



A comprehensive review and performance evaluation of bioinformatics tools for HLA class I peptide-binding prediction

Shutao Mei, Fuyi Li , André Leier, Tatiana T. Marquez-Lago, Kailin Giam, Nathan P. Croft, Tatsuya Akutsu, A. Ian Smith, Jian Li, Jamie Rossjohn, Anthony W. Purcell and Jiangning Song 

Corresponding authors: Shutao Mei, Biomedicine Discovery Institute, Department of Biochemistry and Molecular Biology, Monash University, Melbourne, VIC 3800, Australia. E-mail: Shutao.Mei@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

Shutao Mei is currently 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, immunopeptidomics and immuno-informatics.

Fuyi Li received his BEng and MEng degrees in Software Engineering from Northwest A&F University, China. He is currently a PhD candidate in the Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Australia. His research interests are bioinformatics, computational biology, machine learning and data mining.

André Leier is an assistant professor and group leader in the Department of Genetics, School of Medicine, University of Alabama at Birmingham (UAB), USA. He is also an associate scientist in the UAB Comprehensive Cancer Centre. He received his PhD in Computer Science, University of Dortmund, Germany. He conducted postdoctoral research at Memorial University of Newfoundland, Canada, The University of Queensland, Australia and ETH Zürich, Switzerland. His research interests are in biomedical informatics and computational and systems biomedicine.

Tatiana T. Marquez-Lago is an associate professor and group leader in the Department of Genetics and Department of Cell, Developmental and Integrative Biology, School of Medicine, UAB, USA. Her research interests include multiscale modelling and simulations, artificial intelligence, bioengineering and systems biomedicine. Her interdisciplinary lab studies stochastic gene expression, chromatin organization, antibiotic resistance in bacteria and host-microbiota interactions in complex diseases.

Kailin Giam is a postdoctoral research associate at King's College London, United Kingdom. She received her PhD from The University of Melbourne, Australia. Her research focuses on understanding the immunological aspects of juvenile onset diabetes (type 1 diabetes) using cross-disciplinary strategies.

Nathan P. Croft received his PhD from the University of Birmingham and is currently a postdoctoral research fellow at Monash University, Australia. He focuses on identifying and quantifying the peptides displayed by major histocompatibility complex (MHC) molecules on the surface of infected cells for scrutiny by CD8⁺ T cells. He is also interested in the kinetic profiling of virus and host proteins that are modulated upon infection.

Tatsuya Akutsu received his DEng degree in Information Engineering in 1989 from 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.

A. Ian Smith completed his PhD at Prince Henry's Institute, Melbourne and Monash University, Australia. He is currently the vice-provost (Research & Research Infrastructure) at Monash University. He is also a professorial fellow in the Department of Biochemistry and Molecular Biology at Monash University, where he runs his research group. His research applies proteomic technologies to study the proteases involved in the generation and metabolism of peptide regulators involved in both brain and cardiovascular function.

Jian Li is a professor and group leader in the Monash Biomedicine Discovery Institute and Department of Microbiology, Monash University, Australia. He is a Web of Science 2015–2017 Highly Cited Researcher in Pharmacology & Toxicology. He is currently a National Health and Medical Research Council of Australia (NHMRC) principal research fellow. His research interests include the pharmacology of polymyxins and the discovery of novel, safer polymyxins.

Jamie Rossjohn 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 Australian Research Council (ARC) Australian Laureate fellow. His research is focused on understanding the structural and biophysical basis of MHC restriction, T-cell receptor (TCR) engagement and the structural correlates of T-cell signaling.

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 specializes 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.

Submitted: 11 February 2019; Received (in revised form): 2 April 2019

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 and also an associate investigator of the ARC Centre of Excellence in Advanced Molecular Imaging, Monash University. His research interests include artificial intelligence, bioinformatics, machine learning, big data analytics and pattern recognition.

Abstract

Human leukocyte antigen class I (HLA-I) molecules are encoded by major histocompatibility complex (MHC) class I loci in humans. The binding and interaction between HLA-I molecules and intracellular peptides derived from a variety of proteolytic mechanisms play a crucial role in subsequent T-cell recognition of target cells and the specificity of the immune response. In this context, tools that predict the likelihood for a peptide to bind to specific HLA class I allotypes are important for selecting the most promising antigenic targets for immunotherapy. In this article, we comprehensively review a variety of currently available tools for predicting the binding of peptides to a selection of HLA-I allomorphs. Specifically, we compare their calculation methods for the prediction score, employed algorithms, evaluation strategies and software functionalities. In addition, we have evaluated the prediction performance of the reviewed tools based on an independent validation data set, containing 21 101 experimentally verified ligands across 19 HLA-I allotypes. The benchmarking results show that MixMHCpred 2.0.1 achieves the best performance for predicting peptides binding to most of the HLA-I allomorphs studied, while NetMHCpan 4.0 and NetMHCcons 1.1 outperform the other machine learning-based and consensus-based tools, respectively. Importantly, it should be noted that a peptide predicted with a higher binding score for a specific HLA allotype does not necessarily imply it will be immunogenic. That said, peptide-binding predictors are still very useful in that they can help to significantly reduce the large number of epitope candidates that need to be experimentally verified. Several other factors, including susceptibility to proteasome cleavage, peptide transport into the endoplasmic reticulum and T-cell receptor repertoire, also contribute to the immunogenicity of peptide antigens, and some of them can be considered by some predictors. Therefore, integrating features derived from these additional factors together with HLA-binding properties by using machine-learning algorithms may increase the prediction accuracy of immunogenic peptides. As such, we anticipate that this review and benchmarking survey will assist researchers in selecting appropriate prediction tools that best suit their purposes and provide useful guidelines for the development of improved antigen predictors in the future.

Key words: HLA; peptide binding; bioinformatics; machine learning; sequence analysis; web server; prediction model; performance benchmarking

Introduction

The binding of peptides to specific human leukocyte antigen (HLA) allomorphs and the subsequent recognition of peptide-HLA complexes (pHLAs) by T cells establish the antigenicity of the peptide, and this process forms the basis of immune surveillance [1]. HLA molecules can be divided into two main categories, namely HLA class I (HLA-I) and HLA class II (HLA-II). The HLA-I molecules are encoded by three I loci (HLA-A, -B and -C), and the encoded proteins are expressed on the surface of all nucleated cells. In contrast, HLA-II molecules encoded by HLA class II loci (HLA-DR, -DQ and -DP) can only be expressed in professional antigen-presenting cells (APC) such as dendritic cells, mononuclear phagocytes and B cells [1]. HLA-I molecules mainly bind short peptides of 8–12 amino acids in length, typically derived from proteasome-mediated degradation of intracellular proteins. These pHLAs are then presented on the cell surface for recognition by CD8⁺ T cells. HLA-II molecules tend to bind longer peptides (12–20 amino acids in length) liberated from extracellular proteins within the endosomal compartments. These pHLAs are presented on the surface of professional APC for recognition by CD4⁺ T cells [2]. The interactions between HLA molecules and peptides and subsequent recognition of these complexes by T cells control the magnitude and effectiveness of the immune response. Thus, a major goal in vaccinology and immunotherapy resides in the accurate prediction of peptide-HLA binding and the ability of these complexes to induce a desired immune response [3]. Understanding which

peptides are selected for display in the context of an individual's HLA type can aid the design of vaccines designed to induce protective or therapeutic immunity towards various pathogens [4, 5]. Equally, several studies have found that neo-antigens generated as a result of non-synonymous mutations in cancer cells play a significant role in the dynamics of the anti-tumour immune response [6–9]. Indeed, vaccines based on such neo-antigens have been shown to benefit clinical outcomes [10, 11]. Typically, identifying neo-antigens first requires characterization of the non-synonymous mutations from primary tumours using next-generation sequencing (NGS) platforms such as RNAseq. In the second step, peptide sequences that contain mutations are further assessed by predicting their individual probability to bind to patient HLA allomorphs [12]. By filtering out potential allomorph-specific HLA ligands, the number of candidates can be substantially decreased, thereby accelerating the final step of experimentally verifying neo-epitopes [13, 14]. These and other considerations have increased the interest in predicting peptide binding to HLA molecules over the past few years.

Over the last few decades, the availability of experimentally verified HLA ligand sequences has increased, with sequences often deposited in public peptide ligand databases. Up until recently, most of these data have been generated using *in vitro* binding assays [15, 16], but the use of mass spectrometry (MS)-based identification of purified HLA-binding peptides has now come into the forefront [17–19]. The Immune Epitope Database (IEDB) is the largest public resource for HLA ligands

and T-cell epitopes [20]. It contains detailed information on curated peptides collected from published journal articles with appended metadata including the experimental modality used for data acquisition. Due to the increasing availability of high-quality HLA allele-specific data sets, a number of new computational tools have been developed for predicting peptide binding to HLA molecules. However, the main focus has been on the prediction of HLA-I ligands, since the more complex nature (longer and more heterogeneous peptide sequences) of HLA-II ligands makes their prediction more difficult [21]. Here we classify these publically available tools into three major categories based on the methodologies they used, namely (i) methods based on sequence-scoring functions, including SYFPEITHI [22], RANKPEP [23], PickPocket 1.1 [24], stabilized matrix method (SMM)—peptide: major histocompatibility complex (MHC)-binding energy covariance (SMMPMBEC) [25], PSSMHCpan 1.0 [26] and MixMHCpred 2.0.1 [18, 27]; (ii) methods based on machine learning algorithms, including NetMHC 4.0 [28], NetMHCstabpan 1.0 [29], NetMHCpan 4.0 [30], MHCflurry 1.2.0 [31], MHCnuggets 2.0 [32], ConvMHC [33] and HLA-CNN [34]; and (iii) methods based on the integration of different peptide-binding predictors, including NetMHCcons 1.1 [35] and IEDB-analysis resource-consensus (IEDB-AR-Consensus) [25]. It should be noted that structure-based methods can also contribute to HLA-I peptide-binding prediction. These methods have achieved accurate binding prediction performance by modelling the docking between the HLA protein and peptide ligands [36–51]. However, structure-based approaches are not suitable for all allotypes and rely on homologous structures. Therefore, the present review has not included the structure-based peptide-binding prediction methods.

Several attempts have been made to provide benchmark tests of prediction tools; however, each study had certain limitations: either they did not include a performance evaluation of all reviewed tools, or several state-of-the-art prediction tools were not considered and benchmarked or did not have a detailed algorithm description for each of the reviewed tools [52–56]. For example, a recent review comprehensively discussed currently available peptide-binding prediction tools, but it lacked a uniform validation approach to allow performance comparison of the different predictors [2]. To overcome these issues, here we provide a comprehensive performance benchmarking and assessment of currently available, state-of-the-art tools for predicting peptide binding to HLA-I molecules. In total, 15 prediction tools have been systematically benchmarked in terms of their underlying algorithms, feature selection, performance evaluation strategy and webserver and/or software functionality. Most importantly, we also performed an independent test to evaluate the performance of these tools by using a newly generated peptide data set containing HLA molecule ligands across 19 HLA-I allomorphs. Following our review, we give some suggestions for the design and development of future prediction tools. Lastly, we hope our review will assist and inspire scientists with interest in this field to facilitate their efforts in developing improved tools for the prediction of T-cell epitopes.

Materials and methods

Construction of the positive validation data set

In order to evaluate the performance of currently available peptide-binding prediction tools and provide an overall comparison between them, we extracted the annotations of peptides

including peptide sequences, source proteins where peptides were derived from, binding experimental results and the type of HLA molecules that the peptides bound to, from the latest versions of several widely used public databases including IEDB [57], SYFPEITHI [22], MHCBN [58] and EPIMHC [59]. Of note, SYFPEITHI, MHCBN and EPIMHC only store binary data (i.e. positive or negative) to distinguish whether a peptide has been experimentally verified to be a binder or not, while 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 data set, we collected all positive peptides from all the above four public databases, regardless of any quantitative information provided in IEDB. Next, we removed the sequence redundancy by adopting the following procedures: (i) only selecting confirmed allotype-specific peptides, (ii) removing duplicate peptides if they were associated with the same HLA allotype according to the databases, (iii) removing any peptides contained within the training data sets of the reviewed prediction tools and (iv) removing the peptides that had unnatural amino acids [26]. In addition, since most peptides presented by the HLA-I complexes are of 9–11 amino acids long [60], we only retained those peptides with such lengths in the constructed validation data set. Overall, we obtained an independent validation data set with a total of 21 101 non-redundant peptide ligands across 19 HLA allotypes. A statistical summary of the final constructed validation data set is provided in Table S1.

Construction of the negative validation data set

In order to generate a balanced data set with an equal number of non-binding peptides to those positive peptides associated with a certain HLA allotype, we used non-binding regions of the source proteins of the peptides in the positive data set. This strategy has been commonly used for previous performance benchmarking studies [28, 30, 61, 62]. Specifically, we first randomly selected the source proteins from the positive peptide data set. Then we generated a set of peptide sequences by splitting the sequences of the source proteins into 9, 10 or 11 amino acids-long peptides. Those peptides were further pooled according to their lengths and randomly selected to form the negative data set after filtering for those already contained in our independent positive data set or in the training data sets of the reviewed prediction tools. The random selection of negative peptides was performed in a way such that the numbers of negative peptides were identical to the numbers of positive peptides for each length (9, 10 and 11 mers) for each HLA-I allotype. It should be noted that this strategy might falsely classify peptides as non-binders since their binding potential was not formally assessed. However, as binding to these HLA allotypes are very specific and there would only exist a few allotype-specific HLA binders originating from one source protein, the proportion of misclassified peptides is likely very small [62, 63].

Existing peptide-binding prediction tools

Table 1 provides a summary of currently available tools for HLA-peptide-binding prediction, which are grouped into three major categories in terms of their availability, employed peptide features (for machine learning-based predictors), algorithms and performance evaluation strategies. Figure 1 provides an overall workflow for the three tool categories as an illustration of their underlying general methodologies for constructing peptide-

Table 1. A list of currently available tools for HLA-I peptide-binding prediction assessed in this review

Category	Tool ^a	Year	Software availability ^b	Webserver availability ^b	Max data upload	Step-by-step instruction ^{b,c}	Parameters setting function ^{b,c}	Email of result ^{b,c}	Features	Algorithm ^c	Training-test split ^c	Performance evaluation strategy ^c	Last updated date
Scoring function based	SYFPEITHI	1999	No	Yes	Single sequence	No	No	No		PSSM	NM	N.A.	August 2012
	RANKPEP	2002	No	Yes	N.M.	Yes	Yes	No		PSSM	NM	Independent test	January 2019
	PickPocket 1.1	2009	Yes	Yes	≤5000 sequences and ≤20 000 amino acids	Yes	Yes	Yes		PSSM	NM	LOO, 5-fold CV	January 2017
	SMMPMBEC	2009	Yes	Yes	≤200 sequences or ≤10 MB	Yes	No	Yes		PSSM	NM	5-fold CV	January 2014
	PSSMHCpan 1.0	2017	Yes	No	N.A.	N.A.	N.A.	N.A.		PSSM	NM	10-fold CV, independent test	February 2017
Machine learning based	MixMHCpred 2.0.1	2017	Yes	No	N.A.	N.A.	N.A.	N.A.		PSSM	NM	Independent test	August 2018
	NetMHC 4.0	2014	Yes	Yes	≤5000 sequences and ≤20 000 amino acids	Yes	Yes	Yes	Sequence-based features	NN	4:1	5-fold CV	October 2017
	NetMHCstabpan 1.0	2016	Yes	Yes	≤5000 sequences and ≤20 000 amino acids	Yes	Yes	Yes	Physicochemical features	NN	4:1	5-fold CV, independent test	September 2018
	NetMHCpan 4.0	2017	Yes	Yes	≤5000 sequences and ≤20 000 amino acids	Yes	Yes	Yes	Sequence-based & binary features	NN	4:1	5-fold CV, independent test	January 2018
	MHCnuggets 2.0	2017	Yes	No	N.A.	N.A.	N.A.	N.A.	Sequence-based & physicochemical features	NN	NM	3-fold CV	August 2018
	ConvMHC	2017	No	Yes	N.M.	No	No	No	Sequence-based & physicochemical features	NN	75:1	LOO, 5-fold CV, independent test	July 2018
	HLA-CNN	2017	Yes	No	N.A.	N.A.	N.A.	N.A.	Sequence-based features	NN	7:3	5-fold CV, independent test	August 2017

(Continued.)

Table 1. Continued.

Category	Tool ^a	Year	Software availability ^b	Webserver availability ^b	Max data upload	Step-by-step instruction ^{b,c}	Parameters setting function ^{b,c}	Email of result ^{b,c}	Features	Algorithm ^c	Training-test split ^c	Performance evaluation strategy ^c	Last updated date
	MHCflurry 1.2.0	2018	Yes	No	N.A.	N.A.	N.A.	N.A.	Sequence-based & binary features	NN	9:1	Independent test	January 2019
Consensus	IEDB-AR-Consensus	2012	Yes	Yes	≤200 sequences or ≤10 MB	Yes	No	Yes		Con	NM	Independent data set	May 2012
	NetMHCcons 1.1	2012	Yes	Yes	≤5000 sequences and ≤20 000 amino acids	Yes	Yes	Yes		Con	NM	Independent data set	January 2017

^aThe URL addresses for accessing the listed 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/>; SMMMPMBEC, <https://github.com/ykimbiology/smmmpmbec/>; PSSMHCPan 1.0, <https://github.com/BGI2016/PSSMHCPan/>; MixMHCpred 2.0.1, <https://github.com/GfellerLab/MixMHCpred/>; NetMHC4.0, <http://www.cbs.dtu.dk/services/NetMHC/>; NetMHCstabpan 1.0, <http://www.cbs.dtu.dk/services/NetMHCstabpan/>; NetMHCpan-4.0, <http://www.cbs.dtu.dk/services/NetMHCpan/>; MHCnuggets 2.0, <https://github.com/KarchinLab/mhcnuggets-2.0/>; ConvMHC, <http://jumong.kaist.ac.kr:8080/convmhc/>; HLA-CNN, <https://github.com/uci-cbcl/HLA-bind/>; MHCflurry 1.2.0, <https://github.com/openvax/mhcfllurry/>; IEDB-AR-Consensus, <http://tools.iedb.org/mhci/>; NetMHCcons-1.1, <http://www.cbs.dtu.dk/services/NetMHCcons/>. ^bYes: the publication has developed the function according to the column; No: the publication has not developed the function according to the column. ^cAbbreviations: N.M., not mentioned; N.A., not applicable; PSSM, position-specific scoring matrix; NN, Neural network; Con, Consensus-based method; CV, cross-validation; LOO, leave-one-out.

binding predictors. Most tools constructed the prediction models in a five-step manner [64], which involves data collection and pre-processing as the 1st step, feature encoding and selection as the 2nd step, followed by model construction and optimization, performance evaluation and webserver/stand-alone software construction [65–69].

Computational methods developed based on statistical scoring functions score the candidate peptide sequences by calculating certain features such as the sequence similarity and amino acid frequencies. Other statistical scores depend on the position-specific amino acid profiles of peptides. For instance, the position-specific scoring matrix (PSSM) and BLOSUM 62 matrix [70] are two widely used scoring matrices for specifying the evolutionary information of amino acids at different positions of a peptide sequence [71]. After the PSSM matrices are generated, the score of a peptide can be calculated by multiplying the frequencies of the corresponding amino acids at each position.

Machine learning-based methods classify a peptide as a binder or non-binder by generating a score using the training model based on the extracted representative features. Construction of a machine learning-based model for predicting a peptide's binding probability generally involves four major steps: (i) construction of training data sets where the binding between the HLA allele and peptide ligands have been experimentally verified, (ii) feature encoding based on the peptide and/or HLA allele sequences, (iii) selection of a best-performing machine-learning algorithm and training of the corresponding machine-learning model and (iv) optimization of the model and its performance evaluation. As shown in Table 1, the neural network (NN) is the dominant machine-learning algorithm used by currently available peptide-binding prediction tools. Therefore, we use NN as an example to illustrate the workflow of how to construct machine learning-based tools in Figure 1.

As the 3rd major category of methods, consensus-based methods integrate several peptide-binding predictors in a weighted manner and generate the final prediction score based on the results of all individual predictors. The rationale of these methods is that combining the results of several individual predictors might help improve the performance of the final prediction compared to that of individual predictors [72]. Such consensus-based tools can be implemented based on a combination of similar binding prediction models, as is the case for NetMHCcons 1.1 [35], which will be discussed later.

Scoring function-based tools

The major difference between different scoring function-based tools is the statistical approach that they used to calculate the binding score for a candidate peptide sequence. From this perspective, SYFPEITHI [22] calculates the prediction score of a peptide sequence by adding the corresponding value of each amino acid at each position. For a given HLA allele, amino acids that frequently occur at anchor positions are given the value of 10. The less frequent amino acids are assigned with lower values. The final score of a sequence is the sum of values at each position. As for the result, for a given HLA-I molecule and peptide length, SYFPEITHI calculates the 10 highest-scoring peptides among all possible amino acids with the same length of a given sequence.

RANKPEP [23] predicts the MHC class I-binding peptides using profile motifs by calculating the PSSM of ligands bound to a given HLA allele. Briefly, the ligands of each HLA allele

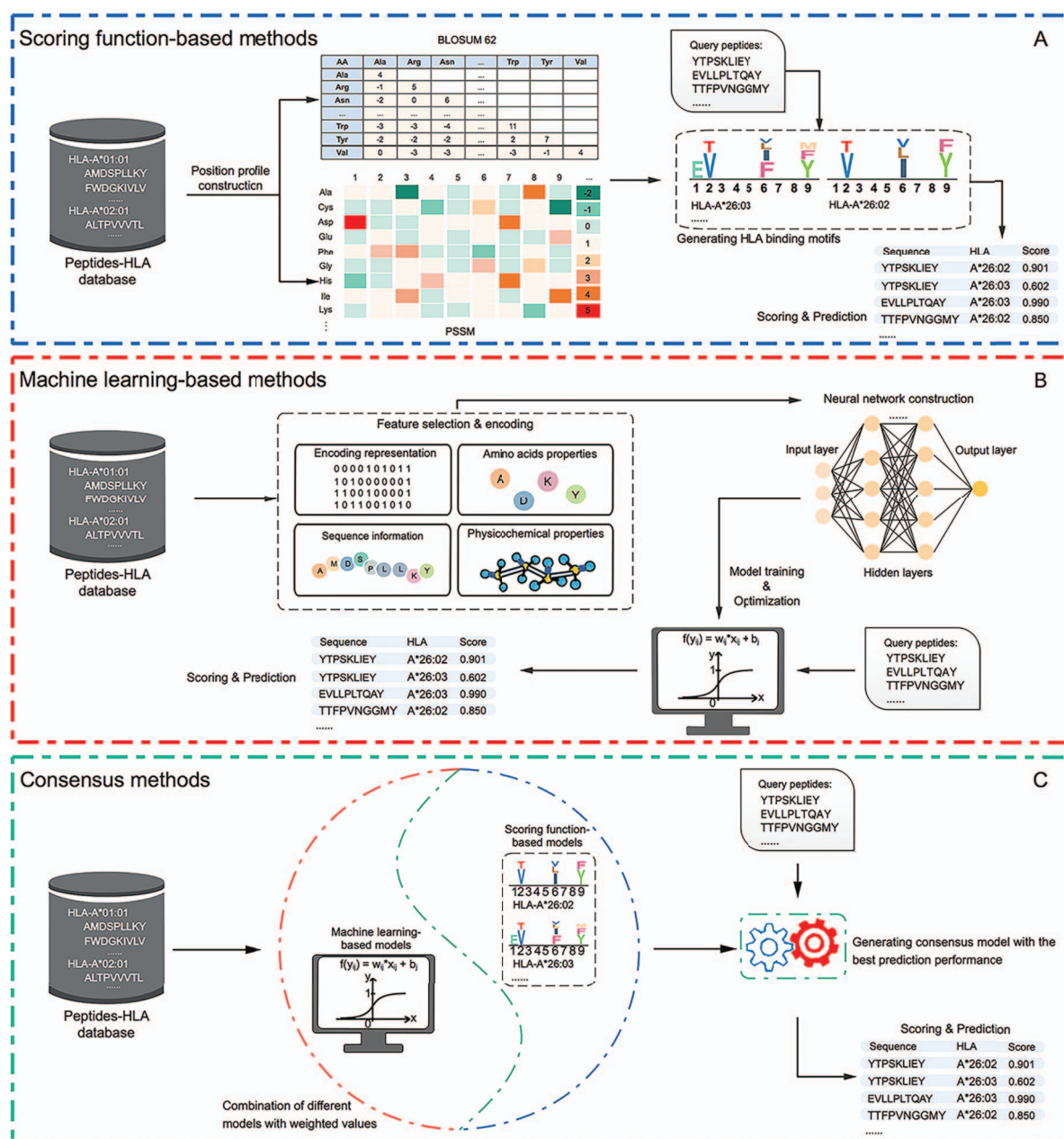


Figure 1. Graphical illustrations of (A) scoring function-based methods; (B) machine learning-based methods and (C) consensus-based methods. For each type of methods, the key steps are summarized and visualized. Scoring function-based methods predict peptide binding using a scoring function to generate the motifs of specific HLA alleles. Machine learning-based methods perform the prediction using well-trained models based on the training data sets. Consensus-based methods can predict peptide binding by integrating different peptide-binding prediction models.

are parsed by the length in five sets of 8, 9, 10, 11 and 12+mers. The PSSM of each length set, as generated by using the PROFILEWEIGHT protocol [73], defines the sequence-weighted frequency of each amino acid observed at each position of the peptide. These values are then normalized to the corresponding expected background frequency of that amino acid in the proteome. The prediction score is calculated by aligning the PSSM with the peptides and adding up the scores that match the residue type and position. The scoring starts at the beginning of each sequence, and the PSSM is slid over the sequence one residue at a time until the end of the sequence. A binding threshold is set to a value. Thus, peptides with a score equal or above the binding threshold are predicted as binders.

PickPocket 1.1 [24] is another scoring function-based tool for predicting peptide binding. It uses the SMM algorithm to construct the PSSM for each HLA allotype according to the peptide ligands in the training data set. To extend the utility of this algorithm for HLA allotypes with limited or no known ligand data, peptide binding was deconvolved to pocket-specific binding events. These pockets, distributed along the antigen-binding cleft of the HLA-I molecule, form specificity determining interactions with amino acid side chains distributed along the length of the ligand. Thus pocket residues library was constructed from HLA allotypes for which a significant amount of peptide-binding information was available to create a PSSM for less well-studied HLA allomorphs. Then, the similarity between the HLA allotype

with insufficient ligand data H_q and each HLA allotype with a known PSSM in the training database H_i was calculated by using the pocket residues of the two HLA molecules as follows:

$$\text{Sim}(s_q, s_i) = \frac{S(s_q, s_i)}{\sqrt{S(s_q, s_q) \cdot S(s_i, s_i)}}, \quad (1)$$

where $S(s_q, s_i)$ denotes the BLOSUM62 [74] similarity score between the two pocket residue sequences s_q and s_i . As a result, the PSSM was calculated as a weighted average based on pocket similarities of all known HLA PSSMs in the training data set.

The SMM method employed in PickPocket has been used for generating other models of the peptide binding to uncharacterized HLA allotypes [75]. SMMPBEC [25, 76] is a tool that uses a PMBEC matrix as a prior to improve the prediction performance when using the SMM method. Instead of using BLOSUM62, the tool employed the PMBEC to calculate the similarity between any two amino acids. For this purpose, a library of combinatorial 9-mer peptide mixtures was used to measure the binding affinity contribution of each residue in a 9-mer peptide to 24 MHC allotypes [76]. All peptides in the library share the same residue type at one position, while the remaining positions are allowed to sample all residue types randomly. In total, 180 mixtures of covering 20 residue types in all positions of a 9-mer peptide were synthesized to test the binding affinity to 24 MHC molecules. The measured binding affinity is defined as IC50 values, which reflect the concentration of peptide yielding 50% inhibition of the binding of a radiolabeled tracer peptide [76]. In order for the final predictions to be represented with a normal distribution and to fit the experimental data well [75], the IC50 value of an amino acid aa at the residue position pos for a given MHC allotype was log transformed to approximate the relative binding energy contribution as follows:

$$\Delta E_{aa, pos, MHC} = \log(\text{IC50}_{aa, pos, MHC}) - \frac{1}{20} \sum_{aa'} \log(\text{IC50}_{aa', pos, MHC}), \quad (2)$$

where $\frac{1}{20} \sum_{aa'} \log(\text{IC50}_{aa', pos, MHC})$ represents the average log transformed IC50 value of all other residues at the same position.

Next, in order to generate the PMBEC matrix, the similarity of two amino acids aa and aa' was defined as the covariance of their relative binding energy contributions, denoted as $\Delta E_{aa, pos, MHC}$ and $\Delta E_{aa', pos, MHC}$, respectively, and can be calculated as follows:

$$\begin{aligned} \text{cov}(aa, aa') &= \frac{1}{24 * 9} \sum_{MHC=1}^{24} \sum_{pos=1}^9 (\Delta E_{aa, pos, MHC} - \Delta E_{aa}) (\Delta E_{aa', pos, MHC} - \Delta E_{aa'}), \end{aligned} \quad (3)$$

where ΔE_{aa} and $\Delta E_{aa'}$ are the averages over all positions and MHC molecules for amino acids aa and aa' , respectively. The prediction outcome is the IC50 value of the peptide transformed from the sum of contributions of residues at each position based on the result of the SMM method that used the PMBEC matrix as a prior.

PSSMHCpan 1.0 [26] is a recently published scoring function-based tool that also uses the PSSM features to predict peptide binding to HLA-I molecules. It can predict both characterized HLA-I allotypes (with binders in the training data set) and uncharacterized HLA-I allotypes (with no binders in the training

data set). The PSSM of each characterized HLA-I allotype is defined as a matrix of M rows ($M=20$, number of amino acid types) and N columns ($N=8-25$, peptide sequence length). Each element $p_{a,i}$ in the matrix is calculated as $p_{a,i} = \log \frac{F_{a,i} + \omega}{BG_a}$, where $F_{a,i}$ is the frequency of amino acid a at position i in a peptide from the training data set; BG_a is the background frequency of amino acid a in the proteome; and ω is a random value ranging from 0 to 1 [77]. Then, the tool defined the binding score of a given peptide by summing the corresponding values of $p_{a,i}$ of the amino acid at each position in the PSSM of the corresponding characterized HLA allotype as follows: binding score = $\frac{\sum_{i=1}^N p_{a,i}}{N}$, where N is the length of the peptide. Finally, PSSMHCpan 1.0 converts the binding score of each peptide to an IC50 value and uses this value as the prediction result for each peptide.

Specifically, for a characterized HLA molecule, the binding score is defined as binding_score = $(\sum_{i=1}^N p_{a,i})/N$. Then the IC50 value is calculated as follows:

$$\text{IC50} = 50000^{\text{Max} - \text{binding_score} / \text{Max} - \text{Min}}, \quad (4)$$

where Max and Min denote the maximum and minimum values of the binding_score, respectively.

For uncharacterized HLA-I allotypes, PSSMHCpan 1.0 first generated a library of HLAs that contained pairs of characterized and uncharacterized HLA proteins with a weight value associated with each pair. Finally, a given peptide of an uncharacterized HLA allotype could be qualitatively predicted with its IC50 value as follows:

$$\text{IC50}_{un} = \frac{\sum_{i=1}^S (w_i \text{IC50}_i)}{\sum_{i=1}^S w_i}, \quad (5)$$

where S is the sum of characterized HLA allotypes that pair the given uncharacterized HLA allele from the HLA similar weight library and w_i and IC50_i denote the weight value and the prediction result of the given peptide corresponding to the characterized HLA allele i , respectively.

MixMHCpred 2.0.1 [18, 27] is another recently published tool. It calculates the PSSM of each allotype based on a large training data set containing more than 115 000 non-redundant peptides across 123 HLA-I molecules derived from MS. This represents the largest training data set among all tools reviewed here. The PSSM of each allotype is calculated by first pooling all peptides assigned to the same allotype. Then single position weight matrices for each allotype and each peptide length type were built for the length $L = 8$ to $L = 14$, including pseudo-counts based on BLOSUM correction and renormalization by the expected background amino acid frequencies [18, 78]. The score for predicting a peptide was calculated by summing the logarithm of the corresponding PSSM at each position of the given peptide defined as follows:

$$S^h(X) = \log \left(\prod_{l=1}^L M_{x,l}^{(h,l)} \right) / \log(L), \quad (6)$$

where h and L is the HLA allotype and the peptide length, respectively. $M_{x,l}^{(h,l)}$ is the value PSSM at the l position of the peptide X based on the PSSM of the h allotype and L length. In the latest version of MixMHCpred, the peptide length distribution and multiple specificity are incorporated into the predictor and lead to the improvement of prediction performance as described in [27].

Machine learning-based tools

The limitation of scoring function-based approaches is that their methods for calculating the prediction scores are relatively simplistic, since they only handle linear features such as sequence similarity and pattern. In the last decade, machine learning-based algorithms have been increasingly used for constructing models to predict peptide binding to HLA allotypes. Such algorithms are capable of identifying non-linear patterns underlying the peptide-binding data [79], which cannot be easily captured by scoring function-based methods. Among various machine learning-based tools, the most commonly used algorithm is NN.

As shown in Table 1, most machine learning-based tools reviewed here utilized NN to construct the prediction models. Generally, the NNs in HLA-peptide-binding prediction models have a layered feed-forward architecture. Briefly, for a typical multi-layer feed-forward NN, the layers are composed of the input, hidden and output layers. Each layer can contain neurons or units to represent the signal. Different units of the layer can be connected to other units of the neighbouring layer through weights and biases. The signal of a unit x_i can be transformed and used as the input of its connecting unit y_j through the function $y_j = w_{ij}x_i + \theta_j$, where w_{ij} is the weight value with respect to the units x_i and y_j , and θ_j is the bias of unit y_j . Then, the input outcome of unit y_j is transformed by an activation function like the sigmoid function $f(y_j) = \frac{1}{1+e^{-y_j}}$. Then, $f(y_j)$ is the output of unit y_j and can be used by the next layer. For optimization of the parameters, back propagation (BP) is a widely used algorithm to optimize the w_{ij} and θ_j . The theory of the BP algorithm is based on the error-correction-learning rule. Briefly, the NN parameters are optimized by modifying the weights and biases via using error function according to the difference between the network output and the actual label [80].

In this section, we will discuss in detail the current strategies of constructing the architecture of NN for peptide-binding prediction.

NetMHC 4.0 [28] uses an ensemble method to generate the NN and assigns the binding core of a given peptide based on the majority vote of the networks in the ensemble. Briefly, the top 10 networks with the highest test set Pearson correlation coefficient within the 50 networks for each training/test set configuration were selected as the final network ensemble. It uses a BP algorithm to update the weights between units [81]. In particular, it uses both BLOSUM62 and sparse encoding schemes to encode the peptide sequences into nine amino acid-binding cores. For peptides longer or shorter than nine amino acids, deletion or insertion methods are applied to reconcile or extend the original sequence to a core of nine amino acids [28]. In addition, other complementary sequence-based features such as the number of deletions/insertions, the length and compositions of the terminal regions flanking the predicted binding core were also incorporated to enable the algorithm to learn the complex binding patterns from the peptide-HLA-I molecule pairs.

NetMHCstabpan 1.0 [29] is a prediction tool that predicts the stability of peptide-HLA-I complexes based on an NN-based algorithm. The stability of the peptide-HLA-I complex is defined as the half-life of the pHLA complex, which is determined by a scintillation proximity-based peptide-HLA-I dissociation assay [82]. Then, the half-life values are transformed to a value ranging between 0 and 1 as follows: $s = 2^{-t_0/th}$, where s is the transformed value, t_0 is the measured pHLA complex half-life and t_0 is a threshold value that is fitted to obtain a suitable distribution of

the data for training purposes, which was set as 1 after optimization of the prediction performance. The tool uses BLOSUM50 or the sparse encoding scheme to encode peptide sequences.

NetMHCpan 4.0 [30] is also an NN-based tool similar to NetMHC 4.0. The major differences between the two tools include two aspects: first, NetMHCpan 4.0 was trained using peptides generated from both binding affinity assays and naturally presented peptide ligands identified by MS; second, the amino acids from the HLA heavy chain that contact the peptide ligand are extracted. These sequences are called pseudo-sequences and enable the tool to predict binding to HLA allotypes with little available binding data. The assumption is that similar HLA allotypes will bind similar peptides. In NetMHCpan 4.0, the similarity between two HLA molecules is defined as the pseudo-distance $d = 1 - \frac{s(A,B)}{\sqrt{s(A,A) \times s(B,B)}}$, where $s(A,B)$ is the BLOSUM50 similarity score between pseudo-sequences A and B. The algorithm takes the peptides and the chosen HLA allotype in terms of a pseudo-sequence as inputs. All peptides are represented as 9-mer binding cores by using the same method described in NetMHC 4.0 [28].

MHCnuggets 2.0 [32] is a prediction tool developed based on deep learning (DL) methods. It contains two DL models: (i) long short-term memory networks [83] and (ii) gated recurrent units [84]. The architecture of both networks is a fully connected single layer of 64 hidden units. The network is regularized with a dropout and recurrent dropout probability is 0.2, which means during each time of optimization, the NN algorithm will randomly choose to ignore 20% of units to avoid overfitting. The input sequence is encoded as a 21-dimensional vector using a sparse encoding scheme. Both models were trained using the Adam optimizer [85] with a learning rate of 0.001. Instead of using pseudo-sequences as in NetMHCpan 4.0, MHCnuggets 2.0 designed a transfer-learning protocol through an empirical, bottom-up approach to regenerate weights between two similar HLA allotypes. This protocol has shown to improve the prediction performance for most alleles predicted by the tool.

ConvMHC [33] is a machine learning-based tool, which uses a deep convolutional NN (DCNN) for pan-specific peptide-MHC class I-binding prediction. The algorithm generates a 'pixel'-like matrix that encodes the residues of a peptide as the height (H) and the sequence of the binding area of a corresponding HLA molecule as the width (W). In this image-like array (ILA) data, each contact area between a peptide and an HLA molecule is defined as a 'pixel'. In addition, a C channel vector that contains the value of physicochemical properties of the amino acid interaction pair in the 'pixel' is also included in this ILA data. ConvMHC selects nine physicochemical properties and converts them into scores assigned to corresponding amino acids. The size of the C channel is 18 as each pair of 'pixels' has 2 amino acids. ConvMHC uses 34 HLA-I contact residues provided by NetMHCpan [86] and is trained on 9-mer peptides. Thus, the size of ILA is $34 (W) \times 9 (H) \times 18$ (channels). The algorithm of ConvMHC is based on the DCNN architecture described in [87] and uses the dropout [88] as the regularization method. The network uses ReLU [89] as an activation function to transfer the non-linear output of each layer and the Adam optimizer with a learning rate of 0.001 for optimization.

HLA-CNN [34], similar to ConvMHC, also uses DCNN as the machine-learning algorithm. The difference between the two is that instead of encoding the input as a 'pixel', HLA-CNN uses the skip-gram model [90], which is a useful technique used in natural language processing, to embed a peptide sequence into a 15-dimensional vector space. The output of this embedding layer is a two-dimensional matrix of size $L \times 15$ (L is the length of a

given peptide). The DCNN architecture of HLA-CNN consists of two convolutional layers and a fully connected layer. The output matrix of the embedding layer is transferred into the first one-dimensional convolutional layer, which has 32 filters with the length 7. This layer then returns the matrices of the size $L \times 32$ to the next convolutional layer, which also has 32 filters with the length of 7. The function of this one-dimensional convolutional layer can be defined as follows:

$$G[i, k] = F_k * H = \sum_{u=0}^M F_k[u, v] H[i - u, M - v], \quad (7)$$

where F_k is the k^{th} filter, H is the input matrix, G is the output matrix, M is the column size of H minus 1 and u ranges from $-\left\lfloor \frac{\text{filter length}}{2} \right\rfloor$ to $\left\lfloor \frac{\text{filter length}}{2} \right\rfloor$. Then the output of the 2nd convolutional layer is reshaped into a single 288-dimensional vector that is fully connected to the next layer. The sigmoid function is used as the activation function. Finally, this layer is fully connected to a logistic regression output unit. Like ConvMHC, the HLA-CNN uses the dropout method for regularization of the convolutional layers and the Adam optimizer with a learning rate of 0.004. Different from ConvMHC, HLA-CNN uses LeakyReLU [91] as the activation function.

MHCflurry 1.2.0 [31] is developed based on a feed-forward network that consists of zero, one or two locally connected layers and a fully connected layer. The sequence is first encoded into a matrix of the size 15×21 using BLOSUM62 as the encoding scheme. To maintain the position of the anchor residue after the sequence encoding, a 'no-residue' character ('X') was introduced to fill the missing residues to generate a 15-mer peptide representation. The activation function of the hidden layers is the hyperbolic tangent function, while the sigmoid function is used in the output layer. For optimization, MHCflurry 1.2.0 uses a modified mean squared error loss function L as the metric to evaluate the training data with affinity measurements and positive predictive value to evaluate MS peptidomics data. The loss function L is defined as

$$L(\hat{y}, y) = \frac{1}{n} \sum_i^n I(\hat{y}_i, y_i), \quad (8)$$

$$I(\hat{y}_i, y_i) = \begin{cases} (\max(\hat{y}_i - y_i, 0))^2 & \text{for MS qualitative measurement and if } y_i < \hat{y}_i \\ (\max(y_i - \hat{y}_i, 0))^2 & \text{for MS qualitative measurement and if } y_i > \hat{y}_i \\ (\hat{y}_i - y_i)^2 & \text{for quantitative affinity measurement} \end{cases} \quad (9)$$

where n denotes the number of measurements, and \hat{y}_i and y_i denote the predicted and measured values for the i -th measurement, respectively. The MS identified ligands were assigned with the value '<500 nM'. The negative data points for MS evaluation were peptide ligands from different HLA allotypes.

Consensus methods

The idea of consensus methods is that prediction performance can be further improved by integrating the outputs from several

individual tools based on a weighted scheme. Several benchmarking studies have shown that an improved prediction performance can be achieved by consensus methods that average the prediction scores from multiple individual predictors [24, 54, 92]. IEDB-AR-Consensus [25] is such a method, whose results are based on prediction outcomes from three sources: (i) NetMHC 4.0; (ii) SMM [75] and (iii) CombLib [93]. IEDB-AR-Consensus is recommended by the IEDB peptide-binding prediction platform [25]. The platform includes several prediction tools, for which the query peptide and the HLA allotype exist in training data sets of the consensus method.

NetMHCcons 1.1 [35] is another representative peptide-binding prediction tool based on the consensus approach. Specifically, two NN-based prediction tools and one PSSM-based tool have been included in this consensus method, namely (i) NetMHC 3.4 [94, 95], (ii) NetMHCpan 2.8 [86, 96] and (iii) Pickpocket 1.1 [24]. NetMHCcons offers three options to predict peptide binding to different HLA allotypes. Each of the included tools was first benchmarked individually and performance evaluated prior to the application of the consensus method. Finally, the NetMHCcons method is defined as follows:

$$\text{NetMHCcons} = \begin{cases} \text{NetMHC} + \text{NetMHCpan} & \text{for } D = 0 \\ \text{NetMHCpan} & \text{for } 0 < D < 0.1 \\ \text{NetMHCpan} + \text{Pickpocket} & \text{for } D \geq 0.1 \end{cases}, \quad (10)$$

where D refers to the distance between the query HLA allotype and its nearest neighbour in the reference HLA allotype list.

Webserver/software functionality

All peptide-HLA-binding prediction tools reviewed here are either accessible via an online web server and/or are available for download as local stand-alone software. This enables researchers to conduct the prediction in an easy and productive manner. In this section, the general functionalities of the currently available tools are discussed.

For prediction tools with accessible web servers, users normally need to submit the peptide sequence(s) of interest and specify the HLA allotype for which binding is to be predicted. RANKPEP, Pickpocket 1.1, SMMPMBEC, NetMHC 4.0, NetMHCpan 4.0, ConvMHC, IEDB-AR-Consensus, NetMHCcons 1.1 and NetMHCstabpan 1.0 allow users to upload a file with multiple sequences in the FASTA format or submit sequences in the FASTA/PEPTIDE format directly. However, SYFPEITHI can only accept a single sequence at each run of prediction. Note that for all tools that can predict multiple sequences at a time, the maximum number of sequences is still limited (i.e. allowing either ≤ 200 sequences or ≤ 10 MB of the uploaded file for SMMPMBEC and IEDB-AR-Consensus, ≤ 5000 sequences and $\leq 20\,000$ amino acids of each sequence for other tools that have mentioned the data size). In addition to being implemented as a web server, Pickpocket 1.1, SMMPMBEC, NetMHC 4.0, NetMHC 4.0, IEDB-AR-Consensus, NetMHCcons 1.1 and NetMHCstabpan 1.0 have also been made available as stand-alone software for download from their websites.

For a tool that has been implemented as an online web server, a well-designed user-friendly interface can undoubtedly

enhance the efficiency during operation and save unnecessary time otherwise consumed in familiarizing with the tool. To this point, detailed instructions of step-by-step operations have been provided for all tools with webserver, with an exception of SYFPEITHI and ConvMHC. In addition, interpretable outputs can also be found in well-marked places at the web servers of Pickpocket 1.1, SMMPMBEC, NetMHC 4.0, NetMHCpan 4.0, IEDB-AR-Consensus, NetMHCcons 1.1 and NetMHCstabpan 1.0, which has further improved the interpretation of the generated results in terms of the detailed explanation of outcomes. Moreover, tools like RANKPEP, Pickpocket 1.1, NetMHC 4.0, NetMHCpan 4.0, NetMHCcons 1.1 and NetMHCstabpan 1.0 also allow for specific parameters, including the prediction thresholds and weights, to be user-defined or to be set to default values. Also, except for RANKPEP, optional email notifications of accomplished predictions with job IDs can be selected in those tools mentioned above as well as in SMMPEMBEC and IEDB-AR-Consensus, which facilitates a future revisit of results. Moreover, users can also choose to download the results directly from these tools for a further in-depth analysis off-line.

Performance evaluation strategy

Performance evaluation strategies including k -fold cross-validation (CV) and independent tests are widely used for performance evaluation of peptide-binding prediction tools and for further optimization. For k -fold CV, the data set of the tool is divided into k partitions. The validation procedure will be performed k times. Each time one of the k partitions is selected as a validation data set, while the remaining (i.e. $k-1$) partitions are used to train the model. The prediction performance of the trained model then would be evaluated using the validation data set. The final performance is calculated as the average performance over those k individual performances. The leave-one-out (LOO) CV is called when k equals the total number of data entries. This method is much more thorough compared with k -fold CV. The data set is first divided into N parts, where N is the number of entries in the entire data set. The training process is carried out N times, and each time one entry will be validated based on the model that has been trained on the remaining ($N-1$) entries. The overall performance is the average over all N training processes. Normally, the CV method is used to test the internal performance and to avoid the overfitting of the model. Independent test is another popular strategy used for performance evaluation of the tools. Here, the independent test data, which is non-overlapping with the training data of the tools, is collected and becomes an independent test data set. This independent test data set is used as a uniform validation data source to test the performance of different tools. Therefore, compared to the CV method, the performance of different tools evaluated on the same independent data test are comparatively more objective, indicative of the tools' generalization ability and can be compared mutually.

As shown in Table 1, except for SYFPEITHI for which there was no information about the evaluation methods used, all remaining tools reviewed here were evaluated by performing CV and/or independent tests. Tools such as PSSMHCpan 1.0, NetMHCpan 4.0, ConvMHC, HLA-CNN and NetMHCstabpan 1.0 used both methods for performance evaluation. Tools that were evaluated using CV only included Pickpocket 1.1, SMMPMBEC, NetMHC 4.0 and MHCnuggets 2.0, while RANKPEP, MixMHCpred 2.0.1, MHCflurry 1.2.0, IEDB-AR-Consensus and NetMHCcons 1.1 were evaluated using the independent test only in their origi-

nal studies. Here, to allow a fair comparison, the performance benchmarking of all these tools is conducted using a curated up-to-date independent test data set aforementioned.

To measure the prediction performance of the different tools, four commonly used performance metrics for evaluating the algorithms' performance were employed [97]. These include accuracy (Acc), sensitivity (Sn), specificity (Sp) and the Matthews Correlation Coefficient (MCC). The MCC was chosen to enable a more balanced assessment of tools that are developed based on different data sets. These four performance metrics are defined as follows:

$$\begin{cases} \text{Sn} = 1 - \frac{N_{-}^{+}}{N_{-}^{+}} & 0 \leq \text{Sn} \leq 1 \\ \text{Sp} = 1 - \frac{N_{+}^{-}}{N_{+}^{-}} & 0 \leq \text{Sp} \leq 1 \\ \text{Acc} = \frac{N_{+}^{+} + N_{-}^{-}}{N_{+}^{+} + N_{-}^{+} + N_{+}^{-} + N_{-}^{-}} & 0 \leq \text{Acc} \leq 1, \\ \text{MCC} = \frac{1 - \left(\frac{N_{+}^{+}}{N_{+}^{+} + N_{-}^{+}} + \frac{N_{-}^{-}}{N_{+}^{-} + N_{-}^{-}} \right)}{\sqrt{\left(1 + \frac{N_{-}^{+} - N_{+}^{+}}{N_{+}^{+}} \right) \left(1 + \frac{N_{+}^{-} - N_{-}^{-}}{N_{-}^{-}} \right)}} & -1 \leq \text{MCC} \leq 1 \end{cases} \quad (11)$$

where N_{+}^{+} , N_{-}^{+} , N_{+}^{-} and N_{-}^{-} represent the numbers of positives, false negatives, negatives and false positives, respectively.

In addition, in this study, the receiver operating characteristic (ROC) curves were also used to visualize the performance of different methods, with the area under the curve (AUC) calculated to quantify their performance.

Results and discussion

Conservation analysis of sequence motifs of the HLA-I ligands

The binding motif of HLA-I allotypes reflects sequence characteristics of peptide ligands that facilitate binding to the antigen-binding cleft of specific HLA-I allotypes. HLA-I allotype-specific scoring matrices can be established based on these conserved peptide-binding motifs. We analysed the positional preferences of amino acids for allotype-specific peptide ligands across our curated validation data set. For each HLA allotype we generated consensus-binding motifs using the pLogo program [98], which can be visualized in Figure 2 and Figure S1. Irrespective of the ligand length, peptides derived from the same HLA-I allotype show the consensus-binding motifs. For instance, the peptides binding to HLA-A*02:04 and HLA-A*24:06 prefer Leu and Tyr at position 2 ('P2'), as well as Leu/Val or Phe at their C-termini (Figure 2D-F; Figure S2A, B, C). In addition, it can be seen that preferential amino acid patterns exist in the peptides binding to closely related alleles, for example for HLA-B*27:x (x = 01, 07, 08 and 09), where Arg is often required at P2 (Figure 2G-I; Figure S2G-O). However, a closer look revealed that HLA-B*27:01 also preferred to have Arg at the P1 position, while other allotypes did not have this requirement. Moreover, the C-termini of ligands binding to HLA-B*27:01 were relatively diverse compared to HLA-B*27:x (x = 07, 08, 09) that had a strong preference for Leu and Phe at this position.

Performance evaluation of different tools for peptide-binding HLA-I prediction

The performance of different tools was assessed in terms of five commonly used metrics, namely AUC, Sn, Sp, Acc and MCC using the validation data set as an input. It should be noted that the training data sets of some reviewed tools are not available, while several other tools have been updated with an expanded training

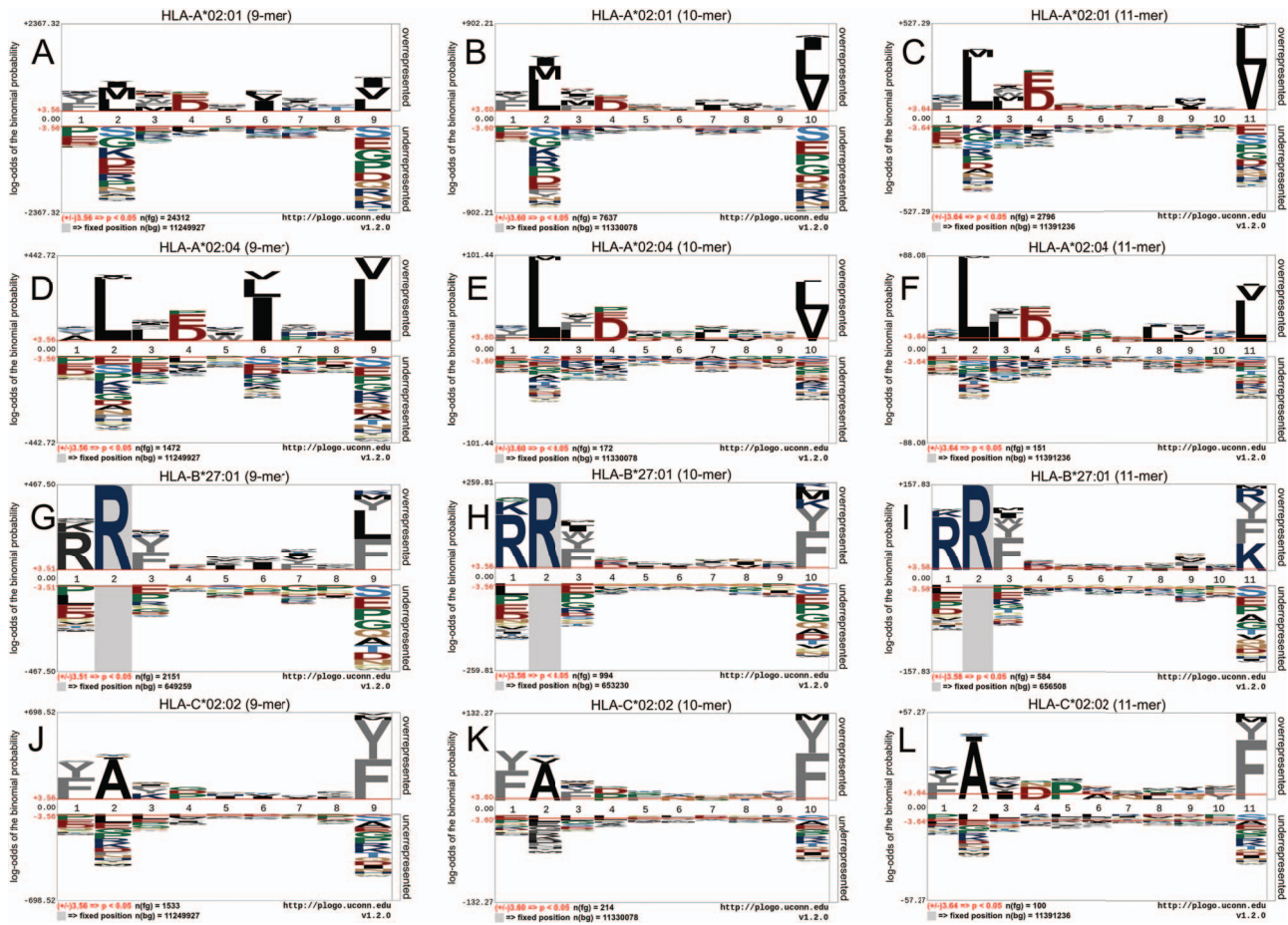


Figure 2. Position and residue specificity of four HLA-I alleles, including (A, B, C) HLA-A*02:01 (9, 10 and 11 mers), (D, E, F) HLA-A*02:04 (9, 10 and 11 mers), (G, H, I) HLA-B*27:01 (9, 10 and 11 mers) and (J, K, L) HLA-C*02:02 (9, 10 and 11 mers).

data set since their first release. Therefore, there might be some overlap between the data sets used for developing some tools and our validation data set. Whenever possible, we downloaded the training data sets of these tools and removed the overlapping entries from our validation data set. Then we submitted the 9-, 10- and 11-mer peptide sequences specific for each HLA-I allotype in the validation data set to the tools. For this evaluation, the tools' parameters were set to the recommended configurations in the corresponding publications or to the default values if no recommendations were given. To illustrate the prediction performance of each tool, the ROC curves with the calculated AUC values were plotted and shown in Figure 3, Figure S2 and Table S2. Moreover, the performance results evaluated in terms of Sn, Sp, Acc and MCC for each tool are given in Table S2. Of note, the evaluation of prediction performance is dependent on the composition of the training data the reviewed predictors have provided.

MixMHCpred 2.0.1 achieved the best performance among all scoring function-based tools as it achieved the highest AUC values among nearly all 19 allotypes, while NetMHCpan 4.0 performed best among all machine learning-based tools, and NetMHCcons 1.0 achieved a better performance than IEDB-AR-Consensus in the consensus category. While no tool universally achieved the best performance for all HLA-I allotypes in the independent test data set, MixMHCpred 2.0.1 performed best for most HLA-I allotypes examined. We speculate that one reason for MixMHCpred 2.0.1 achieving the best performance is

that it is a recently published tool trained with both public HLA-peptide data sources from 40 cell lines and also in-house data from immunoaffinity purification experiments involving 10 additional cell lines. In addition, MixMHCpred 2.0.1 applied both fully unsupervised and semi-supervised machine-learning strategies to identify a total of 52 HLA-I allomorph-specific binding motifs.

Among the machine learning-based tools, NetMHCpan 4.0 achieved the largest AUC values considering all HLA-I allotypes. This is possibly because NetMHCpan 4.0 represents the latest version of NetMHCpan series, which were trained using both experimental affinity measurements and MS-identified peptide ligands. Moreover, NetMHCpan 4.0 uses the pseudo-sequences of HLA-binding pockets to calculate the similarities in ligand binding between different HLA allotypes. Therefore, for HLA allotypes with little binding data NetMHCpan 4.0 is able to achieve a high-prediction performance compared to other machine learning-based tools, e.g. HLA-A*24:06 (9 mers) (AUC=0.989; Figure S2A), HLA-A*24:13 (9 mers) (AUC=0.974; Figure S2B), HLA-B*35:08 (9 mers) (AUC=0.909; Figure S2L) and HLA-C*03:04 (10 mers) (AUC=0.977; Figure S2T). As a comparison, other tools only obtained high AUC values when predicting peptide binding to well-studied HLA-I allotypes. For instance, MHCnuggets 2.0 achieved an AUC value of 0.877, 0.955, 0.942, 0.985 and 0.985 for HLA-B*27:01 (9 mers), -B*27:09 (10 mers), -B*27:09 (11 mers), -B*56:01 (9 mers) and -C*03:04 (9 mers), respectively (Figure 3G; Figure S2J, K, O, S).

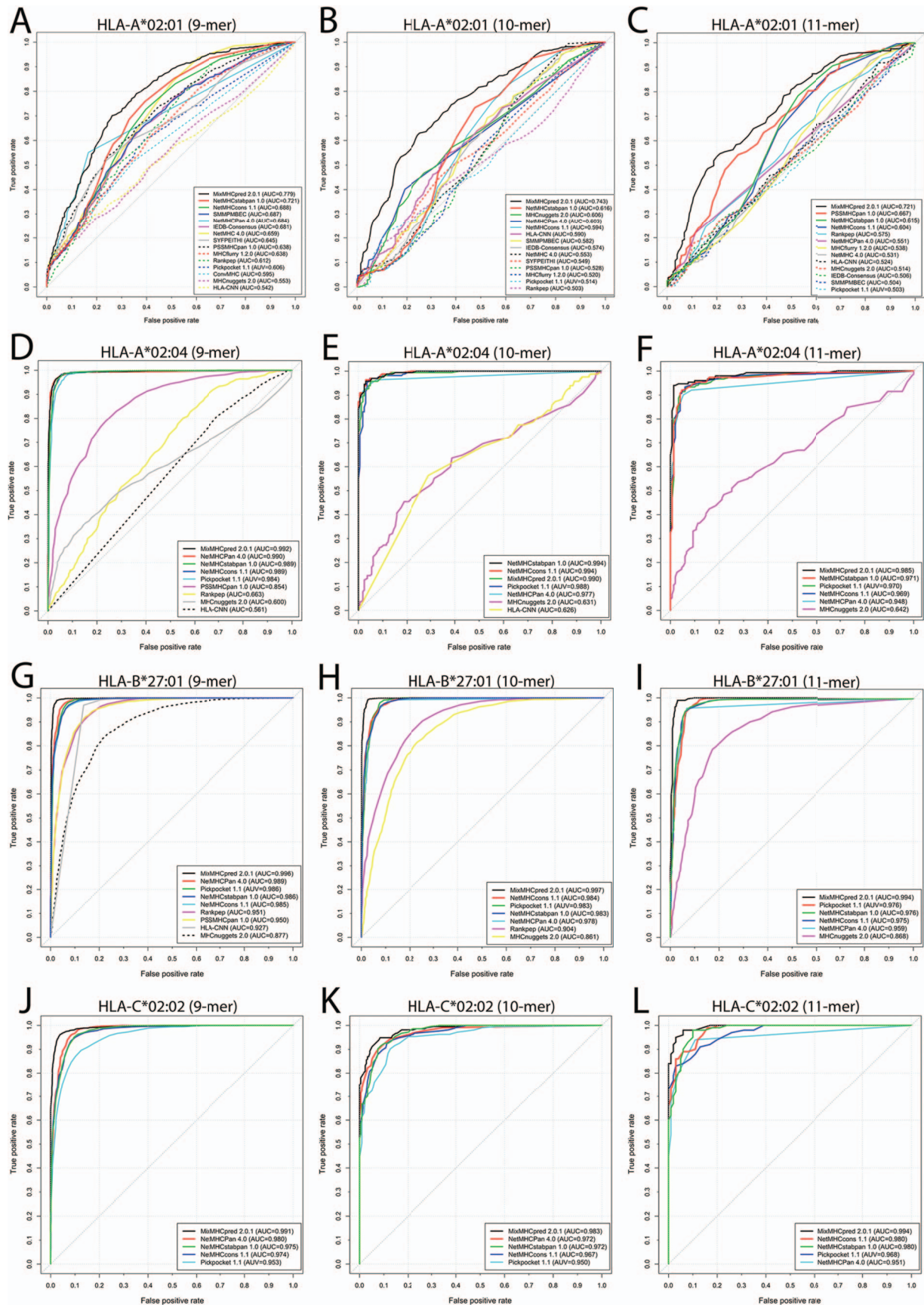


Figure 3. ROC curves and the corresponding AUC values of the reviewed predictors for peptides with lengths of 9, 10 and 11, binding to HLA-I molecules specific for (A, B, C) HLA-A*02:01 (9, 10 and 11 mers), (D, E, F) HLA-A*02:04 (9, 10 and 11 mers), (G, H, I) HLA-B*27:01 (9, 10 and 11 mers) and (J, K, L) HLA-C*02:02 (9, 10 and 11 mers).

As for the consensus methods, NetMHCcons 1.1 achieved better AUC values compared with IEDB-AR-Consensus. The NetMHCcons 1.1 used different types of combinations of peptide-binding prediction tools according to the query pair of peptide and HLA-I allotype as discussed in the section of consensus methods. This might be the primary reason the NetMHCcons 1.1 achieved the best prediction performance. For instance, NetMHCcons 1.1 also utilizes pseudo-sequences of HLA-binding pockets used in NetMHCpan 4.0 to predict binding to HLA allotypes with little binding data, e.g. HLA-A*24:06 (9 mers) (AUC = 0.986; Figure S2A), HLA-A*24:13 (9 mers) (AUC = 0.971; Figure S2B), HLA-B*35:08 (9 mers) (AUC = 0.884; Figure S2L) and HLA-C*03:04 (10 mers) (AUC = 0.989; Figure S2T).

Prospective strategies for improving the prediction performance of antigenic peptides and developing next-generation bioinformatics tools

Based on our independent evaluation of prediction performance, we foresee that peptide-binding prediction will continue to be improved as the volume and quality of training data increases (such as observed for MixMHCpred 2.0.1). Similarly, machine-learning algorithms (such as NetMHCpan 4.0) also benefit from improved training data sets. Such ligand prediction tools have already facilitated the discovery of new epitopes in many diseases and cancers [99–102]. However, only a very small proportion of the predicted binders are naturally targeted by T cells (i.e. T-cell epitopes) [3]. In this context, we also used the IEDB T-cell assay data set, which contains experimentally verified immunogenic and non-immunogenic peptides, to evaluate the performance of the prediction tools reviewed here. As it turned out, independent of the predictor, those peptides that were predicted to be binders are mostly experimentally verified non-immunogenic according to IEDB (for peptides that are verified as non-immunogenic peptides and predicted as binders, more than 95% of are pathogen-derived peptides, data not shown). This result suggests that the prediction of immunogenicity cannot solely rely on the binding affinity between HLA allotypes and peptides. The immunogenicity of a peptide is associated with various properties. These include the available T-cell receptor repertoire, ability of the peptide to be effectively processed and liberated from the parental antigen, the duration and context of presentation and binding to HLA allotypes and so on [4, 103]. A recent study shows that more than 80% of MHC class I-bound peptides derived from virus can be immunogenic in virus-infected mouse. The study demonstrated that the major CD8⁺ T-cell responses are strongly associated with high affinity of peptide to MHC class I molecule. Besides that, the study also showed evidence and pointed out that the peptide abundance and the time of gene expression may play roles in the CD8⁺ T-cell immunity [104]. This finding suggests that most of presented viral MHC class I peptides may be immunogenic, but the T-cell response may be dominated by a few peptides. Thus, how to design a prediction method that can predict dominant T-cell epitopes is still a challenge. We believe that integrating peptide properties related to immunogenicity into an algorithm using cutting-edge machine-learning techniques is a promising direction for improving the prediction of immunogenicity. To this end, we provide several insights for further discussion and future directions.

Firstly, to identify potential epitopes efficiently, peptides that have been experimentally annotated as antigenic, i.e. being able to induce a T-cell response and peptides that are clearly presented yet fail to elicit a response, as high affinity of peptides

is reported to be strongly associated with T-cell responses [104], are the ideal data source for extracting informative features and constructing accurate prediction models. Various features such as physicochemical properties of peptides, the stability of pHLAs complex, antigen processing and the structural contact between pHLAs and TCRs have been shown to influence the immunogenicity of peptides [48, 105–107]. More accurate epitope prediction may be achieved by considering these related features along with peptide binding. In addition to proteomic information, the discovery of immunogenic peptides may also be facilitated using genomic data generated from NGS techniques, which might provide potentially useful information such as identifying somatic mutations within tumour cells to aid the identification of neo-antigens for personalized cancer immunotherapy [108–110].

Secondly, the use of DL is also a promising direction in improving the performance of immunogenic peptide prediction. Compared to conventional machine-learning algorithms, DL-based methods can introduce innovative network architectures and regularization techniques, which allow the functions of the algorithms to be trained to simulate the complexity of peptide binding and T-cell recognition, while avoiding common issues faced by machine-learning methods such as overfitting and slow convergence [111, 112]. Despite the time-consuming model training process, DL-based methods have been shown to outperform other methods in a number of different research areas given sufficient training data [32, 111]. In addition to that, DL methods are particularly suitable for undertaking high-throughput prediction tasks, particularly because they do not need to involve and use feature encodings to represent the original data samples. Thus, DL methods are attractive and promising for analysing the rapidly increasing amount of available data generated by advanced high-throughput techniques.

With the rapid development in the field of machine learning, several novel algorithms have been proposed that exhibit promising prediction performance and are attractive for researchers. For instance, a novel incremental decision tree-learning algorithm, the Hoeffding Anytime Tree, has been recently developed. It is based on the conventional Hoeffding Tree [113] but with a minor modification that allows it to achieve a significantly improved predictive performance on most of the largest classification data sets [114]. Moreover, algorithms like reinforcement learning can improve the prediction accuracy further and autonomously by receiving reward or punishment according to the model performance [115, 116]. This property is also attractive for immunogenicity predictions because the results of experimental tests can be fed back to improve the efficiency and accuracy of the model.

Conclusion

Once liberated from the parental protein, the selection and binding of peptide antigens to available HLA allotypes is the first critical step in influencing the immune response and ultimately the survival of the host. For this reason, tremendous efforts and resources have been applied to accurately predict which pathogen-derived or neo-antigen-derived peptides are selected for antigen presentation. Available tools now do quite a respectable job for predicting a peptide's ability to bind to a given HLA allotype, even in the absence of training data sets for a specific HLA allomorph. The development of these prediction tools enables immunologists to narrow down the search space of antigen candidates that need to be experimentally validated. However, the integration of other host-specific information will

be essential to predict which of the peptides present on the cells surface will be targeted for T-cell responses. In addition, advanced machine-learning algorithms like DL are ideal for processing such extremely large information from the host and generating key features for predicting immunogenicity. Therefore, the combination of skills from immunologists and bioinformaticians will assist the development of immunogenicity prediction in a faster and more efficient way. In this review and survey work, we have introduced and comprehensively evaluated the currently available tools for the prediction of peptides binding to different HLA-I molecules. Additionally, we have discussed, reviewed and assessed all methods in terms of their calculation methods of the prediction score, prediction algorithm(s), functionality and performance evaluation strategy. To obtain a more objective performance evaluation, we constructed an independent test data set to benchmark all tools. Based on the assessment results, MixMHCpred 2.0.1 achieved the highest prediction performance across most of the HLA-I molecules and is the best predictor among scoring function-based tools. Apart from that, NetMHCpan 4.0 and NetMHCcons 1.1 generally achieved the best performance results in the machine learning-based and consensus-based tools, respectively. This study provides useful guidance to researchers who are interested in developing an antigen prediction model in future studies. Additionally, feedback of data on immunogenicity into such models will improve our understanding of the antigen-processing pathways and subsequent T-cell recognition patterns. Finally, we hope that more accurately predicted peptide binding will assist with the development of immunotherapy and vaccine design.

Key Points

- We conducted a comprehensive review and assessment of 15 currently available tools for predicting human leukocyte antigen (HLA) class I (HLA-I)-binding peptides, including 6 scoring function-based, 7 machine learning-based and 2 consensus methods.
- This review and survey systematically analysed these tools with respect to the computational methods of the prediction score, employed algorithms, performance evaluation strategies and software functionality.
- All tools underwent a comprehensive performance assessment based on 19 different HLA-I allotypes and up-to-date independent data sets of experimentally verified HLA-I allotype-specific ligands.
- Extensive benchmarking tests show that MixMHCpred 2.0.1 performs best across most of HLA-I allotypes included in the validation data sets, while NetMHCpan 4.0 and NetMHCcons 1.1 achieve the overall best performance among machine learning-based and consensus-based tools, respectively.
- This study provides a comprehensive analysis and benchmarking of currently available bioinformatics tools for HLA-I peptide-binding prediction and gives directions to the wider research community for developing the next generation of peptide-binding prediction tools.

Supplementary Data

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

Funding

This work was supported by grants from the National Health and Medical Research Council of Australia (NHMRC) (APP1085018, APP1127948, APP1144652 and APP1084283), the Australian Research Council (ARC) (DP120104460, DP150104503) and the Collaborative Research Program of Institute for Chemical Research, Kyoto University (2018-28). J.L., A.W.P. and J.N.S. are supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health (R01 AI111965). J.R. is an ARC Australia Laureate fellow. J.L. and A.W.P. are NHMRC principal research fellows. T.T.M.L. and A.L.'s work was supported in part by the Informatics Institute of the School of Medicine at UAB.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Allergy and Infectious Diseases or the National Institutes of Health.

References

1. Blum JS, Wearsch PA, Cresswell P. Pathways of antigen processing. *Annu Rev Immunol* 2013;**31**:443–73.
2. Gfeller D, Bassani-Sternberg M. Predicting antigen presentation—what could we learn from a million peptides? *Front Immunol* 2018;**9**:1716.
3. 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.
4. Purcell AW, McCluskey J, Rossjohn J. More than one reason to rethink the use of peptides in vaccine design. *Nat Rev Drug Discov* 2007;**6**:404.
5. Koşaloğlu-Yalçın Z, Lanka M, Frentzen A, et al. Predicting T cell recognition of MHC class I restricted neoepitopes. *Oncoimmunology* 2018;**7**:e1492508.
6. Le DT, Uram JN, Wang H, et al. PD-1 blockade in tumors with mismatch-repair deficiency. *N Engl J Med* 2015;**372**:2509–20.
7. Rizvi NA, Hellmann MD, Snyder A, et al. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* 2015;**348**:124–8.
8. Snyder A, Makarov V, Merghoub T, et al. Genetic basis for clinical response to CTLA-4 blockade in melanoma. *N Engl J Med* 2014;**371**:2189–99.
9. Van Allen EM, Miao D, Schilling B, et al. Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science* 2015;**350**:207–11.
10. Ott PA, Hu Z, Keskin DB, et al. An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature* 2017;**547**:217.
11. Sahin U, Derhovanessian E, Miller M, et al. Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. *Nature* 2017;**547**:222.
12. Gfeller D, Bassani-Sternberg M, Schmidt J, et al. Current tools for predicting cancer-specific T cell immunity. *Oncoimmunology* 2016;**5**:e1177691.
13. Linnemann C, Van Buuren MM, Bies L, et al. High-throughput epitope discovery reveals frequent recognition of neo-antigens by CD4+ T cells in human melanoma. *Nat Med* 2015;**21**:81.
14. Bentzen AK, Hadrup SR. Evolution of MHC-based technologies used for detection of antigen-responsive T cells, cancer immunology. *Immunotherapy* 2017;**66**:657–66.

15. Rajasagi M, Shukla SA, Fritsch EF, et al. Systematic identification of personal tumor-specific neoantigens in chronic lymphocytic leukemia. *Blood* 2014;**124**:453–62.
16. Robbins PF, Lu Y-C, El-Gamil M, et al. Mining exomic sequencing data to identify mutated antigens recognized by adoptively transferred tumor-reactive T cells. *Nat Med* 2013;**19**:747.
17. 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**:13404.
18. Bassani-Sternberg M, Chong C, Guillaume P, et al. Deciphering HLA-I motifs across HLA peptidomes improves neoantigen predictions and identifies allosteric regulating HLA specificity. *PLoS Comput Biol* 2017;**13**:e1005725.
19. 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.
20. Vita R, Overton JA, Greenbaum JA, et al. The immune epitope database (IEDB) 3.0. *Nucleic Acids Res* 2014;**43**:D405–12.
21. Nielsen M, Lund O, Buus S, et al. MHC class II epitope predictive algorithms. *Immunology* 2010;**130**:319–28.
22. Rammensee H-G, Bachmann J, Emmerich NPN, et al. SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics* 1999;**50**:213–9.
23. Reche PA, Glutting J-P, Reinherz EL. Prediction of MHC class I binding peptides using profile motifs. *Hum Immunol* 2002;**63**:701–9.
24. 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.
25. Kim Y, Ponomarenko J, Zhu Z, et al. Immune epitope database analysis resource. *Nucleic Acids Res* 2012;**40**:W525–30.
26. 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.
27. 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.
28. Andreatta M, Nielsen M. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics* 2015;**32**:511–7.
29. 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.
30. 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.
31. O'Donnell TJ, Rubinsteyn A, Bonsack M et al, MHCflurry: open-source class I MHC binding affinity prediction. *Cell Syst* 2018;**7**:129–32.
32. Bhattacharya R, Tokheim C, Sivakumar A, et al. Prediction of peptide binding to MHC Class I proteins in the age of deep learning. *bioRxiv* 2017, <https://doi.org/10.1101/154757>.
33. Han Y, Kim D. Deep convolutional neural networks for pan-specific peptide-MHC class I binding prediction. *BMC Bioinformatics* 2017;**18**:585.
34. Vang YS, Xie X. HLA class I binding prediction via convolutional neural networks. *Bioinformatics* 2017;**33**:2658–65.
35. 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.
36. 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**:2239–55.
37. Rognan D, Lauemøller SL, Holm A, et al. Predicting binding affinities of protein ligands from three-dimensional models: application to peptide binding to class I major histocompatibility proteins. *J Med Chem* 1999;**42**:4650–8.
38. Altuvia Y, Margalit H. A structure-based approach for prediction of MHC-binding peptides. *Methods* 2004;**34**:454–9.
39. Liao WW, Arthur JW. Predicting peptide binding affinities to MHC molecules using a modified semi-empirical scoring function. *PLoS One* 2011;**6**:e25055.
40. Knapp B, Giczi V, Ribarics R, et al. PeptX: using genetic algorithms to optimize peptides for MHC binding. *BMC Bioinformatics* 2011;**12**:241.
41. Yanover C, Bradley P. Large-scale characterization of peptide-MHC binding landscapes with structural simulations. *Proc Natl Acad Sci U S A* 2011;**108**:6981–6.
42. Doytchinova IA, Flower DR. Physicochemical explanation of peptide binding to HLA-A* 0201 major histocompatibility complex: a three-dimensional quantitative structure-activity relationship study. *Proteins* 2002;**48**:505–18.
43. Doytchinova IA, Walshe VA, Jones NA, et al. Coupling in silico and in vitro analysis of peptide-MHC binding: a bioinformatic approach enabling prediction of superbinding peptides and anchorless epitopes. *J Immunol* 2004;**172**:7495–502.
44. Jojic N, Reyes-Gomez M, Heckerman D, et al. Learning MHC I—peptide binding. *Bioinformatics* 2006;**22**:e227–35.
45. Antes I, Siu SW, Lengauer T. DynaPred: a structure and sequence based method for the prediction of MHC class I binding peptide sequences and conformations. *Bioinformatics* 2006;**22**:e16–24.
46. Bordner AJ, Abagyan R. Ab initio prediction of peptide-MHC binding geometry for diverse class I MHC allotypes. *Proteins* 2006;**63**:512–26.
47. Tian F, Yang L, Lv F, et al. In silico quantitative prediction of peptides binding affinity to human MHC molecule: an intuitive quantitative structure-activity relationship approach. *Amino Acids* 2009;**36**:535.
48. Saethang T, Hirose O, Kimkong I, et al. PAAQD: predicting immunogenicity of MHC class I binding peptides using amino acid pairwise contact potentials and quantum topological molecular similarity descriptors. *J Immunol Methods* 2013;**387**:293–302.
49. Mukherjee S, Bhattacharyya C, Chandra N. HLaffy: estimating peptide affinities for Class-I HLA molecules by learning position-specific pair potentials. *Bioinformatics* 2016;**32**:2297–305.
50. Wan S, Knapp B, Wright DW, et al. Rapid, precise, and reproducible prediction of peptide-MHC binding affinities from molecular dynamics that correlate well with experiment. *J Chem Theory Comput* 2015;**11**:3346–56.
51. Knapp B, Demharter S, Deane CM, et al. Exploring peptide/MHC detachment processes using hierarchical natural move Monte Carlo. *Bioinformatics* 2015;**32**:181–6.
52. Peters B, Bui H-H, Frankild S, et al. A community resource benchmarking predictions of peptide binding to MHC-I molecules. *PLoS Comput Biol* 2006;**2**:e65.

53. Lin HH, Ray S, Tongchusak S, et al. Evaluation of MHC class I peptide binding prediction servers: applications for vaccine research. *BMC Immunol* 2008;9:8.
54. Zhang H, Lundegaard C, Nielsen M. Pan-specific MHC class I predictors: a benchmark of HLA class I pan-specific prediction methods. *Bioinformatics* 2008;25:83–9.
55. 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.
56. Zhao W, Sher X. Systematically benchmarking peptide-MHC binding predictors: from synthetic to naturally processed epitopes. *PLoS Comput Biol* 2018;14:e1006457.
57. Zhang Q, Wang P, Kim Y, et al. Immune epitope database analysis resource (IEDB-AR). *Nucleic Acids Res* 2008;36:W513–8.
58. 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.
59. 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.
60. Alvarez B, Barra C, Nielsen M, et al. Computational tools for the identification and interpretation of sequence motifs in immunopeptidomes. *Proteomics* 2018;18:1700252.
61. Stranzl T, Larsen MV, Lundegaard C, et al. NetCTLpan: pan-specific MHC class I pathway epitope predictions. *Immunogenetics* 2010;62:357–68.
62. Larsen MV, Lundegaard C, Lamberth K, et al. An integrative approach to CTL epitope prediction: a combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions. *Eur J Immunol* 2005;35:2295–303.
63. Yewdell JW, Bennink JR. Immunodominance in major histocompatibility complex class I-restricted T lymphocyte responses. *Annu Rev Immunol* 1999;17:51–88.
64. Chou K-C. Some remarks on protein attribute prediction and pseudo amino acid composition. *J Theor Biol* 2011;273:236–47.
65. 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;1:9.
66. Chen W, Feng P-M, Lin H, et al. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. *Nucleic Acids Res* 2013;41:e68–8.
67. Chou K-C, Wu Z-C, Xiao X. iLoc-hum: using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. *Mol Biosyst* 2012;8:629–41.
68. Lill JR, van Veelen PA, Tenzer S, et al. Minimal information about an immuno-peptidomics experiment (MIAIPE). *Proteomics* 2018;18:1800110.
69. Li F, Zhang Y, Purcell AW, et al. Positive-unlabelled learning of glycosylation sites in the human proteome. *BMC Bioinformatics* 2019;20:112.
70. 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;1:4.
71. Andreatta M, Alvarez B, Nielsen M. GibbsCluster: unsupervised clustering and alignment of peptide sequences. *Nucleic Acids Res* 2017;45:W458–63.
72. Lam L, Suen S. Application of majority voting to pattern recognition: an analysis of its behavior and performance. *IEEE Trans Syst Man Cybern A Syst Hum* 1997;27:553–68.
73. Thompson JD, Higgins DG, Gibson TJ. Improved sensitivity of profile searches through the use of sequence weights and gap excision. *Bioinformatics* 1994;10:19–29.
74. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* 1992;89:10915–9.
75. Peters B, Sette A. Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method. *BMC Bioinformatics* 2005;6:132.
76. 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.
77. Altschul SF, Gertz EM, Agarwala R, et al. PSI-BLAST pseudocounts and the minimum description length principle. *Nucleic Acids Res* 2008;37:815–24.
78. Nielsen M, Lundegaard C, Wornring P, et al. Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. *Bioinformatics* 2004;20:1388–97.
79. Bassani-Sternberg M, Gfeller D. Unsupervised HLA peptide deconvolution improves ligand prediction accuracy and predicts cooperative effects in peptide-HLA interactions. *J Immunol* 2016;197:2492–9.
80. Suliman A, Zhang Y. A review on back-propagation neural networks in the application of remote sensing image classification. *Journal of Earth Science and Engineering* 2015;5:52–65.
81. Nielsen M, Lund O. NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. *BMC Bioinformatics* 2009;10:296.
82. Harndahl M, Rasmussen M, Roder G, et al. Real-time, high-throughput measurements of peptide-MHC-I dissociation using a scintillation proximity assay. *J Immunol Methods* 2011;374:5–12.
83. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* 1997;9:1735–80.
84. Cho K, Van Merriënboer B, Gulcehre C et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. 2014, arXiv preprint arXiv:1406.1078, doi:10.3115/v1/D14-1179.
85. Kinga D, Adam JB. A method for stochastic optimization. In: *International Conference on Learning Representations (ICLR)*. 2014. arXiv preprint arXiv:1412.6980, Ithaca: San Diego.
86. Nielsen M, Lundegaard C, Blicher T, et al. NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. *PLoS One* 2007;2:e796.
87. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014, arXiv preprint arXiv:1409.1556, arXiv:1409.1556v6.
88. Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res* 2014;15:1929–58.
89. Nair V, Hinton GE. Rectified linear units improve restricted boltzmann machines. In: *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*. 2010, pp. 807–14, Ominipress: Haifa.
90. Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality. In: *Advances in Neural Information Processing Systems*, 2013, pp. 3111–9, Curran Associate Inc.: Nevada.

91. Maas AL, Hannun AY, Ng AY. Rectifier nonlinearities improve neural network acoustic models. In: Proceedings of the International Council for Machinery Lubrication, 2013, 3, Atlanta.
92. Moutafsi M, Peters B, Pasquetto V, et al. A consensus epitope prediction approach identifies the breadth of murine T CD8⁺-cell responses to vaccinia virus. *Nat Biotechnol* 2006;**24**:817.
93. Sidney J, Assarsson E, Moore C, et al. Quantitative peptide binding motifs for 19 human and mouse MHC class I molecules derived using positional scanning combinatorial peptide libraries. *Immunome Res* 2008;**4**:2.
94. 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.
95. Lundegaard C, Lund O, Nielsen M. Accurate approximation method for prediction of class I MHC affinities for peptides of length 8, 10 and 11 using prediction tools trained on 9mers. *Bioinformatics* 2008;**24**:1397–8.
96. Hoof I, Peters B, Sidney J, et al. NetMHCpan, a method for MHC class I binding prediction beyond humans. *Immunogenetics* 2009;**61**:1.
97. Li F, Wang Y, Li C, et al. Twenty years of bioinformatics research for protease-specific substrate and cleavage site prediction: a comprehensive revisit and benchmarking of existing methods. *Briefings Bioinform* 2018;bby077, doi:10.1093/bib/bby077.
98. O'shea JP, Chou MF, Quader SA, et al. pLogo: a probabilistic approach to visualizing sequence motifs. *Nat Methods* 2013;**10**:1211.
99. Samuels Y, Kalaora S, Wolf Y, et al. Combined analysis of antigen presentation and T cell recognition reveals restricted immune responses in melanoma. *Cancer Discov* 2018;**8**:1366–75.
100. Sakabe S, Sullivan BM, Hartnett JN, et al. Analysis of CD8⁺ T cell response during the 2013–2016 Ebola epidemic in West Africa. *Proc Natl Acad Sci U S A* 2018;**115**:E7578–86.
101. Rozanov DV, Rozanov ND, Chiotti KE, et al. MHC class I loaded ligands from breast cancer cell lines: a potential HLA-I-typed antigen collection. *J Proteomics* 2018;**176**:13–23.
102. Fiore-Gartland A, Manso BA, Friedrich DP, et al. Pooled-peptide epitope mapping strategies are efficient and highly sensitive: an evaluation of methods for identifying human T cell epitope specificities in large-scale HIV vaccine efficacy trials. *PLoS One* 2016;**11**:e0147812.
103. Blankenstein T, Coulie PG, Gilboa E, et al. The determinants of tumour immunogenicity. *Nat Rev Cancer* 2012;**12**:307.
104. Croft NP, Smith SA, Pickering J, et al. Most viral peptides displayed by class I MHC on infected cells are immunogenic. *Proc Natl Acad Sci U S A* 2019;**116**:3112–7, doi:10.1145/3219819.3220005.
105. Kim S, Kim HS, Kim E, et al. Neopepsee: accurate genome-level prediction of neoantigens by harnessing sequence and amino acid immunogenicity information. *Ann Oncol* 2018;**29**:1030–6.
106. Chowell D, Krishna S, Becker PD, et al. TCR contact residue hydrophobicity is a hallmark of immunogenic CD8⁺ T cell epitopes. *Proc Natl Acad Sci U S A* 2015;**112**:E1754–62.
107. Zeng J, Treutlein HR, Rudy GB. Predicting sequences and structures of MHC-binding peptides: a computational combinatorial approach. *J Comput Aided Mol Des* 2001;**15**:573–86.
108. 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.
109. Yadav M, Jhunjhunwala S, Phung QT, et al. Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. *Nature* 2014;**515**:572.
110. Li B, Li T, Pignon J-C, et al. Landscape of tumor-infiltrating T cell repertoire of human cancers. *Nat Genet* 2016;**48**:725.
111. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;**521**:436.
112. Min S, Lee B, Yoon S. Deep learning in bioinformatics. *Brief Bioinform* 2017;**18**:851–69.
113. Domingos P, Hulten G. Mining high-speed data streams. In: *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2000, pp. 71–80. Association for Computing Machinery, New York.
114. Manapragada C, Webb G, Salehi M. Extremely Fast Decision Tree. 2018, arXiv preprint arXiv:1802.08780, doi:10.1145/3219819.3220005.
115. Riedmiller M, Gabel T, Hafner R, et al. Reinforcement learning for robot soccer. *Auton Robots* 2009;**27**:55–73.
116. Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning. *Nature* 2015;**518**:529.