Received: 24 October 2016 | Revised: 22 September 2017 | Accepted: 9 January 2018

DOI: 10.1002/pep2.24046

WILEY PeptideScience

REVIEW

Current methods for the prediction of T-cell epitopes

Prattusha Kar | Lanie Ruiz-Perez | Mahreen Arooj | Ricardo L. Mancera

School of Pharmacy and Biomedical Sciences, Curtin Health Innovation Research Institute and Curtin Institute for Computation, Curtin University, Perth, Western Australia 6845, Australia

Correspondence
Ricardo L. Mancera, School of Pharmacy and Biomedical Sciences, Curtin Health Innovation Research Institute and Curtin Institute for Computation, Curtin University, Perth, WA 6845, Australia.
Email: R.Mancera@curtin.edu.au

Abstract

T-cell epitopes are specific peptide sequences derived from foreign or own proteins that can initiate an immune response and which are recognized by specific T-cells when displayed on the surface of other cells. The prediction of T-cell epitopes is of particular interest in vaccine design, disease prevention and the development of immunotherapeutics. There are two principal categories of predictive methods: peptide-sequence based and peptide-structure-based. Sequence-based methods make use of various approaches to identify likely immunogenic amino acid sequences, such as sequence motifs, decision trees, partial least squares (PLS), quantitative matrices (QM), artificial neural networks (ANN), hidden Markov models (HMM), and support vector machines (SVM). Structure-based methods are more diverse in nature and involve approaches such as quantitative structure-activity relationships (QSAR), molecular modelling, molecular docking and molecular dynamics simulations (MD). This review highlights the key features of all of these approaches, provides some key examples of their application, and compares and contrasts the most important methods currently in use.

KEYWORDS
ANN, CTL, MHC, T-cell epitope, TCR

1 | INTRODUCTION

The human adaptive immune system is responsible for the destruction of foreign microorganisms upon entry to the body.[1] Adaptive immunity is mediated by B- and T-cell lymphocytes, which are white blood cells with the ability to distinguish between the noninfectious self and the infectious nonself. To do this, lymphocytes recognize antigenic molecules displayed on the surface of other cells. If an antigenic molecule is recognized as foreign or dangerous, the immune system mounts an attack,[2] in which case it is regarded as an immunogenic antigen.

An epitope is the part of an antigen that is specifically recognized and binds to antibodies as well as to the structurally and functionally similar surface receptors in B-cells and T-cells. Peptide–protein interactions constitute the molecular basis of this recognition, which occurs between the epitope and the binding site of each of these, namely the paratope as the binding site of the antibody, and the binding groove in the case of B- and T-cell receptors.[3] Protein epitopes can be classified as either continuous (linear) or discontinuous (conformational), depending on their molecular structure and interactions with the paratope. The former class can be directly used for the design of vaccines and immunodiagnostics, while discontinuous epitopes may be used to design molecules that mimic the structure and immunogenic properties of an epitope and have utility as a prophylactic, therapeutic vaccine, or for antibody detection in medical diagnostics or experimental research.[4]

The recognition of epitopes triggers the activation of T-cells and the rest of the adaptive immune system, ultimately leading to the clearance of pathogenic microorganisms.[5,6] The process of antigen presentation is the first step in the activation of the immune system, and requires the interaction between the epitope and two cell-surface molecules: T-cell receptors (TCRs) and major histocompatibility complex (MHC) molecules present in the cell displaying the epitope.[7] The majority of epitopes recognized by a TCR (known as T-cell epitopes) are continuous, amphipathic, and helical in structure.[8] The interactions required for antigen presentation also include the binding of an additional co-receptor to the TCR-epitope-MHC complex. Two populations of T-cells are thus distinguished according to the differential expression of co-receptors: CD4, found in helper T-cells (T_h or CD4+ T-cells), and CD8, found in cytotoxic T-cells (CD8+ T-cells).[9] Consequently different populations of T-cells become activated to initiate their actions on pathogens.[10–12]

MHC molecules are a class of membrane proteins encoded by a highly polymorphic gene family present in all vertebrates and

https://doi.org/10.1002/pep2.24046

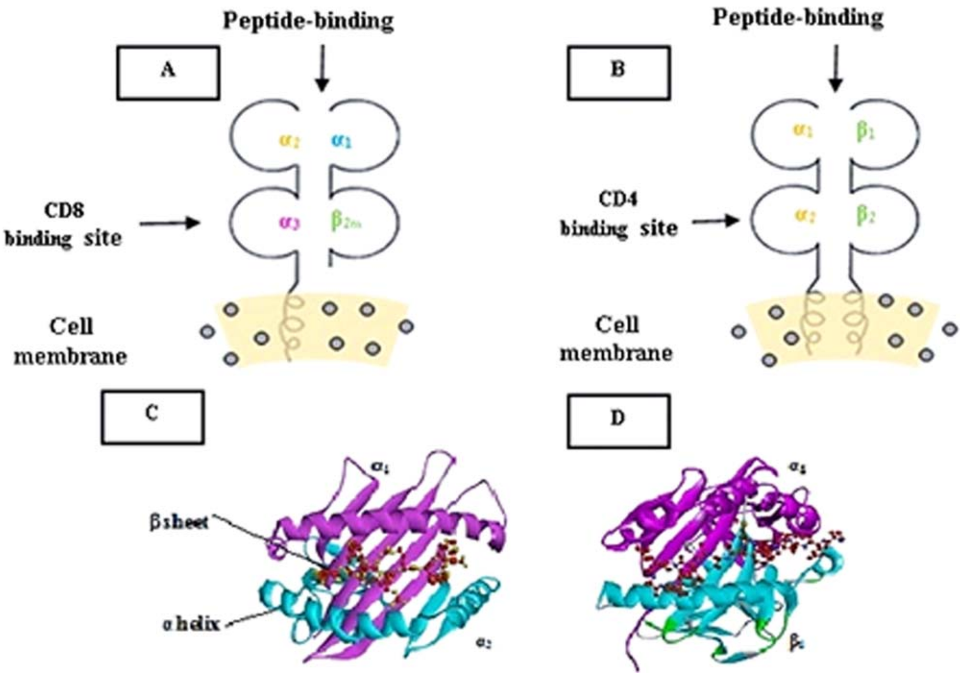

FIGURE 1 MHC I and II molecules in complex with antigenic peptides. A, Structure of MHC-I molecule, B, Structure of MHC-II molecule, C, Peptide binding grove with peptide bound to an MHC-I molecule[22] (PDB code: 5DEF) and D, Peptide binding grove with peptide bound to an MHC-II molecule[23] (PDB code: 1SJE)

responsible for displaying antigens to lymphocytes. In humans, the set of genes that encode MHCs are known as the human leukocyte antigen (HLA) system, in which the HLA-A*02:01 allele has historically been the subject of intense investigation in regards to peptide binding.[13] MHC molecules are further classified into MHC-I and MHC-II according to the type of cells expressing them and thus a variety of subsequent differences, such as their structural subunits, the types of T-cells they engage with, and the class and origin of epitopes they display.

MHC-I molecules are present on the surface of all nucleated cells whereas MHC-II molecules are found mainly on cells of the immune system,[14,15] in particular in antigen-presenting cells (APCs). MHC-I molecules are comprised of two protein chains: α is expressed as a heavy chain comprising three domains, α_1, α_2 and α_3, and β_2-macroglobulin (β_2m), which is referred to as the light chain. The peptide-binding groove is made up of parts of the α_1 and α_2 domains and can bind peptides 8–10 amino acids long.[16] On the other hand, MHC-II molecules are heterodimers consisting of two subunits α and β. Each chain contains two domains (α_1 and α_2, and β_1 and β_2, respectively). The peptide-binding groove is made up of parts of the α_1 and β_1 heterodimers and can bind peptides at least 13 amino acids long but varying in length[17,18] because the peptide ends can extend outside of the groove.[19–21] The generic structures of MHC molecules are shown in Figure 1.

The constitutive α- and β-chains of MHC-II molecules are first joined together in the endothelial reticulum (ER) of APCs and bind transiently to the invariant chain (Ii) to prevent access of endogenous antigens to the binding-groove. MHC-II molecules are then loaded with epitopes derived from the proteolytic digestion of exogenous antigens in the MHC-II compartment (MIIC). The resulting peptide-MHC-II complex is transported to the surface of the APC to be recognized by the TCR on CD4+ T-cells[24–26] (Figure 2B). In the case of the assembly of peptide-MHC-I complex, endogenous antigenic proteins undergo cleavage by the proteasome and the resulting peptides are carried to the ER by TAP (transporter associated with antigen processing), which finally loads them onto the MHC-I molecule. The resulting complex is then transported to the surface of the cell where it is recognized by the TCR on CD8+ T-lymphocites (CTLs) (Figure 2A).

CTLs play a crucial role in the clearance of virally infected cells[27] and secrete chemokines and cytokines to recruit other immune cells as part of cellular immunity.[28] Research conducted in both humans and primates showed that it also plays roles in identifying bacterial infection[29] and inducing apoptosis in tumor cells.[30] Given that proteasomes produce the majority of the peptides that are presented to CTLs (known as CTL epitopes),[25,26,31] they play an essential role in the immune response.[32,33] Nonetheless, from the wide variety of peptides produced by the proteasome during the cleavage of intracellular proteins, only 1 in 200 peptides will actually bind to MHC-I molecules and would be regarded as antigenic.[34,35]

T-cells must scan large numbers of peptide-MHC complexes on multiple cells in order to identify and eliminate threats as quickly as possible, thus effective immunity requires minimal TCR scanning time and maximum antigen coverage.[36] Hence, the affinity between a TCR and a peptide-MHC complex, measured as the binding strength, is comparatively weak, with half-lives typically measured in seconds or microseconds.[36,37] Recently, Dash et al.[38] were able to identify predictive features that define TCR binding to specific, well characterized peptide-MHC complexes. Based on sequence similarities shared among the set

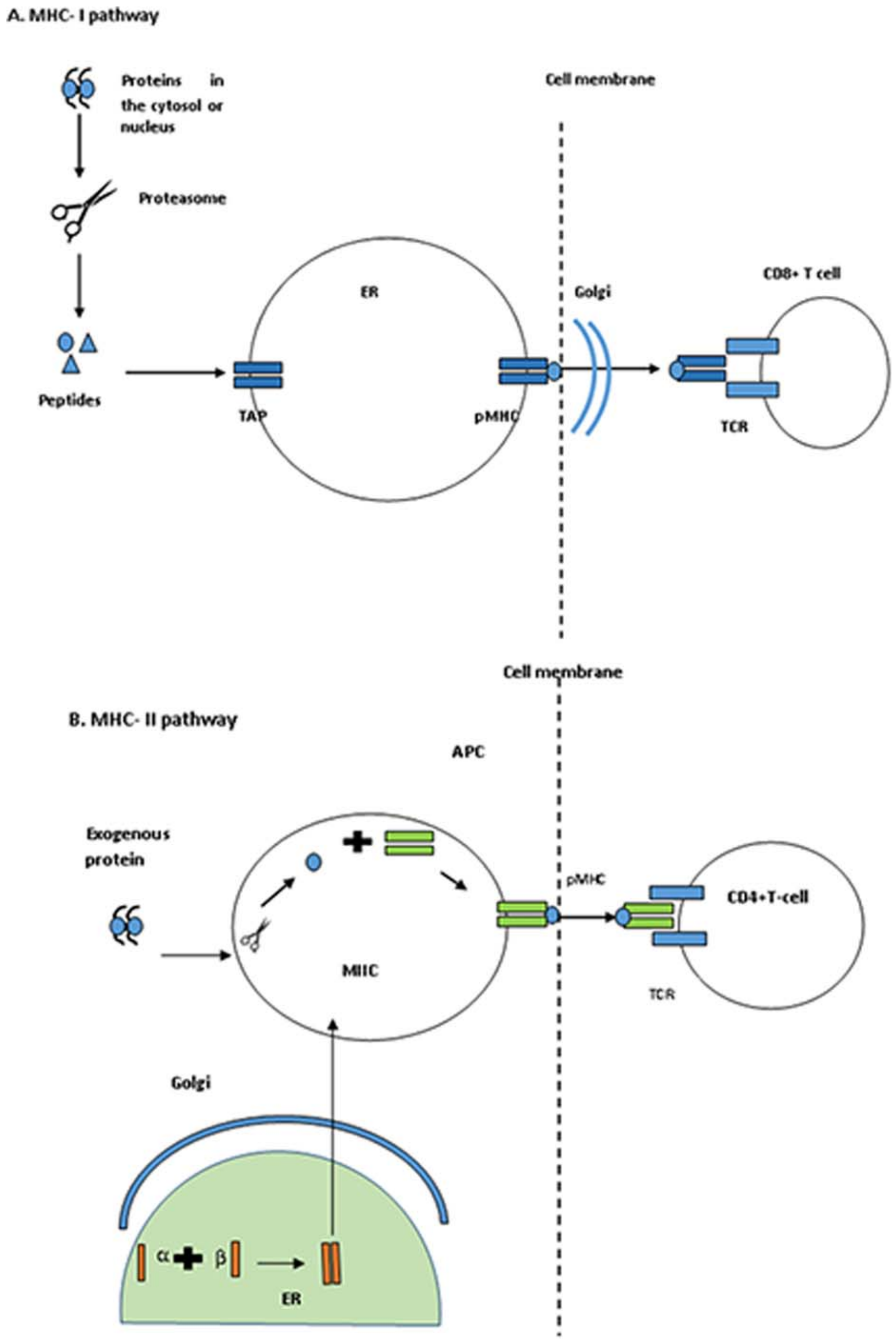

FIGURE 2 MHC-I and II antigen presentation pathways. **A**, In the MHC-I pathway endogenous antigens are cleaved by the proteasome and the resulting peptides are carried to the ER *via* TAP. The peptides then bind to MHC-I molecules and the resulting complex is transported to the surface of the cell to be recognized by the TCR on CD8+ T-cells. **B**, In the MHC-II pathway the α- and β-chains of MHC-II are joined together in the ER and form a complex with the invariant chain (Ii). This complex is then transported to the surface of the cell. Exogenous antigens and Ii are engulfed by the cell and digested into peptides by proteases in the MHC class II compartment (MIIC), which contains lysosome. MHC II molecules complexed with endogenous peptide fragments are then transported to the plasma membrane to be recognized by the TCR on CD4+ T-cells

of TCRs that recognize a specific peptide-MHC complex, a distance measure was developed that allows identification and classification of TCRs according to the presence of sequence motifs in one of the regions responsible for the complementarity to the peptide-MHC complex.

It is important to note that all antigenic peptides have to bind to MHC molecules in order to become T-cell epitopes but not all MHC-binding peptides are T-cell epitopes,[39,40] since not all of them will actually activate the immune system. A distinction thus arises between an antigenic and a immunogenic peptide as several signaling steps and cell intactions are mandatory for eliciting an immune response, the first being the availability of both an MHC molecule capable of presenting the peptide on the cell surface and, the second, a suitable TCR

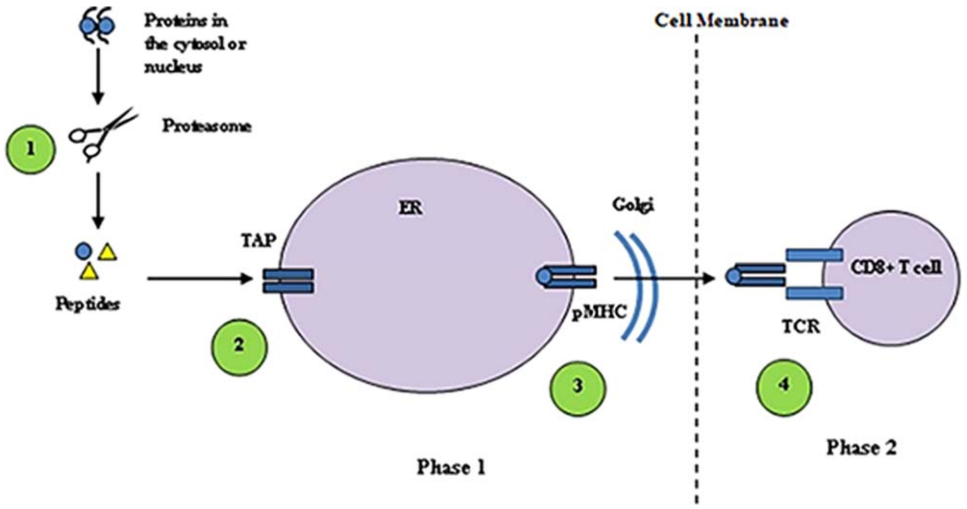

FIGURE 3 The immunogenic pathway associated with MHC-I molecules. **1**. Proteolysis (breakdown of proteins by proteasome). **2.** Transportation of peptides by TAP. **3.** Peptide-MHC binding. **4.** Recognition by the T-cell receptor on CD8+ T-cell (T-cell reactivity)

repertoire.[41] Therefore, immunogenicity can be understood as the propensity of a T-cell to recognize and bind its TCR to a specific peptide-MHC complex and trigger the immune response, whereas antigenicity refers merely to the molecular interaction of TCR with such peptide–MHC complex.

Accordingly, in silico methods for the prediction of immunogenic epitopes rely on the existence of specificity[42] and binding strength[20,43] of the interaction between the epitope and the binding groove of an MHC molecule[26,44] as a prerequisite for T-cell recognition. In silico modelling of peptide immunogenicity has confirmed the importance of positions 4–6 in the binding of an antigenic peptide to MHC-I, highlighting that residues with large and aromatic side chains are also associated with higher immunogenicity .[26,45] Binding affinity of the peptide–MHC complex is the result of the balance between its association and dissociation,[39] and is a measure widely used to identify potential T-cell epitopes, with 500 nM being proposed as the threshold binding affinity for peptide selection,[13,43] with medium binding affinities ranging from 50 to 500 nM, and values <50 nM being considered to reflect high binding affinity.[46] The immunogenicity of a peptide has also been linked to the kinetic stability of the peptide-MHC complex, such that peptides with higher stability (slow dissociation rate) reside longer in the MHC binding groove, maximizing their chances of being recognized by a TCR.[7,47,48] The concept of binding stability being a better predictor of immunogenicity than binding affinity has indeed been proposed by various research groups.[49–51] Furthermore, it has been proposed that combining predictions of stability, binding affinity and propensity of the peptide-MHC complex has the highest probability of accurate prediction of T-cell epitopes.[39]

A single TCR can recognize a peptide-MHC complex with similar structure but different sequence, demonstrating receptor–ligand cross-reactivity. The flexibility in the binding site of TCRs accounts for subtle changes in the topology of the interaction between pMHC and TCR, which has been proposed as the molecular basis of crossreactivity. Nonetheless, the relative orientation of the interaction seems to be conserved, ultimatly determining the effective recognition of the epitope.[26,52–54]

MHC-I molecules exhibit a preference for interacting with certain amino acids at specific positions in a peptide sequence, known as anchor residues. The anchor positions in MHC-I are highly conserved, defining space for only a few different amino acids.[55] In the first ever analysis of the structure of a MHC-I molecule, HLA-A*02:01, it was shown that the primary anchor amino acids (position 2 and leucine or valine in the C-terminus) are important for MHC-binding but still not sufficient, as other amino acids in the middle of the peptide (positions 1, 3, and 7), called secondary amino acids, play a significant role in peptide-binding.[56,57] The orientation (N to C) of the antigenic peptide is fixed for all MHC-I molecules due to the formation of hydrogen bonds at the N- and C-terminal ends of the peptide binding groove. Longer peptides may zigzag or bulge to maintain the position of specific amino acids in the sequence.[21,58,59] This fixed mechanism of binding is the key to determining MHC-peptide binding, which is an important step for discovering CTL epitopes. Many computational methods are indeed trained on the binding affinity data of primarily 9-mer peptides to accurately predict CTL epitopes.[60]

MHC-II molecules can interact with peptides up to 30 amino acids long because their peptide-binding groove is open at both ends (Figure 1), allowing the N- and C-terminal ends of a peptide to extend beyond the binding groove and enabling longer peptides to bind in the groove in many different ways in both ends, through the formation of hydrogen bonds between the side chains of the MHC-II molecule and the main chain of the peptide,[61] with the core binding region fixed to nine residues in most peptides.[58,59] The binding groove in MHC-II molecules contains three primary anchor amino acids, but due to the larger length of the peptide, individual anchor positions are difficult to identify.[57] The presence of such an undetermined core binding region with unidentified number of primary anchor amino acids means that it is more difficult to predict MHC-II than MHC-I epitopes.[42,62] For MHC-II peptides, secondary anchor residues have minimal impact on peptide binding affinity to MHC molecules.[57]

Figure 3 illustrates how the overall immunogenic pathway is split into two major phases. Phase 1 concerns all three steps of antigen presentation previously described. Phase 2 concerns the recognition of the peptide-MHC complex by the TCR, leading to T-cell activation.[63] Several methods have been developed to predict each individual step in Phase 1. The most important methods developed to predict proteasomal cleavage are PAProC,[64] NetChop,[65] MAPPP,[66] and Pcleavage[67] listed in Table 1. TAPPred.[89] is a method developed to predict TAP transport. All other methods listed in Table 1, have been developed to predict peptide-MHC binding. WAPP, NetCTL and SMM are methods which can integrate all three steps in Phase 1 of the immunogenic pathway (prediction of proteasomal cleavage, TAP transport, and MHC-I binding).[91,105] POPISK is a method developed to predict T-cell reactivity, which corresponds to Phase 2 of the immunogenic pathway.[63]

The development of accurate T-cell epitope prediction methods is of great interest to immunologists and the biotechnology industry as T-cell mediated responses constitute a key area for the design of vaccines against infectious diseases and for pre-clinical safety screening of novel potential therapeutic proteins. Nonetheless, the prediction of T-cell epitopes is still challenging due to the high degree of MHC polymorphism, the inconsistency in the data available required to understand and ultimately predict T-cell epitopes[42,106] and the characteristics of the antigen presentation process. In this review article, we will discuss some of the most important sequence- and structure-based methods developed to predict MHC–peptide binding and comment on their abilities and limitations.

2 | T-CELL EPITOPE PREDICTION METHODS

T-cell epitope prediction methods can be classified into two categories: sequence-based and structure based.[107] Very few techniques have been explored for structure-based approach due to the higher complexity involved and the longer computational times required compared to sequence-based approaches.[108]

2.1 | Sequence-based methods

These methods are based on the existence of patterns in the peptide sequences with known binding affinities to MHC molecules.[107] The response achieved and measured in experiments is used to train the algorithms in these methods;[109] however, for T-cell epitopes the large variety of pools of cells presenting a variety of receptors makes computational predictions difficult at best, unless specific TCRs have been selected.

2.1.1 | Anchor residue and sequence motifs-based method

MHC class I and II binding peptides contain residues that fit into polymorphic binding grooves, binding to complementary residues of specific MHC molecules. These anchor residues firmly bind at various positions in the MHC binding groove.[108,110–112] The combination of amino acids at the peptide anchor binding positions is called a sequence motif.[42] The MHC-binding motifs for a given peptide can be identified by comparison to common motifs present in known MHC-binding peptides and absent in nonbinding peptides.[113] The search for anchor residues and sequence motifs is the oldest yet the most extensively used method for the prediction of epitopes.[42,108] Various computational methods have been developed to search for motifs in the amino acid sequence of peptides using a motif library, with SYFPEITHI[68] being one of the most widely used approaches.[108]

Although the prediction of peptide binding motifs based on knowledge of anchor positions has been used to predict T-cell epitopes with MHC-I molecules in proteins from various pathogens,[114] the accuracy of this motif-based approach is not high enough as not all binding peptides have detectable motifs.[42] In addition, use of this approach is not feasible with peptides that bind to MHC-II molecules as the peptide binding sequence motifs are generally more degenerate than those binding to MHC-I molecules. Predicted motifs may contain an equal number of false positives, predicted peptide epitopes may not be actual epitopes, and there may be an equal number of false negatives (i.e., peptides not predicted to be epitopes but which are in fact epitopes).[13] Anchor residues play an integral part in making an effective vaccine by altering their positions.[115]

2.1.2 | Decision trees (DT)

DT constitutes a classification algorithm whereby a decision is reached after a probabilistic sequential test is carried out in every node. A DT is thus used to create a graph model of the MHC–peptide binding motif that later can be used to decide whether a test peptide fits into that motif.[58] Given its rule-based nature this method has better interpretability than other machine learning methods.[115,116] The Bonsai program, based on computational learning, was developed to investigate T-cell preference and predict peptide binding for human MHC-I allele, HLA-A*0201.[108,117] A similar approach was taken to predict peptides binding to H2-K^b, a MHC-I receptor in the murine family.[1][18] This method offers many advantages such as ease of use and its ability to handle large and complicated datasets effectively;[119] however, it does not maintain stability of the system if any input needs to be changed midway through the test, and it is difficult to prepare a DT.[120] DT have been successfully used to construct a classification of the models of leukemia based on protein profiling of bone marrow mononuclear cells[121] and to predict the diagnosis and outcome of dengue fever in its early stages.[122]

2.1.3 | Partial least squares (PLS)

PLS is a multivariate statistical approach[123] that can deal effectively with large data matrices[124] with noisy and highly collinear data.[125] In this approach the amino acid contribution at each position in a sequence motif and the contribution of the side chain interaction is added to calculate the binding affinity of peptides to MHC.[59,126] Crossvalidation (CV) using "leave-one-out" (LOO) is the technique used to validate the quality of the predictions made by the PLS method.[127,128] The PLS method can predict MHC–peptide binding by formulating new variables named principal components (PC), a linear combination of the original values of binding affinity (IC_{50}).[129] The PLS method has been tested and shown to work best among four different predictive methods available online (SYFPEITHI, which uses

TABLE 1 An overall summary of the various methods used for peptide binding and T-cell epitope prediction

Server	URL	Prediction	Method	Reference
SYFPEITHI	http://www.syfpeithi.de/bin/MHCServer.dll/EpitopePrediction.htm	MHC binding	Motif based	68
ProPred	http://www.imtech.res.in/raghava/propred/	MHC-II binding	Matrix	69
ProPred1	http://www.imtech.res.in/raghava/propred1/	MHC-I binding	Matrix	70
MAPPP	http://www.mpiib-berlin.mpg.de/MAPPP/binding.html	Proteasome cleavage	Matrix	66
SMM	http://zlab.bu.edu/SMM/	MHC-II binding	Matrix	55
SVMHC	http://abi.inf.uni-tuebingen.de/Services/SVMHC/	MHC-I and II binding	SVM	71
SVRMHC	http://c1.accurascience.com/SVRMHCdb/	MHC binding	SVM	72
CTLPred	http://www.imtech.res.in/raghava/ctlpred/	MHC-I binding	Consensus-SVM,ANN, QM	73
MHCPred 2.0	http://www.ddg-pharmfac.net/mhcpred/MHCPred/	MHC-I and II binding	PLS	74
NetMHC 4.0	http://www.cbs.dtu.dk/services/NetMHC/	MHC-I binding	ANN	75
EpiMatrix	http://www.epivax.com/epimatrix/	MHC binding	Matrix	76
EpiTOP	http://www.pharmfac.net/EpiTOP/	MHC-II binding	QSAR	77
FDR4	http://crdd.osdd.net/raghava/fdr4/	MHC-II binding	SVM	–
HLADR4Pred	http://www.imtech.res.in/raghava/hladr4pred/	MHC-II binding	Consensus-SVM,ANN	78
HLAPred	http://crdd.osdd.net/raghava/hlapred/	MHC binding	Matrix	–
nHLAPred	http://www.imtech.res.in/raghava/nhlapred/	MHC-I binding	Consensus-ANN,QM	79
ComPred	http://www.imtech.res.in/raghava/nhlapred/comp.html	MHC-I binding	Consensus-ANN,QM	80
AnnPred	http://www.imtech.res.in/raghava/nhlapred/neural.html	MHC-I binding	ANN	80
NetCTL	http://www.cbs.dtu.dk/services/NetCTL/	MHC-I binding	ANN, Matrix	81
NetMHCstab	http://www.cbs.dtu.dk/services/NetMHCstab/	MHC-I peptide binding	ANN	34
NetCTLpan	http://www.cbs.dtu.dk/services/NetCTLpan/	MHC-I binding pan-specific	Consensus-ANN, Matrix	82
NetMHCpan-3.0	www.cbs.dtu.dk/services/NetMHCpan-3.0	MHC-I binding pan-specific	ANN	83
MHC2Pred	http://crdd.osdd.net/raghava/mhc2pred/	MHC-II binding	SVM	80
MOT	http://www.imtech.res.in/raghava/mhc/page4.html	MHC-II binding	Matrix	–
NetMHCIIpan	http://www.cbs.dtu.dk/services/NetMHCIIpan/	MHC-II binding pan-specific	ANN	54
NetMHCIIpan 3.1	http://www.cbs.dtu.dk/services/NetMHCIIpan-3.1/	MHC-II binding pan-specific	ANN	84
NetTepi	http://www.cbs.dtu.dk/services/NetTepi/	MHC binding	HMM	39
NNAlign	http://www.cbs.dtu.dk/services/NNAlign/	MHC-II binding	ANN	85
TEPITOPEpan	http://datamining-iip.fudan.edu.cn/service/TEPITOPEpan/index.html	MHC-II binding pan-specific	PSSM	86
KISS	http://cbio.ensmp.fr/kiss/	MHC-I binding	SVM	87
PickPocket	http://www.cbs.dtu.dk/services/PickPocket/	MHC binding	Position specific weight matrix	88
NetMHCcons	http://www.cbs.dtu.dk/services/NetMHCcons/	MHC-I binding	Consensus-ANN, Matrix	88
PAProC	http://www.paproc.de/	Proteasome cleavage	–	64
NetChop	http://www.cbs.dtu.dk/services/NetChop/	Proteasome cleavage	ANN	65
Pcleavage	http://www.imtech.res.in/raghava/pcleavage/	Proteasome cleavage	SVM	67
TAPPred	http://www.imtech.res.in/raghava/tappred/	TAP Transport	SVM	89
PREDIVAC	http://predivac.biosci.uq.edu.au/	MHC-II binding	Specificity determining resi-	90

(Continues)

TABLE 1 (Continued)

Server	URL	Prediction	Method	Reference
			due (SDR)	
POPISK	http://iclab.life.nctu.edu.tw/POPISK/	T-cell reactivity	SVM	63
WAPP	http://abi.inf.uni-tuebingen.de/Services/WAPP/index_html	Proteasomal cleavage + TAP + MHC-I binding	Matrix + SVM	91
AutoDOCK	http://autodock.scripps.edu/	MHC binding	Molecular docking	92
GOLD	http://www.ccdc.cam.ac.uk/solutions/csd-discovery/components/gold/	MHC binding	Molecular docking	93
GLIDE	http://www.schrodinger.com/Glide/	MHC binding	Molecular docking	94
AMBER	http://ambermd.org/	MHC binding	MD simulation	95
CHARMM	http://charmm.chemistry.harvard.edu/	MHC binding	MD simulation	96,97
NAMD	http://www.ks.uiuc.edu/Research/namd/	MHC binding	MD simulation	98
GROMACS	http://www.gromacs.org/	MHC binding	MD simulation	99
MODELLER	http://salilab.org/modeller/	MHC binding	Homology modelling	100
SWISS-MODEL	http://swissmodel.expasy.org/	MHC binding	Homology modelling	101
Phyre2	http://www.sbg.bio.ic.ac.uk/~phyre2/html/page.cgi?id=index	MHC binding	Homology modelling	102
GenTHREADER	http://www.chemogenomix.com/chemogenomix/GenThreader.html	MHC binding	Threading	103
MUSTER	http://zhanglab.ccmb.med.umich.edu/MUSTER/	MHC binding	Threading	104

binding motifs, MHC-Thread, which is a threading-based method, RANKPEP, which uses position-specific scoring matrices, and ProPred, which uses quantitative matrices) to predict T-cell epitopes.[127] TEpredict[130] and MHCPred2.0[74] are sequence-based methods that use the PLS method to predict binding of peptides to MHC class I and II molecules. TEpredict has been compared to SVRMHC,[72] SVMHC,[71] ProPred1,[70] and SYFPEITHI[68] for T-cell epitope prediction, carried out for allelic variants of class I MHC molecules (A*0101, A*0201, A*0301, and B*0702). Peptide samples were taken from the MHCBN[131,132] and IEDB[133] databases and it was found that TEpredict is as good if not superior to other methods.[130] MHCPred is a method widely utilized to predict T-cell epitopes that uses peptide-MHC affinity values, IC_{50}, as input, converted to pIC_{50} values and then used as the dependent variable in a quantitative structure-activity relationship (QSAR) regression.[123]

2.1.4 | Quantitative matrices (QM)

In QM the contribution of amino acids at each position within the peptide is quantified and a matrix is created using an appropriate formula to calculate a peptide binding score.[73,134] This method is simple and efficient although it has limitations in generating a binding coefficient matrix for each MHC molecule as it requires rigorous testing of several hundreds of peptides, and it assumes that every amino acid residue influences binding individually and ignores the synergistic binding contribution from the entire peptide structure.[135]

EpiMatrix is a method that uses QM to predict MHC binding and rank each 8–10-mer overlapping frame in a protein sequence.[136] An EpiMatrix score is given to a peptide according to its predicted binding affinity for each amino acid, with a higher score indicating greater possibility for MHC binding.[76] ProPred is a graphical online tool which implements QM to predict promiscuous binders, i.e. peptides that bind to several MHC-II alleles for the development of a subunit vaccine, containing only the antigenic parts of the pathogen.[69,137] By contrast, EpiMatrix predicts T-cell epitopes by identifying MHC peptides only.[76] EpiJen is a QM-based server for multi-step T-cell epitope prediction that utilizes successive filters to eliminate negatives at each step of Phase 1 of the immunogenic pathway.[91] QMs are trained first on available data before being validated. As a result, QM can approximate the data by memorizing the training data but do not work well with unseen data, referred to as over-fitting, which limits the application of QM.[138]

Position-specific scoring matrices (PSSM) constitute another category of QM-based method that is simple to use.[11] For each amino acid position in a peptide, PSSM-based methods generate a matrix where each row corresponds to a distinct amino acid and each column corresponds to the position in the peptide sequence, with the matrix being defined as the number of matches of each amino acid at each position of the sequence and the corresponding values being added to derive a score to predict the binding affinity of an unknown peptide.[115,139] PSSM derived from peptide binding to specific MHC molecules capture the complexity of the binding motifs as the peptide and the MHC molecule(s) to which it binds are functionally related.[140] The length variability due to bulges and loops in a peptide shifts the position of amino acids, which can cause the PSSM method to fail.[58] This

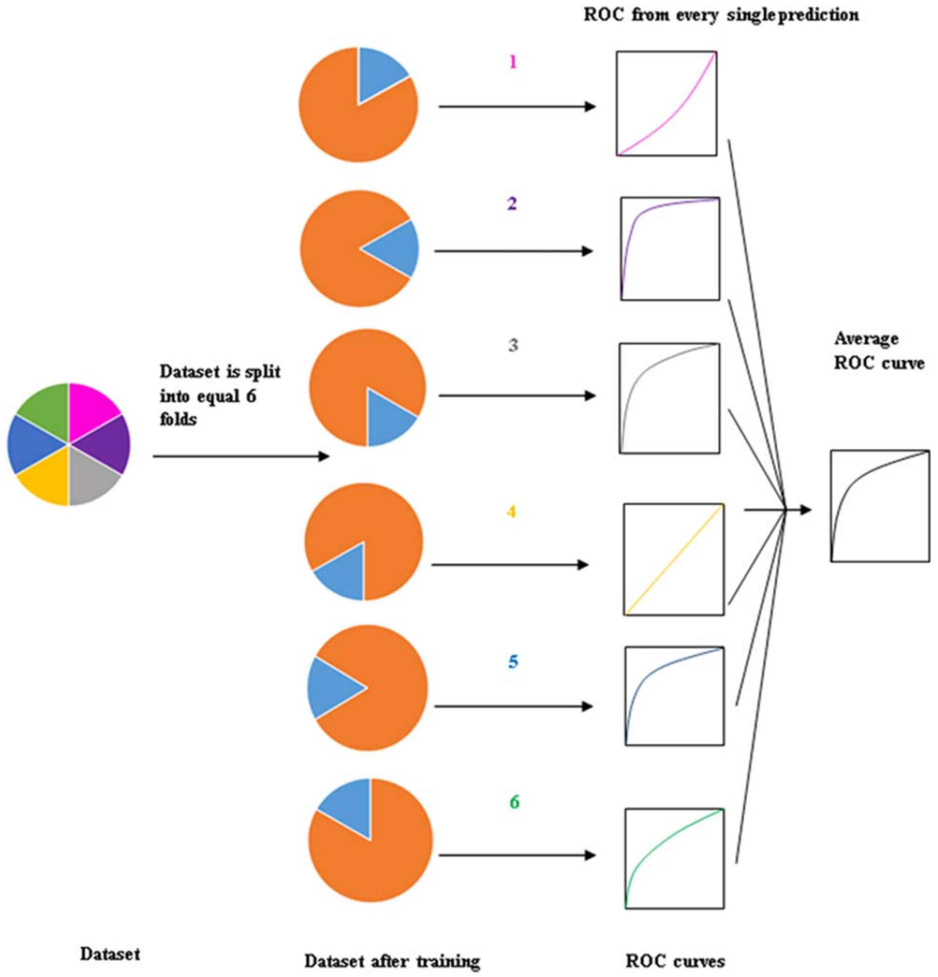
ROC from every single prediction
1
2
3
4
5
6
Dataset is split into equal 6 folds
Average ROC curve
Dataset
Dataset after training
ROC curves

FIGURE 4 Creation of receiver operating characteristic (ROC) curves from the dataset. Data is first split into equal folds (six folds in this case) and then six algorithms work on five folds and are validated on the "leave-one-out" crossvalidation method. 1. Dataset is trained with five folds and validated on LOOC 1st section, 2. Dataset is trained with five folds and validated on LOOC 2nd section, 3. Dataset trained with five folds and validated on LOOC 3rd section, 4. Dataset trained with five folds and validated on LOOC 4th section, 5. Dataset trained with five folds and validated on LOOC 5th section, 6. Dataset trained with 5 folds and validated on LOOC 6th section

method is unable to model the inherent nonlinearity in peptide binding affinity and the inter-relationship between amino acid positions.[11]

A matrix-based method, stabilized matrix-based method (SMM), had been used to predict T-cell epitopes.[55,91,141,142] It utilizes two novel approaches: first, experimental information of clear non-binders is integrated into a distance equation and, secondly, use of a regularization technique is used to try to solve the over-fitting problem.[141] The weight-matrix method is another matrix-based method where the score of a peptide is calculated by adding the score of each residue,[143] and servers such as IEDB[55,142] and ProPred1[70] use this method to predict MHC binding.

2.1.5 | Machine learning (ML)

ML is best suited to situations in computational biology, where a large amount of peptide data is available with many missing or wrong facts also with a risk of having a high degree of uncertainty in selecting the appropriate models with numerous parameters.[144] This technique is further classified into supervised and unsupervised learning. In supervised learning, predefined datasets are available for training the input data, whereas in unsupervised learning predefined data for the training set are not available.[11,116] ML is used where the interactions within data are complex and thus more pattern recognition is derived from the dataset.[11]

Figure 4 depicts of the use of ML where an average receiver operating characteristic (ROC) curve is generated from pattern recognition and validated by the leave-one-out crossvalidation (LOOC) method. ML methods such as ANNs, HMMs, and SVMs have been applied for the prediction of peptide-MHC binding with peptide sequences as input and binding/nonbinding as output.[14,62,78,145]

2.1.6 | Artificial neural networks (ANN)

Inspired by the connections between neurons, ANN is a method based on computational elements (nodes or units) whose connection carries numeric data used for processing nonlinear relationships[14,146] and pattern recognition.[147] Figure 5 shows an ANN that was used to predict the binding score for 9-mer peptides, consisting of an input and an output layer along with hidden layers. In an ANN, the weight of the neuron plays a crucial role, affecting the computation of the neuron. For a

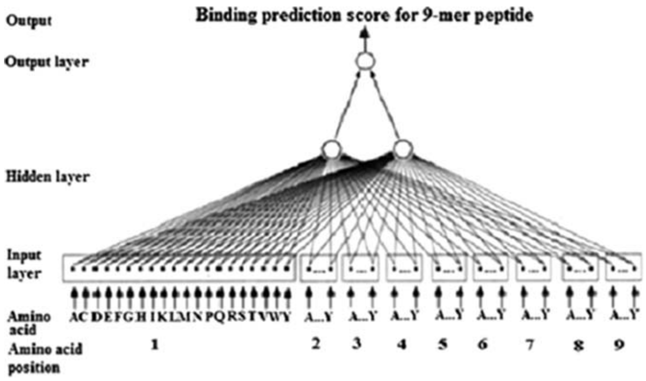

FIGURE 5 Example of an artificial neural network trained to predict the binding score for 9 residues in the core sequence of a peptide (reproduced with permission[14]) Published with permission from Nature Publishing Group

large and complex ANN with numerous neurons, the weights of the neurons can be adjusted with the help of a training algorithm.[148] At the end of every task the output can be considered correct or incorrect. Weights for correct outputs are secured, whereas weights for incorrect outputs are not accepted, so the network experiences "supervised learning".[149] Backpropagation is a common method of training the ANN data by modifying the weights to minimize the error between predicted binding probabilities and target probabilities.[150,151] ANN can process nonlinear, noisy data but the peptide data needs to be aligned as it can contain peptides of variable lengths, with the alignment of peptides before feeding it to ANN being referred to as preprocessing.[138]

In ANN, amino acids are connected via nodes described using classic orthogonal sparse encoding or Blosum series encoding.[75] Blosum50 encoding provides better results over sparse encoding as it is superior when limited data is available and also allows parameters to adjust simultaneously for each sequence.[14] However, use of this method with a small training dataset suffers from over optimization.[152] ANN requires fixed length input data, limiting its use to predictions of the binding of peptides of the same length as those in the training data set.[108]

ANN has been used to predict binding of peptides to MHC-I and MHC-II molecules although it is difficult to predict MHC-II binding of peptides due to their variable length, the undetermined core region and the number of amino acids accepted as primary anchors.[14,62] It has been reported that ANN-based methods accurately classify 80% of binders and almost 20% of nonbinders owing to 80% sensitivity and 80% specificity.[14,62]

NetMHC 4.0,[75] ANNPred,[80] NNAlign,[85] and NetChop[65] are all methods that use ANN models, whereas other methods use consensus models that combine ANN with one or two ML algorithms, such as NetCTL,[81] HLADR4Pred,[78] ComPred,[80] nHLAPred,[79] CTLPred,[73] NetCTLpan,[82] as summarized in Table 1. The consensus model, NetCTL, combines QM and ANN to predict T-cell epitopes utilizing ANN-based proteasome cleavage prediction, QM-based TAP binding and ANN-based MHC class I binding prediction.[81] NetMHCstab is a method that predicts the stability of peptide binding to different MHC molecules using ANN. It has been shown that better predictions can be achieved by combining binding affinity and stability rather than using affinity alone.[34]

2.1.7 | Hidden Markov models (HMM)

HMM constitute a statistical model suitable for demonstrating time-series sequences (strings) with variable lengths.[145] They have been vigorously applied in the analysis of biological sequences.[153–155] HMM can have either supervised or unsupervised learning depending on the training data provided.[116,156]

HMM comprises two stochastic processes: an invisible process of hidden states and a visible process of observable symbols.[155] A fully connected, supervised HMM has been used to predict MHC-peptide binding, specifically to HLA-A*0201. To overcome the limitation of the assignment of weights in the network parameters of ANNs, HMMs have been applied to predict peptide binding to MHC.[145] The advantage of using HMM is that it is possible to investigate peptides of several lengths but separately for binders and nonbinders.[115] In addition, this approach does not need pre-processing of the peptides before training.[138] NetTepi[39] is a computational system developed for T-cell epitope prediction by integrating peptide-MHC-I binding affinity, binding stability and T-cell propensity.

2.1.8 | Support vector machines (SVM)

SVM constitute a supervised method based on statistical theory[157,158] that is used to provide a framework for more sophisticated models that take into account the interactions between peptides and MHC molecules for a better understanding of MHC/peptide binding and to accelerate the process of finding T-cell epitopes.[159] Prediction of peptides that bind to MHC-I using SVM has shown high performance and applicability to a large number of MHC-I types.[71] SVRMHC,[72,157] SVMHC,[71] POPISK,[63] KISS,[87] FDR4, MHC2Pred,[80] and Pcleavage[67] are methods that have been used with the prediction of T-cell epitopes. WAPP is a multistep algorithm to predict T-cell epitopes which applies proteasomal cleavage matrices, transportation of peptides by SVM-based TAP and peptide binding to MHC using a series of successive filters.[91] SVM requires fixed length input data, which poses problems for the prediction of binding to MHC-II, but this limitation has been overcome by using a kernel method derived from SVM that can handle peptides with variable length.[160] A support-vector based Takagi-Sugeno-Kang fuzzy model (TSK-SVR) is a relatively new approach that has been proposed and applied to the quantitative prediction of MHC-peptide binding affinities.[161]

2.1.9 | Deep learning

Deep learning or deep neural network (DNN) is the latest addition to the ML range of methods and is based on the use of multiple hidden layers of nonlinear transformations and sets of complicated propositional logics.[162] Each layer takes an input and produces an output, and that output becomes the input for the next layer.[163] Deep learning methods have shown exceptional performance in the recognition, classification and feature extraction from noisy data, although they have some limitations, requiring large data sets, a selection of the task-appropriate DNN and high computational costs.[164] This method was recently used to predict peptide binding to MHC and T-cell epitopes by directly capturing the interactions between primary (anchor) and secondary amino acid residues involved in the formation of peptide–MHC complexes.[150]

2.2 | Protein structure-based methods

The method is based on knowledge of epitope interactions with MHC molecules on the basis of structural data, and is used to predict the structures of the bound complexes.[107]

2.2.1 | Quantitative structure–activity relationships (QSAR)

QSAR, a commonly used ligand-based drug design approach,[129,165] is a method for building mathematical models of biological activity on the basis of the chemical and structural properties of molecules, and is based on the assumption that biological activity of a molecule depends on its structure.[165–168] All QSAR approaches are thus based on the assumption that molecules with similar structures are likely to have similar biological activity and will bind to the same protein target site.[169,170] By optimizing each residue-to-residue interacting pair, QSAR can improve the prediction of peptide interaction with the MHC-I binding groove.[171]

Regardless of the efficacy of classic QSAR approaches, they cannot be used to analyze all data sets due to the lack of availability of physicochemical parameters for all molecules. In 3D-QSAR, quantitative models of small molecules are developed on the basis of structural and chemical 3D properties.[172] The 3D-QSAR methods are indeed efficient at detecting important features in peptide-MHC interactions.[172] CoMFA and CoMSIA, two 3D-QSAR methods, have been used to predict MHC-peptide interactions.[173]

QSAR is built by implementing PLS,[74,123] a statistical method used in various versions of SYBYL (Tripos Inc.).[127,128,165,167,174] EpiTOP is one of the servers that utilizes QSAR approaches for predicting peptide binding to MHC-II based on proteochemometrics,[175] where ligand binding to several related proteins is considered.[77]

2.2.2 | Protein structure prediction and molecular modelling

Knowledge in atomistic detail of the molecular structure of MHC molecules and their interactions with peptides can be used to build a 3D model of their complexes with other peptides.[138] If the structure of interest is not available or cannot be identified, models can be constructed from scratch in what is called ab initio protein modelling.[176] A range of molecular modelling methods have been developed to investigate the structure, dynamics and thermodynamic properties of proteins, helping to explain how molecular structure and flexibility can be related to the function of biological systems.[177]

Threading, often called fold recognition, involves use of a known peptide–MHC complex structure to predict the bound structures of different peptides to the same MHC molecule by optimizing the alignment between the amino acid sequence and their 3D structural patterns.[108,165,178,179] This method involves using a backbone conformation method whereby the 3D backbone coordinates of the source, known peptide (S_1, S_2, . . ., $S_{n)}$ are substituted with the amino acid sequence of the target peptide (P_1, P_2, ..., $P_{n)}$ by direct replacement of P_i with S_i.[108,180] It has been suggested that threading be used to scan a library of either protein folds or sequences to predict MHC–binding peptides.[181] The known peptide structure used as a template for threading peptides and their binding potential can be analyzed with respect to a wide range of MHC-I molecules.[182] Often accurate prediction of peptide structures is however stalled due to the unavailability of suitable peptide backbone templates.[42] GenTHREADER.[103] and MUSTER[104] are widely used threading algorithms to generate 3D models of proteins.

Homology modelling, also called comparative modelling, is a structure-based method that relies on the relationship between the target protein sequence and at least one known structure of a similar protein.[165] It is the most precise method for predicting a protein structure from its sequence of amino acids,[183] although it can only be applied when the 3D structure of a similar protein is already known.[178] This method consists of a number of steps: alignment of the appropriate 3D template, addition of the polypeptide backbone, modelling of side chains, and mapping of loops.[183–185] Proteins can sometimes be too large or are not amenable to be analyzed by nuclear magnetic resonance (NMR) or X-ray diffraction. In those cases homology modelling is the only way to obtain a structural model of the proteins.[185,186] Peptide residues are selected based on homology to the most similar known MHC-peptide structure. This method has been shown to be effective at generating the in silico 3D structure of a protein to predict conformational epitopes.[187] This requires at least 30% sequence homology to generate a reliable model.[108,183,188,189] MODELLER,[100] SWISS-MODEL,[101] Phyre2[102] are some of the most widely used tools to produce 3D homology models of proteins.

2.2.3 | Molecular docking

Molecular docking is a computational method aimed at predicting the structure and interactions of a ligand-receptor complex. Docking is usually done in two steps: first, by placing the ligand in the binding site of a protein and, second, by estimating the binding affinity with a scoring function.[190–193] Peptides with a higher docking score and a suitable binding conformation can be predicted to be a potential T-cell epitopes for each haplotype, a particular combination of alleles.[194] Fresno, a tailor-made energy scoring function, was developed to predict peptide binding to MHC-I molecules.[93]

The flexibility of the protein molecule and the role of water molecules play critical roles in protein–ligand interactions and the prediction of free energies of binding.[191,192,195] Water molecules may be included as additional participants in docking, particularly if they appear to be an integral part of the protein surface if they are strongly bound and/or evolutionary conserved.[196] WaterMap is a method based on inhomogeneous solvation theory that can efficiently predict thermodynamic properties of water molecules, allowing a better characterization of molecular binding in a water-mediated system, and is also a powerful tool when applied to binding steered by hydrophobic interactions.[197,198] For many years docking simulations were performed with a combination of flexible ligand and rigid receptor, but there are a

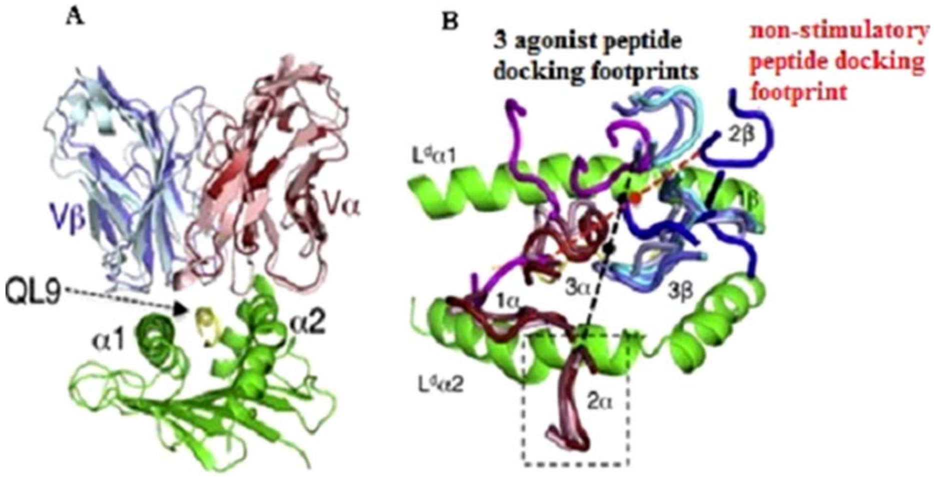
A
Vβ
Vα
QL9
α1
α2
B
3 agonist peptide docking footprints
non-stimulatory peptide docking footprint
2β
Ldα1
1β
1α
3α
3β
Ldα2
2α

FIGURE 6 **A**, Ribbon representations of TCRs, 42F3 and 2C in complex with QL9 (peptide)-H2-L^d aligned on the MHC. **B**, An example of TCR-peptide-MHC docking: Six CDR loops of TCRs, 2C (1α, 2α, 3α, 1β, 2β and 3β) and 42F3, with QL9 peptide on the surface of MHC-I molecule, H2-L^d (reproduced with permission[206]). Published with permission from Elsevier

growing number of methods that can deal with the flexibility of the receptor.[190–193] There have been many developments in docking such as binding with more than one ligand, use of machine learning approaches and nonlinear scoring functions.[195]

AutoDOCK,[92] GLIDE,[94] GOLD,[93,199] and EpiDOCK[200] are just a few structure-based docking programs that have been used for MHC peptide-binding prediction. Use of docking simulations is advantageous because it eliminates the possibility of over-fitting binding data by allowing the development of a predictive model with separate training and test data sets.[201] AutoDOCK was used to predict the binding of 15-mer epitopes to the binding groove of MHC-II molecules for the development of a vaccine against the Nipah virus.[202] EpiDock was reported to recover 90% of known high affinity binders when applied to the prediction of HLA-A*0201-restricted T-cell epitopes from the hepatitis B virus.[203] Computational docking of the MHC-II allele to T-cell epitopes of bacterial meningitis was successfully carried out using GLIDE to identify a common subunit vaccine.[204] TCRFlexDock, a modified docking program based on RosettaDock and ZRANK, successfully predicted nearly 80% of TCR/pMHC binding of a benchmark of 20 test cases with antigenic peptides associated with viruses, cancer, and autoimmunity.[205] Figure 6 illustrates the example of 2C/42F3-QL9-H2-L^d, a docked complex of TCR and the peptide-MHC-I complex, where QL9 is the peptide, 42F3 and 2C are TCRs and H2-L^d is the MHC-I molecule.[206] Molecular docking is an efficient and cost-effective computational method,[192] although simulating the interactions with large ligands remains challenging.[191]

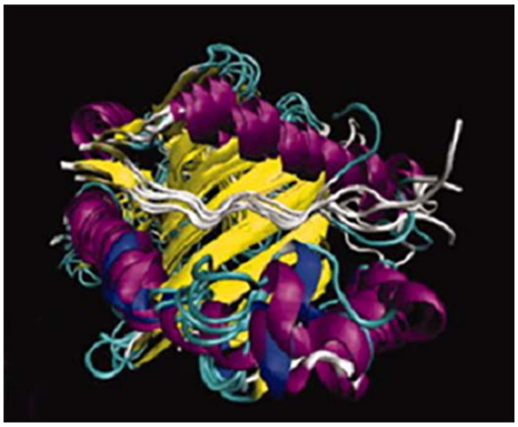

FIGURE 7 Stability of the HLA-DP2 peptide complex. MD simulations conducted over 50 ns showed that the structure of this peptide-MHC-II complex derived from X-ray diffraction studies (PDB code 3lqz) is stable in its binding core (reproduced with permission[15]). Published with permission from John Wiley & Sons

2.2.4 | Molecular dynamics (MD) simulations

MD simulation is a method for studying the individual and collective movement of atoms and molecules inside a realistic molecular system where all intramolecular and intermolecular interactions are represented through the use of force fields.[207,208] This method solves numerically the time-dependent behavior of a molecular system on a microscopic scale[209] and provides detailed information on the structure and conformational changes in proteins and their associated thermodynamic properties.[210] The use of MD simulations does not rely on the availability of binding data (IC50) to make predictions, as it requires only the experimentally determined structure of the proteins involved.[13] MD simulations have enabled the search of epitopes by the direct assessment of the free energy of binding between two peptides.[42] In one of the earliest applications to T-cell epitope prediction, MD simulations were used successfully to simulate the MHC-I protein, HLA-B*2705, characterizing the various structural and dynamic properties of the peptide–MHC complex along with role of secondary anchors in stabilizing MHC-peptide binding.[211,212] MD simulations have also been used to predict the structure of peptide-MHC complexes using HLA-A*02:01-binding peptide epitopes from hepatitis B virus (HBV) proteins.[213] MD simulations have been used for predictive virtual screening of HLA-DP2-peptide binding, revealing that the peptide binding core is stable with time-dependent conformational variations in the flanking residues at the C-terminus, as shown in Figure 7.[15]

MD simulation is the most computationally expensive method, leading to its two major disadvantages. First, short MD simulations often result in inadequate sampling of conformational space with inaccurate predictions of binding affinity and interaction, although the use of supercomputers is rapidly overcoming this limitation. Second, inaccuracies in force fields in representing protein structure and interactions may lead to spurious results.[207] In spite of these potential limitations, MD simulations carried out over sufficiently long times and with suitably validated force fields and methods offer invaluable information about atomistic-level interactions. AMBER,[95] GROMACS,[99] CHARMM,[96,97] and NAMD[98] are some of the most widely used force fields and MD biomolecular simulation programs used for the calculation of binding free energies.

A computational model of the TCR-pMHC-CD4 complex including water and lipid molecules was constructed to study the so-called immune synapse, where the interaction between TCR and peptide-MHC complex occurs (step 4 in Phase 2 of the immunogenic pathway, as shown in Figure 3). Thermodynamic binding properties were estimated by the MM/PBSA, linear interaction energy (LIE) and thermodynamic integration (TI) methods and were found to be comparable to experimental data.[214] MD simulations have also been performed on a TCR-peptide-MHC complex to compare the impact of the N-terminal peptide flanking region (PFR) on peptide binding using two peptides, a 12-mer and an 18-mer respectively, revealing that the 18-mer peptide has a greater peptide binding affinity.[212]

2.3 | Consensus and pan-specific methods

Consensus methods involve the qualitative combination of generally two to three methods to accurately predict a potential vaccine candidate using large numbers of MHC binders and nonbinders. Two such consensus methods, nHLAPred and ComPred, have been developed by integrating QM and ANN methods to improve the accuracy of the prediction of MHC-binding peptides or CTL epitopes of 67 MHC alleles. In this approach ANN has been used to predict MHC binders and nonbinders while QM was used to re-examine these by comparing the cutoff score discriminating between them. If ANN predicted a binder having a low score for QM, then QM assigns it as a nonbinder. Similarly, if ANN predicted a nonbinder with very high score for QM, then QM assigns that nonbinder as binder in the final prediction.[79,80] PERUN is another method developed using an evolutionary algorithm (EA) and ANN for predicting MHC-II binding peptides associated with insulin-dependent diabetes and rheumatoid arthritis.[62] CTLPred is a direct method that applies QM, SVM and ANN to accurately predict CTL epitopes by discriminating T-cell epitopes and nonepitope MHC binders, which indirect MHC binder prediction methods fail to do. This method achieved a sensitivity of 79.4%, which is better than the individual predictions attained with each method, while the consensus approach achieved 88.4%, better than any other method.[73] NetMHCcons is another consensus method that combines ANN-based NetMHC, NetMHCpan, and PSSM-based PickPocket.[88] As a combination of top-performing methods, consensus methods can outperform any single method; however, the method requires outputs from the individual methods simultaneously.[115] PREDIVAC is another pan-specific approach developed to predict T-cell epitopes for vaccine design based on the specificity determining residue (SDR) method. SDRs are a small group of structurally conserved positions in the peptide-binding interaction interface that are responsible for the recognition of the peptide. Predivac predicts MHC-peptide binding by establishing a link between the SDRs in the MHC target protein and MHC proteins of known specificity.[90]

A relatively new method named pan-specific was able to manage with the degree of MHC-II polymorphism[90] and was trained on data covering multiple MHC molecules, and was developed to predict peptide binding with very restricted or no experimental data.[83] This algorithm-based method takes both the peptide sequence and the MHC contact environment into account, going beyond the conventional, sequence-based approach.[82,215] NetMHCIIpan was first developed to predict peptide binding to any HLA-DR, a MHC-II cell surface receptor encoded by HLA.[54] NetCTLpan was later developed for accurate prediction of MHC-I binding.[82] TEPITOPEpan has been evaluated as the second best method after NETMHCIIpan-2.0 (among four pan-specific methods: NetMHCIIpan-2.0, NetMHCIIpan-1.0, MultiRTA and TEPITOPEpan) for predicting binding specificities of unknown HLA-DR molecules.[86] NetMHCIIpan-3.1 is the newest developed pan-specific method which outperforms its earlier version NetMHCIIpan-3. This method can predict the binding affinity while achieving comparable accuracy in identifying the peptide binding core.[84]

3 | DISCUSSION AND CONCLUSIONS

Comparison of the various methods developed for the prediction of T-cell epitopes reveals that QM have the ability to work with both small and large data sets of binding peptides but cannot handle nonlinearity within the data. This approach is not adaptive, requiring a complete redesign to incorporate new data in the matrix. On the other hand, ANN requires a large amount of pre-processed data, can deal with non-linear data and is adaptive and self-learning.[62,138,216]

SVM shows better performance with data classification for MHC binders and nonbinders, as well as in identifying CTL epitopes and nonepitopes compared to ANN and QM.[73,78] When evaluated through Leave One Out Cross-Validation (LOOCV), QM, ANN, and SVM, component methods of CTLPred achieved an accuracy of 70.0, 72.2, and 75.5%, respectively when it was trained and tested on dataset of T cell epitopes and nonepitopes, clearly demonstrates that ML methods (SVM, ANN) have been noted to work better than QM on a blind data set,[73] although QM-based methods have more biological significance as they provide binding information about amino acids at all positions in spite of performing poorly than any other methods.[217] Comparison of QSAR (a structure-based method) and ANN (a sequence-based method) suggests that both methods achieve the same goal of predicting the MHC binding either as 3D-structure or sequence, but QSAR approaches implement a more detailed statistical analysis in the form of multiple linear and continuum regression, discriminant analysis, and PLS.[129] The main drawback of methods for predicting MHC binders instead of T-cell epitopes is that these methods can only predict MHC binding from antigenic peptide sequences but cannot discriminate between T-cell epitopes and nonepitope MHC binders.[73]

A general strategy can be suggested for predicting T-cell epitopes based on using different methods depending on the size of the data available.[1,124,134,138] (1) when no peptide data is available (but the sequence of the relevant MHC molecules is known), structure-based methods such as molecular modelling should be used;[124,176] (2) when only a limited amount of peptide binding data is known (for $<$ 50 peptides), binding motif, PLS and PSSM methods are the best possible method;[115,124] (3) when an intermediate amount of data is known ($\sim$50–100 peptides), QM or SVM is the best choice; (4) when data is available for more than 100 peptides, HMM, ANN or PLS-based methods should be used.[124,138] and (5) when an adequately large data set is available, ANN will be the best approach for peptide binding prediction.[218]

When comparing structure and sequence-based methods, it can be concluded that structure-based methods have the advantage of not being dependent on the availability of experimental binding data, but they are slow methods for characterizing a large number of peptides and are limited by the availability of structural data about the influence of protein conformational changes, pH, ionic strength and other biophysical factors. On the other hand, sequence-based methods are fast but can only achieve high accuracy when extensive peptide-affinity data is available for specific alleles.[219] Nonetheless, the inherent ability of these methods and their specific descriptors remains critical for improving the success of their predictions.

An ideal prediction method would integrate the strengths of individual methods while minimizing their disadvantages. Combining sequence and structure-based methods would be the best consensus method, but this is hindered by the lack of 3D structures.[8] The DNN method has only been explored for the prediction of peptide binding and thus needs to be explored further to determine its full potential.

ORCID

Prattusha Kar http://orcid.org/0000-0001-8490-3749

REFERENCES

[1] C. A. Janeway, Jr, R. Medzhitov, *Annu. Rev. Immunol.* ***2002***, 20, *197*.

[2] D. W. Taylor, D. W. Corne, *An Investigation of the Negative Selection Algorithm for Fault Detection in Refrigeration Systems*, Artificial Immune Systems, Springer **2003**, p. 34.

[3] C. A. Janeway, Jr, P. Travers, M. Walport, *Immunobiology: The Immune System in Health and Disease*, 5th ed., Garland Science, New York **2001**.

[4] J. Ponomarenko, H.-H. Bui, W. Li, N. Fusseder, P. E. Bourne, A. Sette, B. Peters, *BMC Bioinform.* ***2008***, 9, *514*.

[5] J. M. den Haan, R. Arens, M. C. van Zelm, *Immunol. Lett.* ***2014***, 162, *103*.

[6] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, P. Walter, *Molecular Biology of the Cell*, Garland Science, New York **2002**.

[7] P. A. V. D. Merwe, S. J. Davis, *Annu. Rev. Immunol.* ***2003***, 21, *659*.

[8] X. Yang, X. Yu, *Rev. Med. Virol.* ***2009***, 19, *77*.

[9] W.-P. Fung-Leung, M. W. Schilham, A. Rahemtulla, T. M. Kündig, M. Vollenweider, J. Potter, W. van Ewijk, T. W. Mak, *Cell* ***1991***, 65, *443*.

[10] C. A. Janeway, Jr, P. Travers, M. Walport, M. J. Shlomchik, *Antigen Presentation to T Lymphocytes*, New York, Garland Science **2005**.

[11] L. Backert, O. Kohlbacher, *Genome Med.* ***2015***, 7, *1*.

[12] J. S. Blum, P. A. Wearsch, P. Cresswell, *Annu. Rev. Immunol.* ***2013***, 31, *443*.

[13] D. R. Flower, I. Doytchinova, *Appl. Bioinform.* ***2002***, 1, *167*.

[14] M. C. Honeyman, V. Brusic, N. L. Stone, L. C. Harrison, *Nat. Biotechnol.* ***1998***, 16, *966*.

[15] I. Doytchinova, P. Petkov, I. Dimitrov, M. Atanasova, D. R. Flower, *Protein Sci.* ***2011***, 20, *1918*.

[16] C. A. Janeway, P. Travers, M. Walport, J. D. Capra, *Immunobiology: The Immune System in Health and Disease*, New York, Garland Science **2001**.

[17] C. A. Janeway, P. Travers, M. Walport, J. D. Capra, *Immunobiology: The Immune System in Health and Disease*, **2005**.

[18] E. E. Sercarz, E. Maverakis, *Nat. Rev. Immunol.* ***2003***, 3, *621*.

[19] R. Coico, G. Sunshine, *Immunology: A Short Course*, Chichester, West Sussex, UK; Hoboken, NJ, John Wiley & Sons Inc. **2009**.

[20] V. H. Engelhard, *Annu. Rev. Immunol.* ***1994***, 12, *181*.

[21] M. A. Batalia, E. J. Collins, *Pept. Sci.* ***1997***, 43, *281*.

[22] B. Loll, H. Fabian, H. Huser, C. S. Hee, A. Ziegler, B. Uchanska-Ziegler, A. Ziegler, *Arthritis Rheumatol.* ***2016***, 68, *1172*.

[23] Z. Zavala-Ruiz, I. Strug, B. D. Walker, P. J. Norris, L. J. Stern, *Proc. Natl. Acad. Sci. USA* ***2004***, 101, *13279*.

[24] J. Neefjes, M. L. Jongsma, P. Paul, O. Bakke, *Nat. Rev. Immunol.* ***2011***, 11, *823*.

[25] R. H. Schwartz, *Annu. Rev. Immunol.* ***1985***, 3, *237*.

[26] C. A. Weber, P. J. Mehta, M. Ardito, L. Moise, B. Martin, A. S. De Groot, *Adv. Drug Deliv. Rev.* ***2009***, 61, *965*.

[27] D. Mirano-Bascos, M. Tary-Lehmann, S. J. Landry, *Eur. J. Immunol.* ***2008***, 38, *1231*.

[28] Q. M. Sheikh, D. Gatherer, P. A. Reche, D. R. Flower, *Bioinformatics* ***2016***, 32, *3233*.

[29] W.-F. Lo, H. Ong, E. S. Metcalf, M. J. Soloski, *J. Immunol.* ***1999***, 162, *5*.

[30] H. Ito, M. Seishima, *BioMed Res. Int.* ***2010***, 2010, *1*.

[31] W. Zhang, Y. Niu, H. Zou, L. Luo, Q. Liu, W. Wu, *PloS One* ***2015***, 10, *e0128194*.

[32] P. Saxová, S. Buus, S. Brunak, C. Keşmir, *Int. Immunol.* ***2003***, 15, *781*.

[33] Y-F. Lu, H. Sheng, Y. Zhang, Z-y. Li, *J. Zhejiang Univ. Sci. B* ***2013***, 14, *816*.

[34] K. W. Jørgensen, M. Rasmussen, S. Buus, M. Nielsen, *Immunology* ***2014***, 141, *18*.

[35] S. Chen, D. Gfeller, S. A. Buth, O. Michielin, P. G. Leiman, C. Heinis, *Chembiochem* ***2013***, 14, *1316*.

[36] J. M. Pentier, A. K. Sewell, J. J. Miles, *Advances in T-cell epitope engineering. Investigating and harnessing T-cell functions with engineered immune receptors and their ligands*, **2015**, *68*.

[37] M. Irving, V. Zoete, M. Hebeisen, D. Schmid, P. Baumgartner, P. Guillaume, P. Romero, O. Michielin, *J. Biol. Chem.* **2012**, 287, *23068*.

[38] P. Dash, A. J. Fiore-Gartland, T. Hertz, G. C. Wang, S. Sharma, A. Souquette, *Nature* **2017**, 547, *89*.

[39] T. Trolle, M. Nielsen, *Immunogenetics* **2014**, 66, *449*.

[40] M. C. Feltkamp, M. P. Vierboom, W. M. Kast, C. J. Melief, *Mol. Immunol.* **1994**, 31, *1391*.

[41] N. C. Toussaint, M. Feldhahn, M. Ziehm, S. Stevanović, O. Kohlbacher, Editors. *T-cell Epitope Prediction Based on Self-tolerance*. Proceedings of the 2nd ACM Conference on Bioinformatics, Computational Biology and Biomedicine, **2011** ACM.

[42] A. Patronov, I. Doytchinova, *Open Biol.* **2013**, 3, *120139*.

[43] S. Paul, D. Weiskopf, M. A. Angelo, J. Sidney, B. Peters, A. Sette, *J. Immunol.* **2013**, 191, *5831*.

[44] T. M. Williams, *J. Mol. Diagn.* **2001**, 3, *98*.

[45] J. J. A. Calis, M. Maybeno, J. A. Greenbaum, D. Weiskopf, A. D. De Silva, A. Sette, C. Keşmir, B. Peters, *PLoS Comput. Biol.* **2013**, 9, *e1003266*.

[46] M. J. Blythe, I. A. Doytchinova, D. R. Flower, *Bioinformatics* **2002**, 18, *434*.

[47] V. Jawa, L. P. Cousens, M. Awwad, E. Wakshull, H. Kropshofer, A. S. De Groot, *Clin. Immunol.* **2013**, 149, *534*.

[48] T. F. Gregers, B. Fleckenstein, F. Vartdal, P. Roepstorff, O. Bakke, I. Sandlie, *Int. Immunol.* **2003**, 15, *1291*.

[49] M. Harndahl, M. Rasmussen, G. Roder, I. Dalgaard Pedersen, M. Sørensen, M. Nielsen, S. Buus, *Eur. J. Immunol.* **2012**, 42, *1405*.

[50] C. A. Lazarski, F. A. Chaves, S. A. Jenks, S. Wu, K. A. Richards, J. M. Weaver, A. J. Sant, *Immunity* **2005**, 23, *29*.

[51] S. H. Van Der Burg, M. Visseren, R. Brandt, W. M. Kast, C. Melief, *J. Immunol.* **1996**, 156, *3308*.

[52] L. A. Colf, A. J. Bankovich, N. A. Hanick, N. A. Bowerman, L. L. Jones, D. M. Kranz, K. C. Garcia, *Cell* **2007**, 129, *135*.

[53] K. C. Garcia, M. Degano, L. R. Pease, M. Huang, P. A. Peterson, L. Teyton, *Science* **1998**, 279, *1166*.

[54] M. Nielsen, C. Lundegaard, T. Blicher, B. Peters, A. Sette, S. Justesen, S. Buus, O. Lund, *PLoS Comput. Biol.* **2008**, 4, *e1000107*.

[55] M. Nielsen, C. Lundegaard, O. Lund, *BMC Bioinform.* **2007**, 8, *238*.

[56] P. Lasso, C. Cárdenas, F. Guzmán, F. Rosas, M. C. Thomas, M. C. López, J. M. González, C. J. Puerta, *Peptides* **2016**, 78, *68*.

[57] J. Liu, G. F. Gao, *Major Histocompatibility Complex: Interaction with Peptides. eLS*, John Wiley & Sons Ltd. **2011**.

[58] C. Meydan, H. H. Otu, O. U. Sezerman, *BMC Bioinform.* **2013**, 14, *S13*.

[59] E. M. Lafuente, P. A. Reche, *Curr. Pharma. Des.* **2009**, 15, *3209*.

[60] C. Lundegaard, O. Lund, M. Nielsen, *Bioinformatics* **2008**, 24, *1397*.

[61] M. N. Davies, D. R. Flower, *Drug Discov. Today* **2007**, 12, *389*.

[62] V. Brusic, G. Rudy, G. Honeyman, J. Hammer, L. Harrison, *Bioinformatics* **1998**, 14, *121*.

[63] C.-W. Tung, M. Ziehm, A. Kämper, O. Kohlbacher, S.-Y. Ho, *BMC Bioinform.* **2011**, 12, *446*.

[64] A. K. Nussbaum, C. Kuttler, K.-P. Hadeler, H.-G. Rammensee, H. Schild, *Immunogenetics* **2001**, 53, *87*.

[65] M. Nielsen, C. Lundegaard, O. Lund, C. Keşmir, *Immunogenetics* **2005**, 57, *33*.

[66] J. Hakenberg, A. K. Nussbaum, H. Schild, H.-G. Rammensee, C. Kuttler, H.-G. Holzhütter, *Appl. Bioinform.* **2002**, 2, *155*.

[67] M. Bhasin, G. Raghava, *Nucleic Acids Res.* **2005**, 33, *W202*.

[68] H.-G. Rammensee, J. Bachmann, N. P. N. Emmerich, O. A. Bachor, S. Stevanović, *Immunogenetics* **1999**, 50, *213*.

[69] H. Singh, G. Raghava, *Bioinformatics* **2001**, 17, *1236*.

[70] H. Singh, G. Raghava, *Bioinformatics* **2003**, 19, *1009*.

[71] P. Dönnes, O. Kohlbacher, *Nucleic Acids Res.* **2006**, 34, *W194*.

[72] J. Wan, W. Liu, Q. Xu, Y. Ren, D. R. Flower, T. Li, *BMC Bioinform.* **2006**, 7, *463*.

[73] M. Bhasin, G. Raghava, *Vaccine* **2004**, 22, *3195*.

[74] P. Guan, C. K. Hattotuwagama, I. A. Doytchinova, D. R. Flower, *Appl. Bioinform.* **2006**, 5, *55*.

[75] M. Nielsen, C. Lundegaard, P. Worning, S. L. Lauemøller, K. Lamberth, S. Buus, S. Brunak, O. Lund, *Protein Sci.* **2003**, 12, *1007*.

[76] J. R. A. Schafer, B. M. Jesdale, J. A. George, N. M. Kouttab, A. S. De Groot, *Vaccine* **1998**, 16, *1880*.

[77] I. Dimitrov, P. Garnev, D. R. Flower, I. Doytchinova, *Bioinformatics* **2010**, 26, *2066*.

[78] M. Bhasin, G. Raghava, *Bioinformatics* **2004**, 20, *421*.

[79] M. Bhasin, G. Raghava, *J. Biosci.* **2007**, 32, *31*.

[80] S. Lata, M. Bhasin, G. P. Raghava, *Immunoinformatics: Predicting Immunogenicity In Silico*, *Humana Press*, *Totowa*, *NJ* **2007**, *p. 201*.

[81] M. V. Larsen, C. Lundegaard, K. Lamberth, S. Buus, O. Lund, M. Nielsen, *BMC Bioinform.* **2007**, 8, *424*.

[82] T. Stranzl, M. V. Larsen, C. Lundegaard, M. Nielsen, *Immunogenetics* **2010**, 62, *357*.

[83] M. Nielsen, M. Andreatta, *Genome Med.* **2016**, 8, *1*.

[84] M. Andreatta, E. Karosiene, M. Rasmussen, A. Stryhn, S. Buus, M. Nielsen, *Immunogenetics* **2015**, 67, *641*.

[85] M. Andreatta, C. Schafer-Nielsen, O. Lund, S. Buus, M. Nielsen, *PLoS One* **2011**, 6, *e26781*.

[86] L. Zhang, Y. Chen, H.-S. Wong, S. Zhou, H. Mamitsuka, S. Zhu, *PLoS One* **2012**, 7, *e30483*.

[87] L. Jacob, J.-P. Vert, *Bioinformatics* **2008**, 24, *358*.

[88] E. Karosiene, C. Lundegaard, O. Lund, M. Nielsen, *Immunogenetics* **2012**, 64, *177*.

[89] M. Bhasin, S. Lata, G. Raghava, *Immunoinformatics: Predicting Immunogenicity In Silico*, *Humana Press*, *Totowa*, *NJ* **2007**, *p. 381*.

[90] P. Oyarzún, J. J. Ellis, M. Bodén, B. Kobe, *BMC Bioinform.* **2013**, 14, *52*.

[91] I. A. Doytchinova, P. Guan, D. R. Flower, *BMC Bioinform.* **2006**, *7*, *131*.

[92] G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew, A. J. Olson, *J. Comput. Chem.* **1998**, 19, *1639*.

[93] A. Logean, A. Sette, D. Rognan, *Bioorg. Med. Chem. Lett.* **2001**, 11, *675*.

[94] R. A. Friesner, J. L. Banks, R. B. Murphy, T. A. Halgren, J. J. Klicic, D. T. Mainz, M. P. Repasky, P. S. Shenkin, *J. Med. Chem.* **2004**, 47, *1739*.

[95] D. A. Case, T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang, R. J. Woods, *J. Comput. Chem.* **2005**, 26, *1668*.

[96] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, M. Karplus, *J. Comput. Chem.* **1983**, 4, *187*.

[97] B. R. Brooks, C. L. Brooks, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, M. Karplus, *J. Comput. Chem.* **2009**, 30, *1545*.

[98] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale, K. Schulten, *J. Comput. Chem.* **2005**, 26, *1781*.

[99] H. J. Berendsen, D. van der Spoel, R. van Drunen, *Comput. Phys. Commun.* **1995**, 91, *43*.

[100] B. Webb, A. Sali, *Curr. Protoc. Bioinform.* **2014**, 47, *5.6. 1–5.6. 32*.

[101] M. Biasini, S. Bienert, A. Waterhouse, K. Arnold, G. Studer, T. Schmidt, *Nucleic Acids Res*. ***2014***, *gku340*.

[102] L. A. Kelley, S. Mezulis, C. M. Yates, M. N. Wass, M. J. Sternberg, *Nat. Protoc*. ***2015***, 10, *845*.

[103] D. T. Jones, *J. Mol. Biol*. ***1999***, 287, *797*.

[104] S. Wu, Y. Zhang, *Proteins Struct. Funct. Bioinform*. ***2008***, 72, *547*.

[105] P. Dönnes, O. Kohlbacher, *Protein Sci*. ***2005***, 14, *2132*.

[106] D. V. Desai, U. Kulkarni-Kale, *Immunoinformatics* ***2014***, 333, *1184*.

[107] D. R. Jandrlić, G. M. Lazić, N. S. Mitić, M. D. Pavlović, *J. Biomed. Inform*. ***2016***, 60, *120*.

[108] J. C. Tong, T. W. Tan, S. Ranganathan, *Brief. Bioinform*. ***2006***, 8, *96*.

[109] E. S. Bergmann-Leitner, S. Chaudhury, N. J. Steers, M. Sabato, V. Delvecchio, A. S. Wallqvist, C. F. Ockenhouse, E. Angov, *PloS One* ***2013***, 8, *e71610*.

[110] D. Rajamani, S. Thiel, S. Vajda, C. J. Camacho, *Proc. Natl. Acad. Sci. USA* ***2004***, 101, *11287*.

[111] U. Hobohm, A. Meyerhans, *Eur. J. Immunol*. ***1993***, 23, *1271*.

[112] K. C. Parker, M. A. Bednarek, J. E. Coligan, *J. Immunol*. ***1994***, 152, *163*.

[113] Y. Altuvia, J. A. Berzofsky, R. Rosenfeld, H. Margalit, *Mol. Immunol*. ***1994***, 31, *1*.

[114] M. P. Davenport, I. A. H. Shon, A. V. Hill, *Immunogenetics* ***1995***, 42, *392*.

[115] H. Luo, H. Ye, H. W. Ng, L. Shi, W. Tong, D. L. Mendrick, *Bioinform. Biol. Insights* ***2015***, *9*, *21*.

[116] K. Kadam, S. Sawant, U. Kulkarni-Kale, V. K. Jayaraman, *Genomics III Methods, Techniques and Applications*, India, iConcept Press Ltd **2014**, 1.

[117] C. Savoie, N. Kamikawaji, T. Sasazuki, S. Kuhara, editors. *Use of BONSAI Decision Trees for the Identification of Potential MHC Class I Peptide Epitope Motifs*. Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing **1998**.

[118] M. R. Segal, M. P. Cummings, A. E. Hubbard, *Biometrics* ***2001***, 57, *632*.

[119] Y.-Y. Song, L. Ying, *Shanghai Arch. Psychiatry* ***2015***, 27, *130*.

[120] N. Nayab, *Bright Hub* ***2011 Sep***.

[121] Y. Xu, J. Zhuo, Y. Duan, B. Shi, X. Chen, X. Zhang, *Int. J. Clin. Exp. Pathol*. ***2014***, 7, *5569*.

[122] L. Tanner, M. Schreiber, J. G. H. Low, A. Ong, T. Tolfvenstam, Y. L. Lai, L. C. Ng, E. E. Ooi, *PLoS Negl. Trop. Dis*. ***2008***, 2, *e196*.

[123] P. Guan, I. A. Doytchinova, C. Zygouri, D. R. Flower, *Nucleic Acids Res*. ***2003***, 31, *3621*.

[124] D. R. Flower, *Trends Immunol*. ***2003***, 24, *667*.

[125] S. Wold, *Chemometr. Intell. Lab. Syst*. ***1995***, 30, *109*.

[126] S. T. Chang, D. Ghosh, D. E. Kirschner, J. J. Linderman, *Bioinformatics* ***2006***, 22, *2761*.

[127] I. A. Doytchinova, D. R. Flower, *Bioinformatics* ***2003***, 19, *2263*.

[128] C. K. Hattotuwagama, C. P. Toseland, P. Guan, D. J. Taylor, S. L. Hemsley, I. A. Doytchinova, D. R. Flower, *J. Chem. Inform. Model*. ***2006***, 46, *1491*.

[129] I. A. Doytchinova, D. R. Flower, *Curr. Proteom*. *2008*, *5*, *73*.

[130] D. Antonets, A. Maksyutov, *Mol. Biol*. ***2010***, 44, *119*.

[131] S. Lata, M. Bhasin, G. P. Raghava, *BMC Res. Notes* ***2009***, 2, *61*.

[132] M. Bhasin, H. Singh, G. Raghava, *Bioinformatics* ***2003***, 19, *665*.

[133] B. Peters, J. Sidney, P. Bourne, H.-H. Bui, S. Buus, G. Doh, *PLoS Biol*. ***2005***, 3, *e91*.

[134] K. Yu, N. Petrovsky, C. Schönbach, J. Y. Koh, V. Brusic, *Mol. Med*. ***2002***, 8, *137*.

[135] B. Zhao, K. R. Sakharkar, C. S. Lim, P. Kangueane, M. K. Sakharkar, *Int. J. Integr. Biol*. ***2007***, 1, *127*.

[136] A. S. De Groot, H. Sbai, C. Saint Aubin, J. McMurry, W. Martin, *Immunol. Cell Biol*. ***2002***, 80, *255*.

[137] U. Gowthaman, J. N. Agrewala, *J. Proteome Res*. ***2008***, 7, *154*.

[138] V. Brusic, V. B. Bajic, N. Petrovsky, *Methods* ***2004***, 34, *436*.

[139] J. C. Tong, *Position-Specific Scoring Matrices (PMMS): Encyclopedia of Systems Biology*, New York, Springer **2013**, p. 1722.

[140] P. A. Reche, J.-P. Glutting, E. L. Reinherz, *Hum. Immunol*. ***2002***, 63, *701*.

[141] B. Peters, W. Tong, J. Sidney, A. Sette, Z. Weng, *Bioinformatics* ***2003***, 19, *1765*.

[142] B. Peters, A. Sette, *BMC Bioinform*. ***2005***, *6*, *1*.

[143] R. E. Toes, A. K. Nussbaum, S. Degermann, M. Schirle, N. P. Emmerich, M. Kraft, C. Laplace, H. Schild, *J. Exp. Med*. ***2001***, 194, *1*.

[144] P. Baldi, S. Brunak, *Bioinformatics: The Machine Learning Approach*, Cambridge, Massachusetts, MIT Press **2001**.

[145] H. Mamitsuka, *Proteins Struct. Funct. Genet*. ***1998***, 33, *460*.

[146] S. Lek, M. Delacoste, P. Baran, I. Dimopoulos, J. Lauga, S. Aulagnier, *Ecol. Model*. ***1996***, 90, *39*.

[147] B. Yegnanarayana, *Sadhana* ***1994***, 19, *189*.

[148] C. Gershenson, *arXiv Preprint cs/0308031*. ***2003 Aug 20***.

[149] R. Smith, *Know Thine Enemy-Current Methods of Epitope Prediction*, London, UCL **2007**.

[150] P. P. Kuksa, M. R. Min, R. Dugar, M. Gerstein, *Bioinformatics* ***2015***, 31, *3600*.

[151] S. S. Soam, F. Khan, B. Bhasker, B. N. Mishra, *Bioinformation* ***2009***, 3, *403*.

[152] S. Lata, B. Sharma, G. Raghava, *BMC Bioinform*. ***2007***, 8, *263*.

[153] H. Noguchi, R. Kato, T. Hanai, Y. Matsubara, H. Honda, V. Brusic, T. Kobayashi, *J. Biosci. Bioeng*. ***2002***, 94, *264*.

[154] E. Birney, *IBM J. Res. Dev*. ***2001***, 45, *449*.

[155] B.-J. Yoon, *Curr. Genom*. ***2009***, 10, *402*.

[156] P. Blunsom, *Lect. Notes* ***2004***, 15, *18*.

[157] W. Liu, X. Meng, Q. Xu, D. R. Flower, T. Li, *BMC Bioinform*. ***2006***, 7, *1*.

[158] C. Cortes, V. Vapnik, *Mach. Learn*. ***1995***, 20, *273*.

[159] Y. Zhao, C. Pinilla, D. Valmori, R. Martin, R. Simon, *Bioinformatics* ***2003***, 19, *1978*.

[160] J. Salomon, D. R. Flower, *BMC Bioinform*. ***2006***, 7, *501*.

[161] V. Uslan, H. Seker, *Appl. Soft Comput*. ***2016***, 43, *210*.

[162] Y. Bengio, *Found. Trends Mach. Learn*. ***2009***, 2, *1*.

[163] L. Rampasek, A. Goldenberg, *Cell Syst*. ***2016***, 2, *12*.

[164] P. Mamoshina, A. Vieira, E. Putin, A. Zhavoronkov, *Mol. Pharma*. ***2016***, 13, *1445*.

[165] P. Aparoy, K. Kumar Reddy, P. Reddanna, *Curr. Med. Chem*. ***2012***, 19, *3763*.

[166] J. Verma, V. M. Khedkar, E. C. Coutinho, *Curr. Top. Med. Chem*. ***2010***, 10, *95*.

[167] I. A. Doytchinova, V. A. Walshe, N. A. Jones, S. E. Gloster, P. Borrow, D. R. Flower, *J. Immunol*. 172, *7495*.

[168] D. A. Winkler, *Brief. Bioinform*. ***2002***, 3, *73*.

[169] M. Akamatsu, *Curr. Top. Med. Chem*. ***2002***, 2, *1381*.

[170] E. N. Muratov, E. V. Varlamova, A. G. Artemenko, P. G. Polishchuk, V. E. Kuz'Min, *Mol. Inform.* ***2012***, 31, *202*.

[171] I. Dimitrov, P. Garnev, D. R. Flower, I. Doytchinova, *BioMed Res. Int.* ***2010***, 2010, *1*.

[172] L. Zhihua, W. Yuzhang, Z. Bo, N. Bing, W. Li, *J. Comput. Biol.* ***2004***, 11, *683*.

[173] I. A. Doytchinova, D. R. Flower, *J. Med. Chem.* ***2001***, 44, *3572*.

[174] K. Roy, S. Kar, R. N. Das, *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment*, Chennai, India, Academic Press **2015**.

[175] M. Lapinsh, P. Prusis, A. Gutcaits, T. Lundstedt, J. E. Wikberg, *Biochim. Biophys. Acta Gen. Subj.* ***2001***, 1525, *180*.

[176] J. Lee, S. Wu, Y. Zhang, *Ab Initio Protein Structure Prediction. From Protein Structure to Function with Bioinformatics*, Dordrecht, Amsterdam, Springer **2009**, p. 3.

[177] T. Schlick, *Molecular Modeling and Simulation: An Interdisciplinary Guide: An Interdisciplinary Guide*, New York, USA, Springer Science & Business Media **2010**.

[178] T. Akutsu, K. L. Sim, *Genome Inform.* ***1999***, 10, *23*.

[179] M. Castelli, F. Cappelletti, R. A. Diotti, G. Sautto, E. Criscuolo, M. Dal Peraro, N. Clementi, *Clin. Dev. Immunol.* ***2013***, 2013, *1*.

[180] Y. Altuvia, O. Schueler, H. Margalit, *J. Mol. Biol.* ***1995***, 249, *244*.

[181] Y. Altuvia, H. Margalit, *Methods* ***2004***, 34, *454*.

[182] O. Schueler-Furman, Y. Altuvia, A. Sette, H. Margalit, *Protein Sci.* ***2000***, 9, *1838*.

[183] M. J. Bower, F. E. Cohen, R. L. Dunbrack, *J. Mol. Biol.* ***1997***, 267, *1268*.

[184] R. Abagyan, S. Batalov, T. Cardozo, M. Totrov, J. Webber, Y. Zhou, *Proteins Struct. Funct. Genet.* ***1997***, 29, *29*.

[185] T. Schwede, J. Kopp, N. Guex, M. C. Peitsch, *Nucleic Acids Res.* ***2003***, 31, *3381*.

[186] P. Bourne, H. Weissig, *Methods Biochem. Anal.* ***2003***, 41, *D475*.

[187] N. Poorinmohammad, H. Mohabatkar, *J. Arthropod-Borne Dis.* ***2015***, 9, *116*.

[188] A. Ingale, *J. Comput. Sci. Syst. Biol.* ***2012***, 2010, *53*.

[189] A. Nayeem, D. Sitkoff, S. Krystek, *Protein Sci.* ***2006***, 15, *808*.

[190] X.-Y. Meng, H.-X. Zhang, M. Mezei, M. Cui, *Curr. Computer-Aided Drug Des.* ***2011***, 7, *146*.

[191] A. Dhanik, L. E. Kavraki, *Protein–Ligand Interactions: Computational Docking. eLS*, Wiley Online Library **2001**.

[192] N. Brooijmans, *Docking Methods, Ligand Design, and Validating Data Sets in the Structural Genomic Era. Structural Bioinformatics*, Wiley, United States **2009**. p. 635.

[193] D. B. Kitchen, H. Decornez, J. R. Furr, J. Bajorath, *Nat. Rev. Drug Discov.* ***2004***, 3, *935*.

[194] Y. Hou, Y. Guo, C. Wu, N. Shen, Y. Jiang, J. Wang, *PloS One* ***2012***, 7, *e39344*.

[195] E. Yuriev, P. A. Ramsland, *J. Mol. Recogn.* ***2013***, 26, *215*.

[196] S. F. Sousa, A. J. M. Ribeiro, J. T. S. Coimbra, R. P. P. Neves, S. A. Martins, N. S. H. N. Moorthy, P. A. Fernandes, M. J. Ramos, *Curr. Med. Chem.* ***2013***, 20, *2296*.

[197] E. Yuriev, J. Holien, P. A. Ramsland, *J. Mol. Recogn.* ***2015***, 28, *581*.

[198] R. Abel, T. Young, R. Farid, B. J. Berne, R. A. Friesner, *J. Am. Chem. Soc.* ***2008***, 130, *2817*.

[199] M. L. Verdonk, J. C. Cole, M. J. Hartshorn, C. W. Murray, R. D. Taylor, *Proteins Struct. Funct. Bioinform.* ***2003***, 52, *609*.

[200] M. Atanasova, A. Patronov, I. Dimitrov, D. R. Flower, I. Doytchinova, *Protein Eng. Des. Select.* ***2013***, *gzt018*.

[201] A. Patronov, I. Dimitrov, D. R. Flower, I. Doytchinova, *BMC Struct. Biol.* ***2012***, 12, *1*.

[202] M. S. Sakib, M. R. Islam, A. Hasan, A. Nabi, *Adv. Bioinform.* ***2014***, 2014, *1*.

[203] A. Logean, D. Rognan, *J. Computer-Aided Mol. Des.* ***2002***, 16, *229*.

[204] M. Munikumar, I. V. Priyadarshini, D. Pradhan, A. Umamaheswari, B. Vengamma, *Interdiscipl. Sci. Comput. Life Sci.* ***2013***, 5, *155*.

[205] B. G. Pierce, Z. Weng, *Protein Sci.* ***2013***, 22, *35*.

[206] J. J. Adams, S. Narayanan, B. Liu, M. Birnbaum, A. Kruse, N. Bowerman, W. Chen, K. C. Garcia, *Immunity* ***2011***, 35, *681*.

[207] D. R. Flower, K. Phadwal, I. K. Macdonald, P. V. Coveney, M. N. Davies, S. Wan, *Immunome Res.* ***2010***, 6, *S4*.

[208] G. Scarabelli, G. Morra, G. Colombo, *Biophys. J.* ***2010***, 98, *1966*.

[209] S. Wan, P. V. Coveney, D. R. Flower, *Immunoinformatics: Predicting Immunogenicity In Silico, Humana Press, Totowa, NJ* ***2007***. *p. 321*.

[210] M. N. Davies, D. R. Flower, *Immunoinformatics: Predicting Immunogenicity In Silico*, Humana Press, Totowa, NJ **2007**. p. 309.

[211] D. Rognan, L. Scapozza, G. Folkers, A. Daser, *Biochemistry* ***1994***, 33, *11476*.

[212] B. Knapp, U. Omasits, B. Bohle, B. Maillere, C. Ebner, W. Schreiner, B. Jahn-Schmid, *Mol. Immunol.* ***2009***, 46, *1839*.

[213] J.-S. Lim, S. Kim, H. G. Lee, K.-Y. Lee, T.-J. Kwon, K. Kim, *Mol. Immunol.* ***1996***, 33, *221*.

[214] S. Wan, D. R. Flower, P. V. Coveney, *Mol. Immunol.* ***2008***, 45, *1221*.

[215] H. Zhang, C. Lundegaard, M. Nielsen, *Bioinformatics* ***2009***, 25, *83*.

[216] R. H. Cheng, T. Miyamura, *Structure-based Study of Viral Replication: With CD-ROM*, New Jersey, USA, World Scientific **2008**.

[217] S. Gupta, P. Kapoor, K. Chaudhary, A. Gautam, R. Kumar, G. P. Raghava, *PLoS One* ***2013***, 8, *e73957*.

[218] V. Mathura, P. Kangueane, *Bioinformatics: A Concept-Based Introduction*, New York, USA, Springer Science & Business Media **2008**.

[219] I. Antes, S. W. Siu, T. Lengauer, *Bioinformatics* ***2006***, 22, *e16*.

[220] D. K. Cole, E. S. J. Edwards, K. K. Wynn, M. Clement, J. J. Miles, K. Ladell, J. Ekeruche, A. K. Sewell, *J. Immunol.* ***2010***, 185, *2600*.

AUTHOR BIOGRAPHIES

Ms Prattusha Kar Prattusha did her Bachelor and Master's degrees in electrical engineering in India. She started her PhD under Prof. Mancera at the School of Pharmacy and Biomedical Sciences, Curtin University in 2016. Her research focuses on biomolecular modelling.

Ms Lanie Ruiz-Perez Lanie Ruiz-Perez was born in Colombia and obtained a biological engineering degree from the National University of Colombia (Medellin) in 2011. During this she carried out work in the area of human immunovirology. Currently, she is pursuing a Master's degree by researching the interaction of drug-like molecules with lipid bilayers using biomolecular modelling

Prof. Ricardo Mancera Ricardo Mancera received a BSc in biological and pharmaceutical chemistry from the National University of Mexico (1993), and a PhD in theoretical chemistry from the University of Cambridge (1997). He was a Founding Senior Scientist at De Novo Pharmaceuticals in Cambridge before moving to Curtin University, Australia, where he is now Professor of Biophysical Chemistry and Computational Biophysics. His research focuses on the use of molecular dynamics simulation methods to study biomolecular structure and interactions.

How to cite this article: Kar P, Ruiz-Perez L, Arooj M, Mancera RL. Current methods for the prediction of T-cell epitopes. *Peptide Science*. 2018;110:e24046. https://doi.org/10.1002/pep2.24046