-valued where smaller values imply a greater likelihood of binding. These values can be binarized using a threshold value, which means that samples with affinity < threshold are regarded as 'binding'. Interestingly, threshold values could be different for different alleles, although 500 nm is also used as the universal threshold across alleles [13, 32]. The use of the binding affinity value could be problematic since it measures only the peptide–MHC binding while neglecting other biological features of the underlying antigen presentation process. An alternative is the qualitative mass spectrometry (MS) derived eluted ligand (EL) based label. This label covers the binding event and prior steps in the processes, making it arguably more reliable than the binding affinity label. The MS EL label has five values in IEDB: positive, positive-high, positive-intermediate, positive-low and negative. Some methods differ in how they process the positive-intermediate and positive-low labels and in some instances, these labels are even deleted [29]. Although the majority of the predictors consider one of the label types, recent literature [28, 33] reveals that some of the more recent methods cover both types of predictions.

Another important aspect of the training datasets is the rate of binding (positive) and nonbinding (negative) samples, which is often heavily imbalanced. We study the rate based on the 11 907 epitopes that we extracted from IEDB. We find 28 alleles for which the total number of positive and negative samples is below 50, which renders the generation of reliable predictive models virtually impossible. Among the remaining alleles, we find extremely imbalanced data for 28 alleles. The number of five where the number of positive samples is over five times higher than the number of negative samples and 23 where the number of negatives is over five times higher than positives. The remaining 51 alleles include a more balanced distribution of binding and nonbinding samples where the difference is less than five folds. Given these extreme differences in the rate of binding to nonbinding samples, some studies use randomly generated peptides to maintain a consistent negative to positive rate [34, 35]. Typically, these peptides are generated from the nonbinding regions of the proteins that have the binding peptides [22, 36, 37].

Finally, as the peptides that are presented to HLA type I molecule are typically in the range of 8 to 11 residues long [33], the current tools limit their predictions to the peptides in this length range [22]; see the 'Inputs' column in Table S1.

Encoding of predictive inputs

One of the main differences between predictors is how they encode their predictive inputs (Table S1). The underlying challenge is to convert the raw inputs that have variable lengths (i.e. sequences of the epitope and HLA-I allele) into fixed-size feature vectors. This is necessary since virtually all machine learning algorithms require fixed-size inputs. The HLA-I allele sequences are typically collected from the immuno polymorphism database [38]. Although binding of the allele sequence with the epitope occurs in a specific site on the allele, this information is typically unavailable. A common way to circumvent this issue is to encode features based on key sites selected based on prior works that include (i) positions {7, 9, 24, 45, 59, 62, 63, 66, 67, 69, 70, 73, 74, 76, 77, 80, 81, 84, 95, 97, 99, 114, 116, 118, 143, 147, 150, 152, 156, 158, 159, 163, 167, 171} [39]; (ii) positions {7, 9, 13, 24, 31, 45, 59, 62, 63, 65, 66, 67, 69, 70, 71, 73, 74, 76, 77, 80, 81, 84, 95, 97, 99, 110, 114, 116, 118, 138, 143, 147, 150, 152, 156, 158, 159, 163, 167, 171} [40] and (iii) specific regions on two alpha helices of allele, calculated by MUSCLE [41, 42]. Although most of the methods generate input features from the sequences of peptides and alleles, some add other relevant biological information [40]. One drawback of the latter tools is that they require these inputs to make predictions, which limits their applications to scenarios where this information is available.

Two common ways to derive features from sequences are to use position-specific encoding and sequence context-based encoding; see the 'Peptide encoding' column in Table S1. The former method uses fixed-size vectors (typically 20-dimensional) to represent amino acids at each position in the input sequence. There are several widely used ways to produce this encoding: one-hot, BLOSUM50, BLOSUM62 and position-specific scoring matrix (PSSM). The one-hot approach applies a unit binary vector, e.g.

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbad150/7126340 by Virginia Commonwealth University Libraries user on 21 April 2023

[1, 0, 0, ..., 0] for Alanine, [0, 1, 0, ..., 0] for Arginine, where the position of 1 in this vector denotes the amino acid type. The length of this vector equals the number of the considered amino acid types, which could be 20 or more if special types, such as X, are covered. BLOSUM*n* is a matrix in which every position describes the similarity between a pair of residues, where parameter *n* is the similarity threshold. The PSSM is calculated from multi-sequence alignments against a large library of sequences, typically generated using PSI-BLAST [43]. Elements of this matrix denote the frequency of each amino acid type at a given query sequence position in the alignment and can be used to reflect evolutionary conservation. The other common approach is sequence context-based encoding, often called embedding [44]. This approach originated in the natural language processing area. The embedding model captures relationships between a residue at a certain position in the sequence and its surrounding residues in the same sequence.

Predictive models

The predictive models can be categorized in two ways, based on the granularity of the models (allele-specific versus pan-specific) and based on the type of approach/algorithm used to derive the model; see the 'Algorithm' column in Table S1.

Table S1 shows that seven current methods are allele-specific while 20 are pan-specific. As the name suggests, the allele-specific approach is a collection of models that are specialized to predict specific alleles [13, 45, 46]. Given the uneven distribution of the per-allele data (see Training datasets), these methods often cluster alleles together and train models for an allele with the largest amount of binding peptide data in each cluster. Next, they transfer these models to other alleles in this cluster based on the transfer learning concept [13, 47]. The transfer learning takes a pre-computed model (trained on the most common allele in a cluster) and uses it as a starting point to specialize it to another allele from that cluster using the limited amount of data for that allele. The allele-specific models can potentially provide strong predictive performance, depending on the amount of per-allele training data and the quality of the clustering of alleles. They perform particularly well for alleles that have a large and representative set of training peptides. However, results for small sample size alleles from clusters of alleles that do not share sufficient mutual similarity inevitably result in using a wrong starting point to perform the transfer learning. This will ultimately harm the predictive performance of the resulting models for these small sample-size alleles. Moreover, this approach may not be able to establish models for all alleles, particularly for those which are poorly represented (or in extreme cases unseen) in the training dataset. On the other hand, the pan-specific methods utilize a single predictive model that is derived from the binding peptides across different alleles [14, 16, 48]. They provide predictions for all alleles that have sequence information and learn the relationship between data features inside the predictive models without relying on clustering. Neither of the two approaches was established as universally more accurate.

Some of the earlier predictors rely on rather simple probabilistic models (scoring functions), which are based on position specificity [45, 49]. They typically produce a prediction score for a given peptide by adding or averaging scores computed for the individual amino acids at each considered position. A more recent trend is to derive predictive models using machine learning algorithms. In the latter case, the model is learned from training data by a given machine learning algorithm by optimizing the fit of the model-produced outputs to the known labels. Table S1 reveals that the dominant type of machine learning algorithm used is the neural network. A large variety of neural networks were utilized, from classical feed-forward networks [47, 50] to modern deep convolutional networks [15, 16, 48]. For instance, the multiple versions of the widely used NetMHCpan use a rather simple shallow feed-forward neural network with one hidden layer [27, 51–53]. However, other recent tools, such as Seq2Neo [54], MHCnuggets [13] and ACME [30] rely on more sophisticated deep convolutional topologies. Another interesting approach to designing predictive models is consensus predictors. They combine results produced by multiple input predictors of the HLA-I peptides to provide improved predictive performance when compared with the input methods. The underlying principle is that the input predictors produce complementary results, where this complementarity can be exploited to fix some of the mistakes that they make individually. An early example is NetMHCcons [55], which combines predictions generated by three predictors: NetMHC 3.4 [56], NetMHCpan 2.8 [51] and Pickpocket 1.1 [57].

Availability, inputs and outputs

Important factors that may contribute to a higher user uptake of these predictors are the public availability of the underlying predictive models, formulation and potential restrictions on the inputs and the format and scope of the outputs.

Although virtually all tools are available as standalone software, only some can be used as webservers; see the 'Software' and 'Webserver' columns in Table S1. Examples of the webserver-available methods include PickPocket, netMHCstabpan and the most recent version of NetMHCpan. Webservers allow the end users to run the predictions on the server side (using the developer's hardware) and free them from installation and programming. Potential limitations of the web servers include the inability to run batch jobs (i.e. process multiple predictions in a single run) and limited throughput. The standalone software, which has to be installed and run on the user's hardware, is arguably more suitable for computer-savvy users and those who would like to incorporate these methods into broader bioinformatics pipelines. Some software does not provide a user interface, forcing the users to familiarize themselves with command-line coding. This difficulty is often compounded by the requirements to possess specific hardware and operating systems to install and run the software.

The other significant aspect is the restrictions on the inputs and the format and scope of the outputs (Table S1) [58–61]. Most of the predictors accept the allele name-peptide pairs as inputs, whereas some also accept the allele name-protein sequence pairs. The pan-specific tools may take the allele sequence-peptide pairs, without specifically being required to identify the allele. A couple of notable observations include MHCflurry, which accepts inputs solely from the terminal (not in a file), and MixMHCpred that can be used to predict input 8- to 14-mer peptides. The types of outputs are determined by the labels used to derive the underlying predictive model (see Training datasets).

COMPARATIVE ASSESSMENT

We perform an empirical comparative assessment for a carefully selected set of eight representative predictors. We cover two popular/highly-cited older tools: PickPocket [57] and netMHCstabpan [62]; MixMHCpred [63], which is the best-performing predictor from the most recent comparative survey [22]; and five latest tools: MHCflurry 2.0 [28], MATHLA [48], MHCSeqNet [29], MHCflurry 2.0 [28] and NetMHCpan 4.1 [27]. The latter is a significantly improved version of NetMHCpan-4.0, another tool that ranked particularly well in a recent survey [22].

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbad150/7126340 by Virginia Commonwealth University Libraries user on 21 April 2023

Validation dataset

Before assessing the selected tools, we evaluate the consistency between the binding affinity scores and the MS EL labels. Firstly, we utilize commonly applied selection criteria to filter high-quality data from IEDB (downloaded on 24 May 2022): (i) HLA-I alleles that are of HLA-A, B and C subtypes; (ii) peptides with 8 to 14 amino acids in length; and (iii) peptides with both affinity scores and MS labels. Next, the origin dataset extraction strategy is applied for ligand: the single allele data (SA, where each peptide is associated with a single MHC restriction) and the multi allele data (MA, where each peptide has multiple options for MHC restriction). As we explain in Training datasets, we mark peptides with an affinity <500 nm as positive samples; the remaining peptides are considered negative samples. The MS labels are taken as the reference, and we compare the affinity classifications against them, i.e. we use affinity labels to 'predict' MS labels. The affinity-based annotations predict an AUC of about 0.94 for MS labels across different selections of peptides. This analysis highlights the need for careful curation of the labels to avoid scenarios where models are trained and/or tested using low-quality labels, which inevitably would lead to the generation of substandard quality models.

We use the 'SA + MA' datasets to perform a comparative assessment of the eight predictors. Moreover, we further curate these data to minimize inconsistencies between the affinity-based and the MS-based annotations of labels. In particular, we solve conflicts between annotations of the same type by majority vote, and we remove samples that have inconsistent annotations between the affinity and the MS labels. The final curated origin dataset (origin dataset) consists of 228 711 entries with 225 523 positive entries. To simulate the real scene (negative samples are more than positive samples), we filled in the original dataset from the random peptide library with a ratio of 1:5 ('f5' dataset) as the test dataset below. Here are the processing steps of the f5 dataset: (i) first, randomly select proteins from the proteome and cut them according to different lengths to form a random peptide library [64] with different lengths (8–14); (ii) grouping peptides of the same allele and the same length; (iii) in each peptide group, if the negative samples are more than five times the number of positive samples, randomly select negative samples that are five times the number of positive samples, and form the 'f5' dataset of this group together with positive samples. If the negative sample is less than five times the positive sample, randomly select peptide segments from the random peptide library of corresponding length, so that the number of negative samples is five times of positive sample, forming the 'f5' dataset of this group; and (iv), finally, different groups of 'f5' datasets are combined to form the 'f5' dataset in this review.

Overall performance on different datasets

First of all, we test the performance of these eight tools on the 'f5' dataset. The classification performance of each tool is measured using AUC and AUPR. What's more, we use SRCC to evaluate the ranking consistency of predicted binding possibilities with their true affinity labels. The prediction results are shown in Figure 1. The AUC, AUPR and SRCC of MHCflurry are all the highest, followed by NetMHCpan and MATHLA. Except for MHCSeqNet, other tools have shown good prediction results.

Performance limited by allele type and peptide length

To observe the effect of the allelic subtype restriction conditions on the results, we extract 151 subsets based on allele subtypes. Figure S1 shows the number of subsets that perform best for each tool. As can be seen from Figure S1, MHCflurry is still the best-performing tool, followed by NetMHCpan and MATHLA. NetMHCstabpan also performs well in SRCC, with the best performance on 21 sub-datasets. Figure 2 and Figure S2 compares the distribution of the average predicted results for different allele loci. The prediction results at locus C are not as good as those of A and B, which indicates that the current tools are relatively conservative for the prediction of C loci. Kolmogorov Smirnov test [65] is used to test whether the mean prediction outputs of different loci are in the same distribution. As shown in Figure 2, the *P*-values show that the average prediction results of loci A and loci B belong to the same distribution with a confidence of more than 16%. Compared with loci A, the confidence of the same distribution between loci B and loci C is higher. This phenomenon may indicate that loci A and loci B are similar in structure, whereas loci B and loci C are also similar to each other, but there is a large structural difference between loci A and loci C. At present, the best prediction effect is still on loci A.

Besides, we divide and evaluate the datasets based on the length of peptides. It is found that the number of data items with a peptide length of 9 accounts for 53.0%. This indicates that HLA-I prefers to present peptides of length 9, followed by length 10. Figure 3 is the density maps of the average prediction results of eight tools on the 'f5' dataset. From Figure 3 and Figure S3, we can see that except for the subset with a peptide length of 8, the prediction accuracy decreases with the increase of peptide length. On the subset with peptide length of 8, the prediction accuracy is much lower than that of the peptide length of 9. This suggests that the binding properties of the 8-mer peptide are specific. This finding is consistent with the previous studies [66, 67], which implicates that the binding of the 8-mer peptide is related to the structural changes of some HLA alleles.

Finally, we divide the datasets into subgroups based on the restrictions of allele subtypes and peptide lengths. The experimental results are similar to that with the allelic subtype restriction, as shown in Figure S4. It should be noted that the binding peptides in all subsets with more than 100 samples are 9 or 10 amino acids in length, which also indicates the preference of HLA-I for presenting peptides with 9 and 10 amino acids.

Comparison of performance under different positive and negative sample ratios

In reality, the number of bound and unbound peptides is very heterogeneous during HLA-I peptide presentation. To evaluate the performance under different proportions of positive and negative samples, we populate the test set with different proportions of negative peptides. First, we randomly intercept one million peptides from UniProt [68] database for each length as a random peptide library. The random peptide library is then used to populate the test sets, and the resulting data sets have positive and negative sample ratios of 1:1, 1:5, 1:10 and 1:50, respectively. The results of these test sets are shown in Figure S5. From Figure S5, we can see that with the gradual increase of negative samples, the performance of these eight tools has changed, but the overall trend is similar. With the increase in the percentage of negative samples, MHCflurry and MATHLA perform better on AUC and SRCC, and MHCflurry and NetMHCpan perform better on AUPR, these tools are better suited to work in more demanding situations.

Capability of binding motifs

The binding motif of allele subtypes can reflect the preference for binding peptides at different positions. We use the 'ggseqlogo' R package [69] to draw real binding patterns for 68 subsets with

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbad150/7126340 by Virginia Commonwealth University Libraries user on 21 April 2023

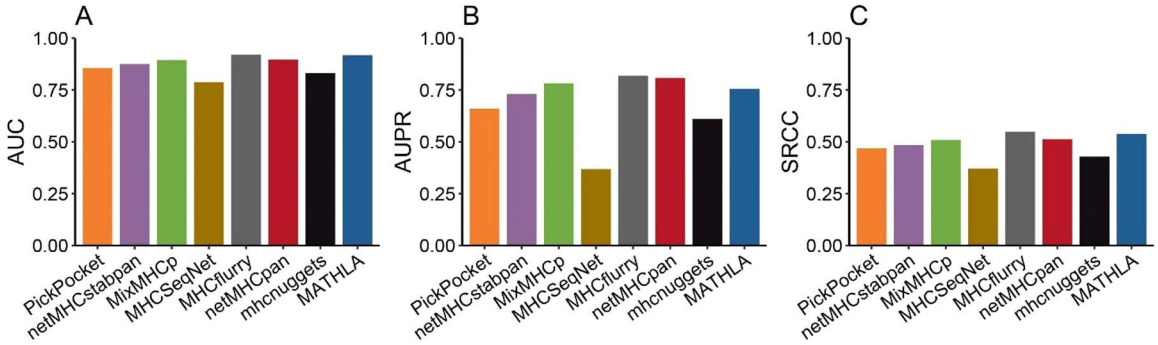

Figure 1. The overall performance of eight tools in different metrics on the 'f5' dataset.

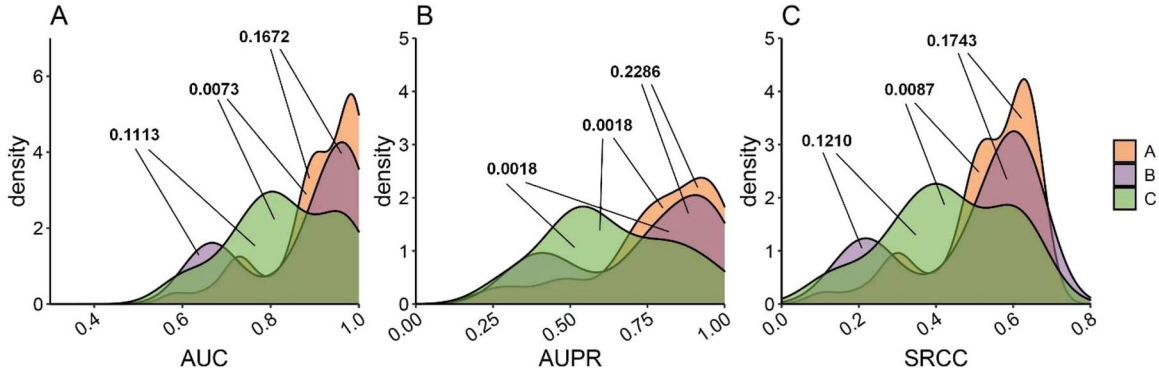

Figure 2. The average distribution of predicted results for alleles with different locus. The number between different distributions indicates the confidence that two distributions belong to the same distribution.

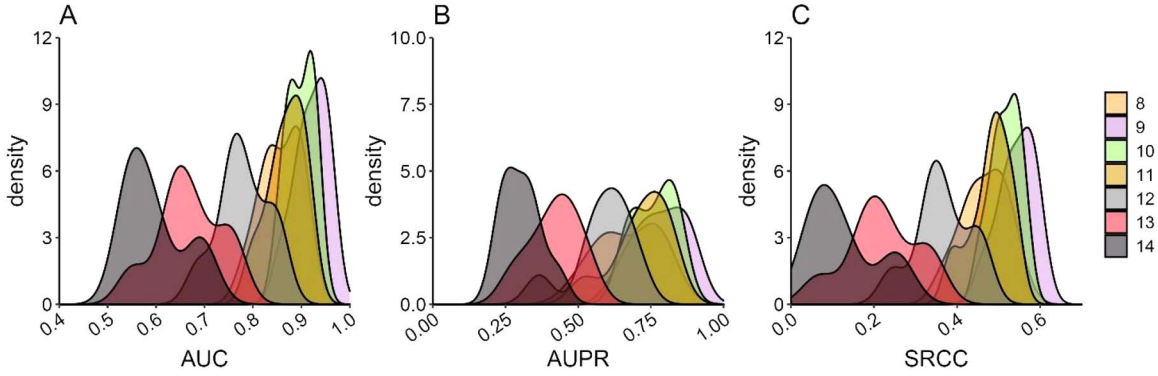

Figure 3. Distribution of predicted results for different peptide lengths.

more than 100 positive samples, as shown in Figure S6. For instance, HLA-A*01:01 prefers the 9-mer peptides which have Thr/Ser at the second position, Asp at the third position from the N terminal and Tyr at the first position from the C terminal, whereas HLA-A*01:01 with binding peptides of length 10 shows the similar preference. Statistics show that almost all of the alleles show a preference for N-terminal position 2 and C-terminal position 1. The alleles show similar positional binding preferences for peptides of different lengths. Thus, the ability to capture motifs is also considered to be an important measure of the quality of a tool. Firstly, we extract more than 1000 positive samples from 68 subgroups, including seven groups, and draw their true motifs. Secondly, each of these alleles is predicted to bind to a million random peptides. Then we pick out the top 1000 peptides in each tool according to the predicted scores for each allele and draw their binding motifs with these peptides. As shown in Figure 4, almost all the tools can capture allele preference, but they differ slightly in detail.

To find the relationship between the similarity of allelic sequences and their binding motifs, we re-cluster the 45 alleles in dataset one according to full sequence, pseudo-sequence 1 and pseudo-sequence 2 (described in 2.2), as shown in Figure S7. It can be seen that the full sequence-based allele clustering can better reflect the similarity of their binding motifs, whereas the other two sequence representation methods are less effective.

Sensitivity of data volume

To evaluate the effect of data volume on the performance of each tool, we divide the 'f5' dataset into different subsets based on different alleles and different peptide lengths. The scatter plots of the relationship between different data volumes and their measured values are shown in Figure S8. The predictions of these tools are very unstable at small data volumes, but they tend to stabilize as the data volume increases. We believe that better

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbad150/7126340 by Virginia Commonwealth University Libraries user on 21 April 2023

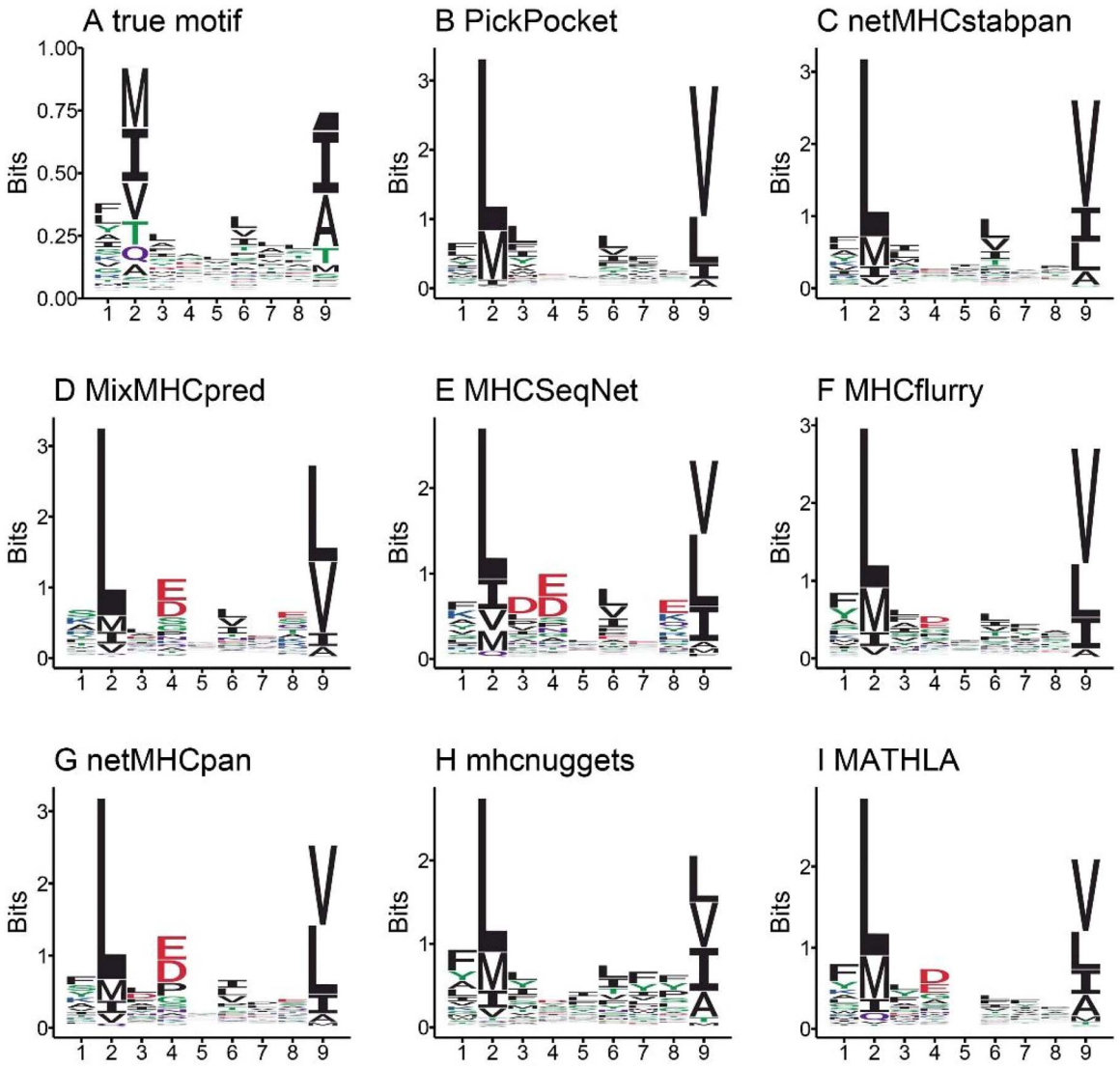

Figure 4. Comparison of true binding motifs and predicted motifs by eight tools on HLA-A*02:01 with peptides of length 9.

and more stable computational tools will be developed with the gradual increase of experimental data in the future.

Comparison of running speed

We compare the running speeds of these tools for different predicted data volumes in the same running environment, as shown in Figure 5. The running speed here only includes the process of the model inputting data from files and outputting the results. As the amount of input data increases, MHCFlurry's running time increases dramatically, reaching over 3000 s at 10^6 input data. When the input data amount increases to 10^6, NetMHpan and MHCSeqNet have similar running times of 2392.16 and 2311.94 s, respectively. The other five tools are faster and they run at comparable speeds. It is worth mentioning that although the MATHLA tool has a fast prediction speed, its input data needs to be converted to the npz format, which will take more waiting time that cannot be ignored.

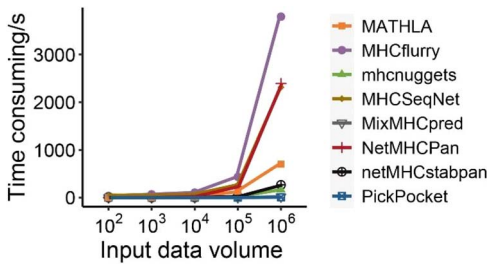

Figure 5. Comparison of running time consumption of each tool with the increase of input data. Each node is the average of five repeated experiments.

CONCLUSIONS

HLA-I molecules play an important role in human immunity. They present partial cleavage products of endogenous proteins to the T cells so that the immune system can recognize whether there is a mutation and facilitate further immune response. Accurate prediction of peptides that can be presented by HLA-I molecules provides useful information, particularly in the context of the recognition of neoepitopes. Recent years have seen an acceleration in the development of predictors, especially those that rely on modern machine learning algorithms like deep neural networks. The use of these cutting-edge algorithms aims to boost

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbad150/7126340 by Virginia Commonwealth University Libraries user on 21 April 2023

predictive performance. We review a comprehensive collection of 27 predictors, focusing on recent tools that apply modern algorithms. We compare them in terms of model inputs and outputs, algorithms and data encoding used and their availability. This analysis provides informative clues to the user to identify tools that are available via the most suitable means (code versus webserver), that rely on more sophisticated algorithms, and that satisfy their requirements concerning the inputs and outputs.

We also empirically compare a carefully selected set of eight predictors that covers popular older tools and several newest methods. This comparison benefits from our analysis of the reliability of data labels, which resulted in the curation of high-quality benchmark datasets. Moreover, we apply a variety of metrics and consider several different angles of assessment (alleles, negative to positive data ratios, runtime, etc.) to provide a comprehensive picture of various aspects of predictive quality. On the 'f5' dataset, the overall dataset, allele-specific datasets and peptide-length specific datasets, MHCflurry is always the most accurate tool for prediction. We test predictive performance under different positive-to-negative sample ratios. The corresponding results reveal that the performance of different tools changes in different ways with varying ratios. With the increase in the percentage of negative samples, MHCflurry and MATHLA perform better on AUC and SRCC, and MHCflurry and NetMHCpan perform better on AUPR. Furthermore, our analysis finds that each of the eight predictors captures meaningful binding motifs, which explains why they are capable of making accurate predictions. Finally, we show that some methods are faster than others, with the older PickPocket, netMHCstabpan, MHCnuggets and MixMHCpred being the fastest.

Altogether, our empirical analysis demonstrates that although MHCflurry can perform better on more alleles, there is no universally best predictor for all time. This systematic survey provides invaluable insights that allow us to identify the most suitable methods for a given prediction scenario. We also provide useful guidance for the development of future methods. As the predictive performance is sensitive to the amount of data, we anticipate that future methods that will benefit from larger training datasets will inevitably provide more accurate predictions.

Key Points

- We survey a comprehensive collection of 27 tools focusing on their input and output data characteristics, key aspects of the underlying predictive models and their availability.
- We provide a comprehensive comparison of convenience and accuracy of these methods under specific prediction scenarios, such as for analysis of specific alleles and length of peptides, influence of negative to positive data ratios, metrics of predictive performance and constraints on runtime.
- We provide practical observations and discuss directions for future developments in this research area.

SUPPLEMENTARY DATA

Supplementary data are available online at https://academic.oup.com/bib.

FUNDING

This work was supported in part by the National Natural Science Foundation of China (Grant No. 61832019) and the Hunan Provincial Science and Technology Program (Grants 2019CB1007 and 2021RC4008).

REFERENCES

1. Thakur A, Sharma A, Alajangi HK, *et al.* In pursuit of next-generation therapeutics: antimicrobial peptides against superbugs, their sources, mechanism of action, nanotechnology-based delivery, and clinical applications. *Int J Biol Macromol* 2022;**218**:135–56.
2. Yuvaraj N, Srihari K, Chandragandhi S, *et al.* Analysis of protein-ligand interactions of SARS-Cov-2 against selective drug using deep neural networks. *Big Data Min Anal* 2021;**4**(2):76–83.
3. Castle JC, Kreiter S, Diekmann J, *et al.* Exploiting the mutanome for tumor vaccination. *Cancer Res* 2012;**72**(5):1081–91.
4. Durgeau A, Virk Y, Corgnac S, Mami-Chouaib F. Recent advances in targeting CD8 T-cell immunity for more effective cancer immunotherapy. *Front Immunol* 2018;**9**:14.
5. Neefjes J, Jongsma MLM, Paul P, Bakke O. Towards a systems understanding of MHC class I and MHC class II antigen presentation. *Nat Rev Immunol* 2011;**11**(12):823–36.
6. Vita R, Overton JA, Greenbaum JA, *et al.* The Immune Epitope Database (IEDB) 3.0. *Nucleic Acids Res* 2015;**43**(D1):D405–12.
7. Lundegaard C, Lund O, Buus S, Nielsen M. Major histocompatibility complex class I binding predictions as a tool in epitope discovery. *Immunology* 2010;**130**(3):309–18.
8. Kosaloglu-Yalcin Z, Lanka M, Frentzen A, *et al.* Predicting T cell recognition of MHC class I restricted neoepitopes. *Onco Targets Ther* 2018;**7**(11):e1492508.
9. Vita R, Mahajan S, Overton JA, *et al.* The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res* 2019;**47**(D1):D339–43.
10. Bassani-Sternberg M, Bräunlein E, Klar R, *et al.* Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. *Nat Commun* 2016;**7**(1):13404. https://doi.org/10.1038/ncomms13404.
11. Ramarathinam SH, Croft NP, Illing PT, *et al.* Employing proteomics in the study of antigen presentation: an update. *Expert Rev Proteomics* 2018;**15**(8):637–45.
12. Nielsen M, Lund O, Buus S, Lundegaard C. MHC class II epitope predictive algorithms. *Immunology* 2010;**130**(3):319–28.
13. Shao XM, Bhattacharya R, Huang J, *et al.* High-throughput prediction of MHC class I and II neoantigens with MHCnuggets. *Cancer Immunol Res* 2020;**8**(3):396–408.
14. Han Y, Kim D. Deep convolutional neural networks for pan-specific peptide-MHC class I binding prediction. *BMC Bioinform* 2017;**18**(1):1–9.
15. Vang YS, Xie X. HLA class I binding prediction via convolutional neural networks. *Bioinformatics* 2017;**33**(17):2658–65.
16. 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**(1):1–10.
17. Bui HH, Schiewe AJ, von Grafenstein H, Haworth IS. Structural prediction of peptides binding to MHC class I molecules. *Proteins* 2006;**63**(1):43–52.
18. Mukherjee S, Bhattacharyya C, Chandra N. HLaffy: estimating peptide affinities for Class-1 HLA molecules by learning position-specific pair potentials. *Bioinformatics* 2016;**32**(15):2297–305.

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbad150/7126340 by Virginia Commonwealth University Libraries user on 21 April 2023

19. Antunes DA, Abella JR, Devaurs D, *et al.* Structure-based methods for binding mode and binding affinity prediction for peptide-MHC complexes. *Curr Top Med Chem* 2018;**18**(26):2239–55.
20. Liu G, Li D, Li Z, *et al.* PSSMHCpan: a novel PSSM-based software for predicting class I peptide-HLA binding affinity. *Giga Sci* 2017;**6**(5):gix017.
21. Bravi B, Tubiana J, Cocco S, *et al.* RBM-MHC: a semi-supervised machine-learning method for sample-specific prediction of antigen presentation by HLA-I alleles. *Cell Syst* 2021;**12**(2):195–202.e9.
22. 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**(4):1119–35.
23. Mahajan S, Yan Z, Jespersen MC, *et al.* Benchmark datasets of immune receptor-epitope structural complexes. *BMC Bioinform* 2019;**20**(1):1–7.
24. Andreatta M, Trolle T, Yan Z, *et al.* An automated benchmarking platform for MHC class II binding prediction methods. *Bioinformatics* 2018;**34**(9):1522–8.
25. Trolle T, Metushi IG, Greenbaum JA, *et al.* Automated benchmarking of peptide-MHC class I binding predictions. *Bioinformatics* 2015;**31**(13):2174–81.
26. Zhao W, Sher X. Systematically benchmarking peptide-MHC binding predictors: from synthetic to naturally processed epitopes. *PLoS Comput Biol* 2018;**14**(11):e1006457.
27. 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**(W1):W449–54.
28. O'Donnell TJ, Rubinsteyn A, Laserson U. MHCflurry 2.0: improved pan-allele prediction of MHC class I-presented peptides by incorporating antigen processing. *Cell Syst* 2020;**11**(1):42–48.e7.
29. Phloyphisut P, Pornputtapong N, Sriswasdi S, *et al.* MHCSeqNet: a deep neural network model for universal MHC binding prediction. *BMC Bioinform* 2019;**20**(1):1–10.
30. 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;**35**(23):4946–54.
31. Martini S, Nielsen M, Peters B, Sette A. The Immune Epitope Database and Analysis Resource Program 2003-2018: reflections and outlook. *Immunogenetics* 2020;**72**(1-2):57–76.
32. Campbell KM, Steiner G, Wells DK, *et al.* Prediction of SARS-CoV-2 epitopes across 9360 HLA class I alleles BioRxiv. 2020:2020.03 30.30.016931.
33. Alvarez B, Barra C, Nielsen M, Andreatta M. Computational tools for the identification and interpretation of sequence motifs in immunopeptidomes. *Proteomics* 2018;**18**(12):e1700252.
34. 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**(12):2459–77.
35. Reynisson B, Barra C, Kaabinejadian S, *et al.* Improved prediction of MHC II antigen presentation through integration and motif deconvolution of mass spectrometry MHC eluted ligand data. *J Proteome Res* 2020;**19**(6):2304–15.
36. Andreatta M, Nielsen M. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics* 2016;**32**(4):511–7.
37. Stranzl T, Larsen MV, Lundegaard C, Nielsen M. NetCTLpan: pan-specific MHC class I pathway epitope predictions. *Immunogenetics* 2010;**62**(6):357–68.
38. Robinson J, Mistry K, McWilliam H, *et al.* IPD—the immuno polymorphism database. *Nucleic Acids Res* 2010;**38**(suppl_1):D863–9.
39. Nielsen M, Lundegaard C, Blicher T, *et al.* NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and-B locus protein of known sequence. *PloS One* 2007;**2**(8):e796.
40. Sarkizova S, Klaeger S, le PM, *et al.* A large peptidome dataset improves HLA class I epitope prediction across most of the human population. *Nat Biotechnol* 2020;**38**(2):199–209.
41. Smith KJ, Reid SW, Stuart DI, *et al.* An altered position of the α2 helix of MHC class I is revealed by the crystal structure of HLA-B* 3501. *Immunity* 1996;**4**(3):203–13.
42. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004;**32**(5):1792–7.
43. Hu G, Kurgan L. Sequence similarity searching. *Curr Protoc Protein Sci* 2019;**95**(1):e71.
44. Collobert R, Weston J, Bottou L, *et al.* Natural language processing (almost) from scratch. *J Mach Learn Res* 2011;**12**(ARTICLE):2493–537.
45. Rammensee H-G, Bachmann J, Emmerich NPN, *et al.* SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics* 1999;**50**(3):213–9.
46. Reche PA, Glutting J-P, Reinherz EL. Prediction of MHC class I binding peptides using profile motifs. *Hum Immunol* 2002;**63**(9):701–9.
47. O'Donnell TJ, Rubinsteyn A, Bonsack M, *et al.* MHCflurry: open-source class I MHC binding affinity prediction. *Cell Syst* 2018;**7**(1):129–132.e4.
48. Ye Y, Wang J, Xu Y, *et al.* MATHLA: a robust framework for HLA-peptide binding prediction integrating bidirectional LSTM and multiple head attention mechanism. *BMC Bioinform* 2021;**22**(1):1–12.
49. 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 Bioinform* 2009;**10**(1):1–11.
50. Venkatesh G, Grover A, Srinivasaraghavan G, Rao S. MHCAttnNet: predicting MHC-peptide bindings for MHC alleles classes I and II using an attention-based deep neural model. *Bioinformatics* 2020;**36**(Supplement_1):i399–406.
51. Hoof I, Peters B, Sidney J, *et al.* NetMHCpan, a method for MHC class I binding prediction beyond humans. *Immunogenetics* 2009;**61**(1):1–13.
52. Nielsen M, Andreatta M. NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. *Genome Med* 2016;**8**(1):1–9.
53. 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**(9):3360–8.
54. Diao K, Chen J, Wu T, *et al.* Seq2Neo: a comprehensive pipeline for cancer neoantigen immunogenicity prediction. *Int J Mol Sci* 2022;**23**(19):11624.
55. Karosiene E, Lundegaard C, Lund O, Nielsen M. NetMHCcons: a consensus method for the major histocompatibility complex class I predictions. *Immunogenetics* 2012;**64**(3):177–86.
56. Lundegaard C, Lamberth K, Harndahl M, *et al.* NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11. *Nucleic Acids Res* 2008;**36**:W509–12.
57. Zhang H, Lund O, Nielsen M. The PickPocket method for predicting binding specificities for receptors based on receptor pocket

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbad150/7126340 by Virginia Commonwealth University Libraries user on 21 April 2023

similarities: application to MHC-peptide binding. *Bioinformatics* 2009;**25**(10):1293–9.
58. Mei S, Li F, Xiang D, *et al.* Anthem: a user customised tool for fast and accurate prediction of binding between peptides and HLA class I molecules. *Brief Bioinform* 2021;**22**(5):bbaa415. https://doi.org/10.1093/bib/bbaa415.
59. Yang X, Zhao L, Wei F, *et al.* DeepNetBim: deep learning model for predicting HLA-epitope interactions based on network analysis by harnessing binding and immunogenicity information. *BMC Bioinform* 2021;**22**(1):1–16.
60. Zhang Y, Zhu G, Li K, *et al.* HLAB: learning the BiLSTM features from the ProtBert-encoded proteins for the class I HLA-peptide binding prediction. *Brief Bioinform* 2022;**23**(5):bbac173.
61. Chu Y, Zhang Y, Wang Q, *et al.* A transformer-based model to predict peptide–HLA class I binding and optimize mutated peptides for vaccine design. *Nat Mach Intell* 2022;**4**(3):300–11.
62. 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**(4):1517–24.
63. Gfeller D, Guillaume P, Michaux J, *et al.* The length distribution and multiple specificity of naturally presented HLA-I ligands. *J Immunol* 2018;**201**(12):3705–16.
64. Chamoli T, Khera A, Sharma A, *et al.* Peptide utility (PU) search server: a new tool for peptide sequence search from multiple databases. *Heliyon* 2022;**8**(12):e12283.
65. Justel A, Peña D, Zamar R. A multivariate Kolmogorov-Smirnov test of goodness of fit. *Stat Probab Lett* 1997;**35**(3):251–9.
66. Rist MJ, Theodossis A, Croft NP, *et al.* HLA peptide length preferences control CD8+ T cell responses. *J Immunol* 2013;**191**(2):561–71.
67. Maenaka K, Maenaka T, Tomiyama H, *et al.* Nonstandard peptide binding revealed by crystal structures of HLA-B* 5101 complexed with HIV immunodominant epitopes. *J Immunol* 2000;**165**(6):3260–7.
68. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* 2019;**47**(D1):D506–D515.
69. Wagih O. ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics* 2017;**33**(22):3645–7.

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbad150/7126340 by Virginia Commonwealth University Libraries user on 21 April 2023