Title of Thesis

*"AI based prediction of HLA-DRB1*04:01 binder for designing subunit vaccines"*

by
Sumeet Patiyal

Under the Supervision of
Prof. Gajendra P.S. Raghava

Indraprastha Institute of Information Technology Delhi

2022

*"AI based prediction of HLA-DRB1*04:01 binder for designing subunit vaccines"*

by
Sumeet Patiyal
PhD17204

Submitted
in partial fulfilment of the requirements for the degree of
Master of Technology

to

Indraprastha Institute of Information Technology Delhi
2022

Certificate

This is to certify that the thesis titled *"**AI based prediction of HLA-DRB1*04:01 binder for designing subunit vaccines***" being submitted by Sumeet Patiyal to the Indraprastha Institute of Information Technology Delhi, for the award of the Master of Technology, is an original research work carried out by him under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

May 2022

Prof. Gajendra P.S. Raghava
Head of Department
Department of Computational Biology
Indraprastha Institute of Information Technology, Delhi
New Delhi 110020

Acknowledgment

This work would not have been possible without the guidance and support of several individuals who in one way or the other assisted in the preparation and completion of this thesis.

I would like to express my sincere gratitude and respect towards Prof. G.P.S Raghava from Indraprastha Institute of Information and Technology, Delhi for being my supervisor. I would also like to thank Ms. Anjali Dhall, my lab colleague for providing me the time support with enthusiasm whenever required.

My supervisor has inspired me to do good work and has supported me all along the way. Dozens of people have helped and taught me immensely through this year long journey. I would like to thank my friends and batchmates for providing such a good environment for helping me whenever I got stuck.

Lastly, I would also like to thank my parents and family for motivating from time to time throughout the course of my thesis which enabled me to pursue my research in an efficient and structured manner.

Sumeet Patiyal

PhD17024

Department of Computational Biology

Table of Contents

List of Figures

List of Tables

ABSTRACT

HLA gene complex is a highly polymorphic region in the human genome and mutations associated with these regions can lead to many deadly disorders such as bare lymphocyte syndrome, whereas presence of few HLA-class II alleles makes an individual more prone to some diseases. One of these class-II alleles named HLA-DRB1*04:01 is associated with many autoimmune disorders such as multiple sclerosis, rheumatoid arthritis, type 1 diabetes, Lyme disease, etc. Moreover, a particular variant of HLA-DRB1*04:01 gene is found to be abundant in the asymptomatic carriers of SARS-CoV-2. Hence, it is the need of the hour to develop a more accurate method with the ability to classify HLA-DRB1*04:01 binding peptides. We have developed a systematic approach to predict, scan, and design the binders of class-II HLA allele HLA-DRB1*04:01 and provided as a webserver. It is an updated version HLADR4Pred developed in year 2004. In this study, we have compiled the positive (HLA-DRB1*04:01 binder) and negative dataset (HLA-DRB1*04:01 non-binder) from IEDB. We have a total 12676 peptides in the positive and 86300 peptides in the negative dataset. At first, we generated composition and binary profile based features using the Pfeature standalone package. After that we have implemented various machine learning techniques to develop prediction models by using different types of features. Secondly, we have segregated the complete dataset into training and validation dataset, where training dataset comprises 80% of the complete dataset and the remaining 20% was assigned as validation dataset. We have trained the models on the training dataset by applying a five-fold cross validation technique and performed external validation by evaluating our models on the validation dataset. Number of performance measures have been calculated to assess the performance of each model developed on different features. We observed that the extra tree classifier based model developed on dipeptide composition based features outperformed other classifiers and achieved maximum AUROC of 0.96 on both training and validation dataset. After that, we have combined similarity search using BLAST with our best performing model to develop the hybrid method, which attains the highest performance i.e. AUROC of 0.98 and 0.99 on training and validation dataset, respectively. Finally, we have incorporated the hybrid model in our webserver named HLADR4Pred2 available at https://webs.iiitd.edu.in/raghava/hladr4pred2/. Along with that we have also provided the python- and Perl based standalone package which is available at webserver (https://webs.iiitd.edu.in/raghava/hladr4pred2/standalone.php) and at GitHub (https://github.com/raghavagps/hladr4pred2).

Chapter 1
Introduction

1. Introduction

The human leukocyte antigen (HLA) complex is a highly polymorphic genomic region located at chromosome 6 in the human genome (1,2). Majority of genes located in this region are encode several proteins of immune defence system (3). The HLA system is classified into three major categories I, II and III, where I (HLA-A, -B, -C) and II (HLA-DP, -DQ, -DR) genes are polymorphic in nature (4). IMGT/HLA the largest repository of HLA related sequences report thousands of human major histocompatibility complex associated alleles and genomic sequences (5). HLA are the crucial components of our immune system and stimulate immune responses to fight against several pathogens and autoimmune disorders (6,7). HLA class-I molecules display intracellular peptides to CD8+ T cells whereas HLA class-II molecules composed two polypeptide chains (α and β) and presents extracellular peptides to CD4+ T cells. HLA-class II alleles mainly presented on antigen presenting cells for instance, B cells, macrophages, DCs etc (8–10).

The binding groove of MHC-II molecules is open from both sides which enables long length peptides to enlarge the binding grooves from the flanking regions as shown in Figure 1. (11). The peptides binds to the MHC-II molecules sharing specific anchor residues. Typically, MHC-II alleles have four anchor residues P1, P4, P6 and P9, the peptide binds to allele-specific binding groove and it may vary due to high polymorphism (12). Moreover, the anchor residues of class-II MHC alleles also vary therefore, it allows a wide range of peptides to bind to its surface. Majority of MHC class-II alleles presented peptides which were derived from the pathogenic proteins (13,14). MHC Class-II alleles carry a peptide and express it on the cell surface; further it interact with T cell receptors (Figure 1C) and activate CD4+ T-cells the immune responses via secreting cytokines such as IFN-gamma, TNF and GM-CSF.

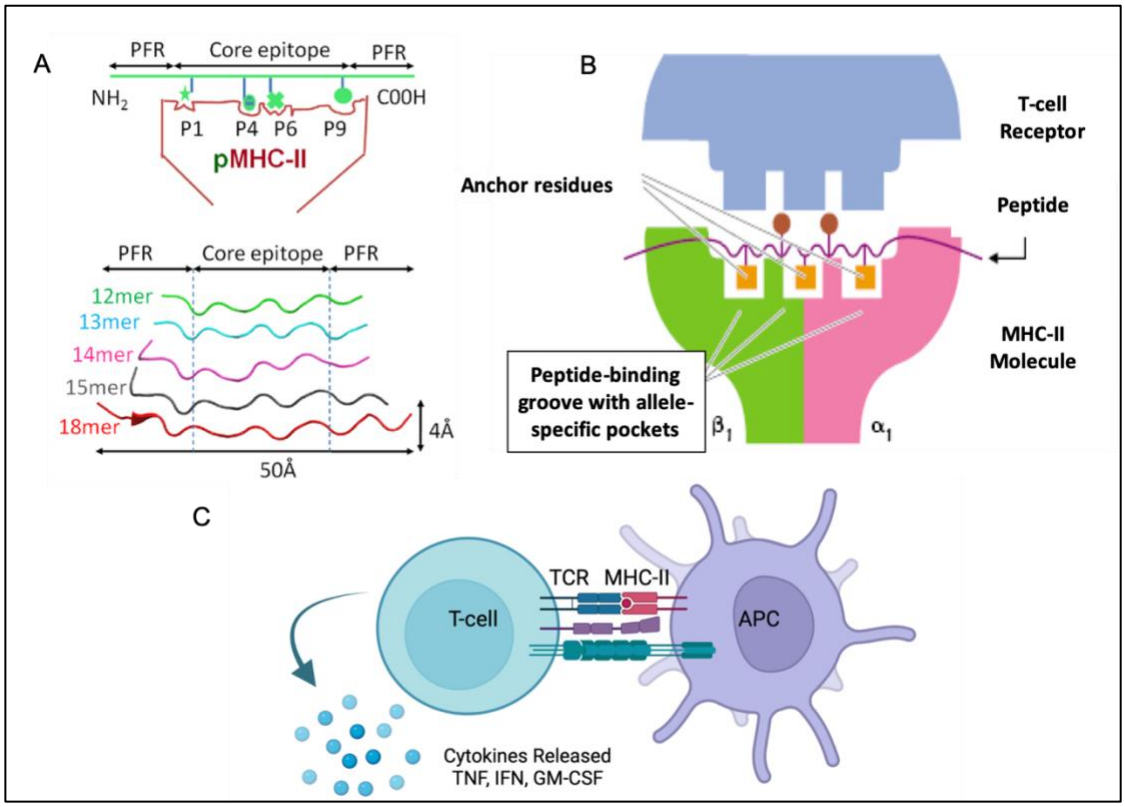


Figure 1: Pictorial representation of peptide conformations presented by MHC-II molecules, anchor residues of peptides bound to the allele-specific pockets of MHC-II molecule (Figure source (11,15))

Several studies report that more than 200 immune-mediated disorders are associated with HLA molecules; however, HLA is the major genetic factor in developing autoimmune diseases. In the past, studies have shown that the HLA-DR4 gene is highly correlated with several diseases (16–19), especially HLA-DRB1∗04:01 is associated with the development of multiple sclerosis (20,21), autoimmune disorders (AID) (22), type 1 diabetes (23), Lyme disease (24), COVID-19 severity, and rheumatoid arthritis (25) as shown in Figure 2. HLA-DR4 molecules plays a significant role in autoimmune disorders initiation and progression. Therefore, it is of utmost importance to determine the epitopes which bind to HLA-DRB1*04:01 in order to understand or cure several autoimmune disorders (26–30). Studies also reveal that patients positive with HLA-DR4 associated alleles have maximum chances of having autoimmune disorders therefore it could be a significant as genetic biomarker. Researcher developed a number of experimental techniques for the detection of HLA-peptide bindings, but they are time-exhaustive and cost-effective (31,32).

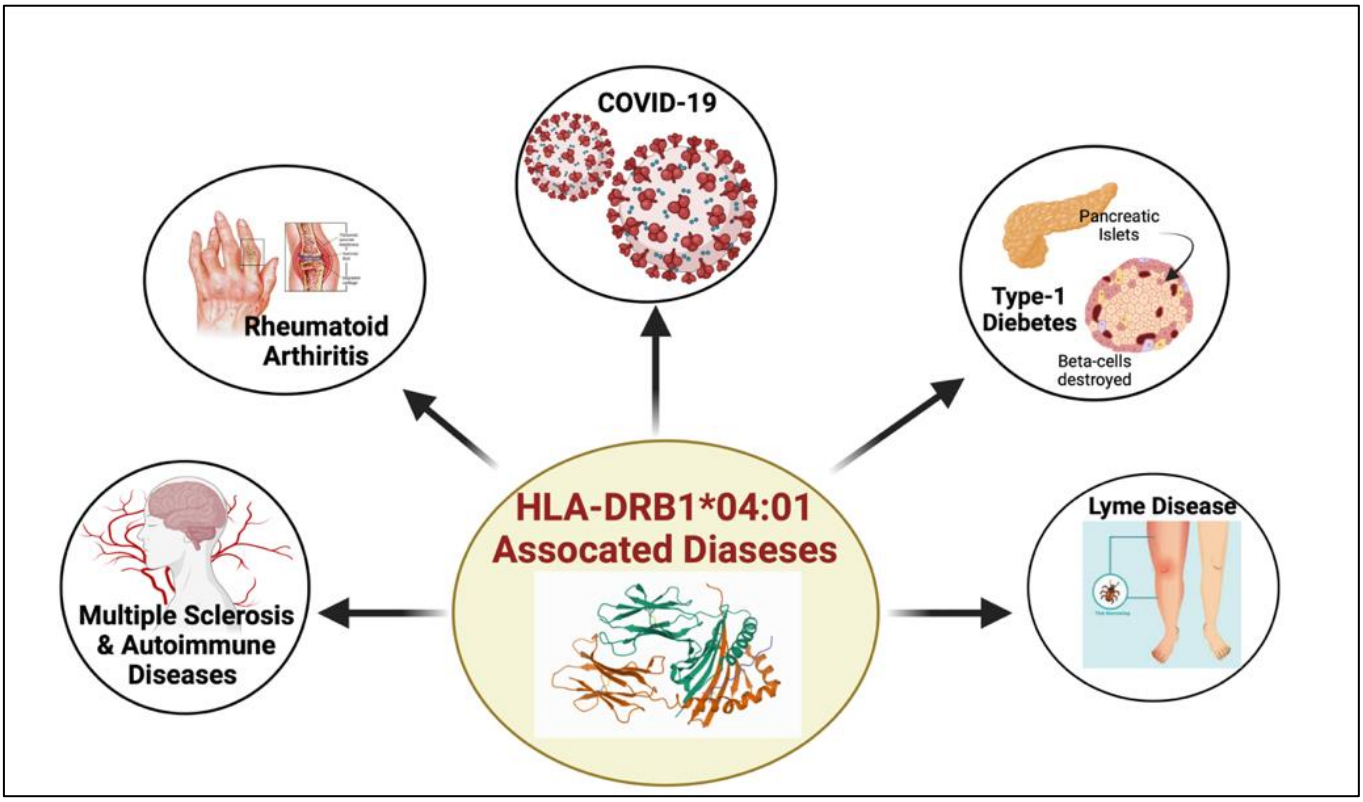


Figure 2: Pictorial representation of association of HLA-DRB1*04:01 allele with number of diseases.

Therefore, many attempts have been made to develop computational tools to predict the binding peptides associated with class-I HLA-alleles. However, fewer methods have been developed for HLA class-II molecules binder prediction due to the variable length of binding peptides and uncertain core/anchor residue positions (33–36). From last few years, several in-silico tools have been developed for the prediction of HLA-DR binding peptides, based on the sequence and structure information.

Bhasin et al., developed SVM based approach for the prediction of HLA-DRB1*04:01-binding peptides and archived 86% accuracy (37). PROPRED method uses quantitative matrices for the prediction of HLA-DRB1*04:01-binding peptide (38). Whereas, SMM-align uses stabilization matrix alignment method for the prediction of peptide-MHC binding affinities (39). ARB matrix binding prediction tool utilizes average relative binding matrix method for direct prediction of binding affinity and IC50 values (40). In addition, NNAlign_MA (41), NetMHCpan (42) and NetMHCIIpan (43) uses motif convolution and mass spectrometry data for the better prediction of HLA-II binding peptides (42,44).

In this study, we have developed a computational approach to classify the HLA-DRB1*04:01 binding peptides using the sequence information. We have obtained the experimentally validated HLA-DRB1*04:01 binding peptides of length 9-22 amino acids from

IEDB to train and evaluate the prediction models. We have implemented various machine learning classifiers and hyper tuned the parameters to improve the performance of the generated model. We hope that this study will benefit the researchers working in the field of cellular immunology, vaccine design, immunodiagnostics, immunotherapeutic, and molecular understanding of autoimmune susceptibility. In order to serve the scientific community, we have developed the user-friendly webserver "HLADR4Pred 2.0" available at URL https://webs.iiitd.edu.in/raghava/hladr4pred2. We have also developed python and Perl-based standalone available at webserver https://webs.iiitd.edu.in/raghava/hladr4pred2/standalone.php and at GitHub https://github.com/raghavagps/hladr4pred2 with how-to-use instructions.

Chapter 2
Review of literature

It has been shown in literature that some of the HLA-DR proteins are highly expressed in patients with autoimmune diseases such as rheumatoid arthritis, hence it is of utmost important to identify the peptides interacting with class II MHC molecules to understand the mechanism and can aid in designing the novel therapeutics to treat associated diseases. Class II MHC molecules binding peptides interact with T-cell receptors and initiate subsequent T-cell activation which further drive immune responses. The malfunctioning of immune responses can lead to several immune related and autoimmune disorder. Several studies report that the MHC-II binding antigenic peptides/epitope can used as major immunotherapy target in the treatment of cancer and other diseases. Therefore, it is mandatory to identify particular peptides/epitopes which are strong binders to MHC-II molecules and have the capacity to interact with T-cell receptors. A number of factors such as binding groove pocket, polar, charged or hydrophobic residues are responsible for the binding of peptides with the MHC-II molecule. In past several in vitro studies have been used to determine the binding peptides of MHC-II molecules.

In-vitro tests assessing the binding of synthetic peptides to HLA molecules were the most common MHC binding assays from the 1990s through 2010 (45). Whereas, mass spectrometry sequencing first recorded in 1991 for HLA eluted peptides sequencing has been more common in the recent five years (46–48). In addition, high throughput peptide-MHC II binding tests can help guide epitope selection by assessing peptide binding affinity and binding promiscuity across various MHC II alleles. Such experimental procedures, like protein deimmunization, are eventually required to test computational predictions of promising vaccine candidates. Biochemical experiments using recombinant human MHC II molecules can offer quick, quantitative information on immunogenic epitope identification, deletion, and design (49–51). As demonstrated by X-ray crystallographic studies the MHC class II epitope binding sites are made up of a binding-groove and multiple pockets provided by alpha-sheet and two beta-helices (52,53) and both ends of the class II binding groove are open. As a result, peptides that bind to class II molecules have a wide range of lengths, ranging from 13 to 25 residues.

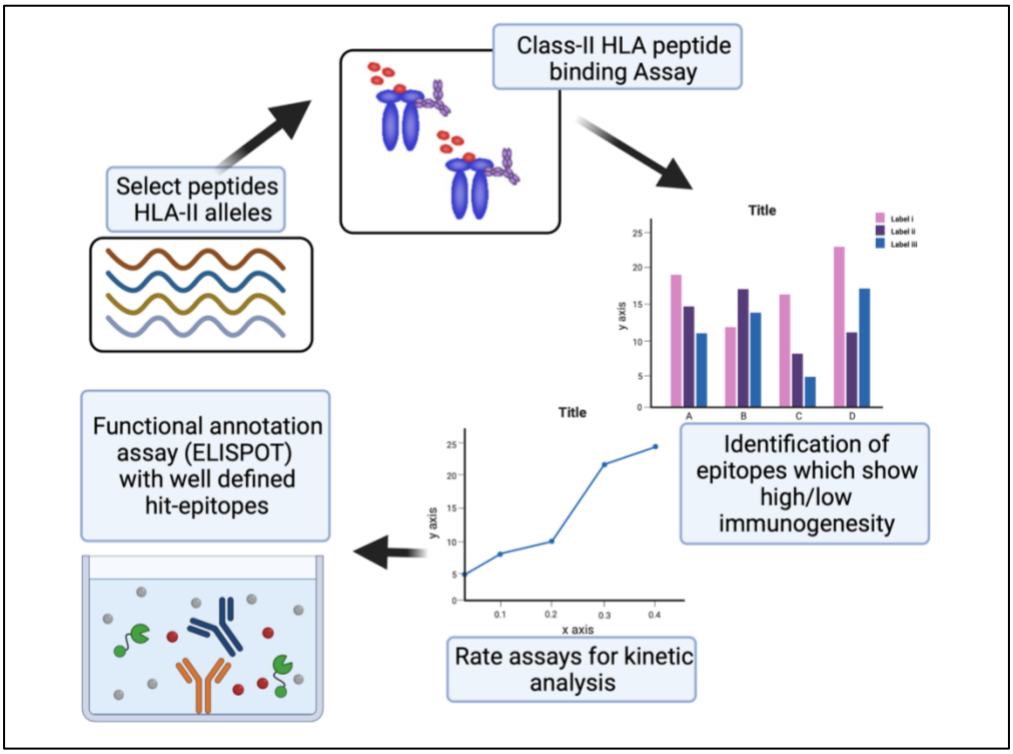


Figure 3: Class-II HLA binding assay

[Source: BioRender, https://www.proimmune.com/mhc-class-ii-binding-assays/]

But, identification of MHC II binding peptides through experimental approaches a very cost- and time-intensive process. On the contrary, computational approaches are the time- and cost-effective to find out the binders. However, prediction of MHC class II binders is a strenuous process as compared to the prediction of binders in case of class I molecules because of the differences in the length of the peptides, unidentified core, and versatility in the anchor residues that interact with the grooves. There are number of methods have been developed in the past for the prediction of MHC class II binding peptides based on different algorithms. Initially, prediction methods were developed on the motifs (42,54), followed by matrix based (38,39,55–57). Then, machine learning based methods (37,56,58–60) have overtaken with improved accuracy. Consensus IEDB (61) is also a method which make predictions by drawing consensus from the already existing methods such as SMM-align (39), and NN-align (56). Structure-based methods (62–64) have also been developed in the past for MHC class-II binders as shown in Table 1 . On the other hand, there are methods which are specific to HLA-DR binding peptides based on motifs (65–68), weighted-matrices (62,69–71), and machine-learning (37). Table 1 comprises the tools developed for prediction of binders for MHC class II as well as specific for HLA-DR alleles in the last two decades.

Table 1: Compilation of tools for the prediction of MHC class II binding peptides

Tool	Year	Web Link	Working	Alleles	Description
Propred (38)	2001	https://webs.iiitd.edu.in/raghava/propred/	YES	51 HLA-DR allele	Prediction of promiscuous HLA-DR binders
HLA-DR4Pred (37)	2004	https://webs.iiitd.edu.in/raghava/hladr4pred/	YES	HLA-DRB1*0401(MHC class II allele)	SVM and ANN based HLA-DR4*04:01 binder prediction tool
Consensus (61)	2008	http://tools.immuneepitope.org/mhcii/	YES	20 MHC-II alleles	IEDB tool for predicting MHC Class II binders
MULTIPRED 2 (59)	2005	http://antigen.i2r.a-star.edu.sg/multipred/	NO	class I-A, B and class II DR supertypes	ANN based method for HLA-binder prediction
SMM-align (39)	2007	NA	NO	14 HLA-DR (human MHC) and three mouse H2-IA alleles	Prediction of MHC class II binding affinity using matrix alignment method
MHC2pred	-	https://webs.iiitd.edu.in/raghava/mhc2pred/	YES	42 MHC-II alleles	SVM based method for MHC-II binder prediction
NN-align (56)	2009	https://services.healthtech.dtu.dk/service.php?NetMHCIIpan-4.0	YES	14 HLA-DR (human MHCII)	Artificial neural network-based alignment algorithm for MHC class II peptide binding prediction
EpiTOP (57)	2010	http://www.pharmfac.net/EpiTOP/	NO	12 HLA-DRB1 alleles	Prediction of MHC class II binding using quantitative matrix
Tepitopepan (62)	2012	http://www.biokdd.fudan.edu.cn/Service/TEPITOPEpan/	NO	51 HLA-DR Molecules	Uses pocket binding specificities for HLA-DR binder prediction
EpiDock (63)	2013	http://www.ddg-pharmfac.net/epidock/EpiDockPage.html	YES	23 MHC-II alleles	Molecular docking based tool for MHC-II binding prediction
NetMHC-II (60)	2018	https://services.healthtech.dtu.dk/service.php?NetMHCIIpan-3.2	YES	MHC class II isotypes HLA-DR, HLA-DP and HLA-DQ, as well as mouse molecules (H-2)	MHC-II Binding affinity prediction
MHCII3D (64)	2020	https://pbwww.services.came.sbg.ac.at/mhcii3d/	NO	25 MHC-II alleles	Structure based prediction of MHC-II binding peptides

Chapter 3
Materials & Methods

3.1. Dataset Creation and Preprocessing

We have extracted experimentally validated HLA-Class II allele HLA-DRB1*04:01 binding peptides from the immune epitope database (IEDB) (72). Initially, total number of binding peptides extracted from IEDB was 19665 with length varying from 8 to 32 amino acids. After removing the identical peptides and peptides containing non-natural amino acids, we left with 12880 unique peptides. As shown in Figure 4, the peptide length analysis exhibited that 98.4% i.e. 12676 peptides were having length between 9-22 amino acids, hence we selected 12676 peptides and constitute our positive dataset.

In such prediction methods, one of the major challenges is to obtain the experimentally validated negative dataset, i.e. non-binders of HLA-DRB1*04:01 allele. In order handle that we have downloaded the HLA-class II binders from IEDB except the binders of HLA-DRB1*04:01 which resulted into 154534 peptides. After applying the aforementioned constraints, we were left with 86300 peptides having length between 9-22 amino acids. To avoid the biasness in the negative dataset, we have made another dataset comprises of 12676 peptides having length between length 9-22 generated randomly using Swiss-Prot database release 2022_01 (73).

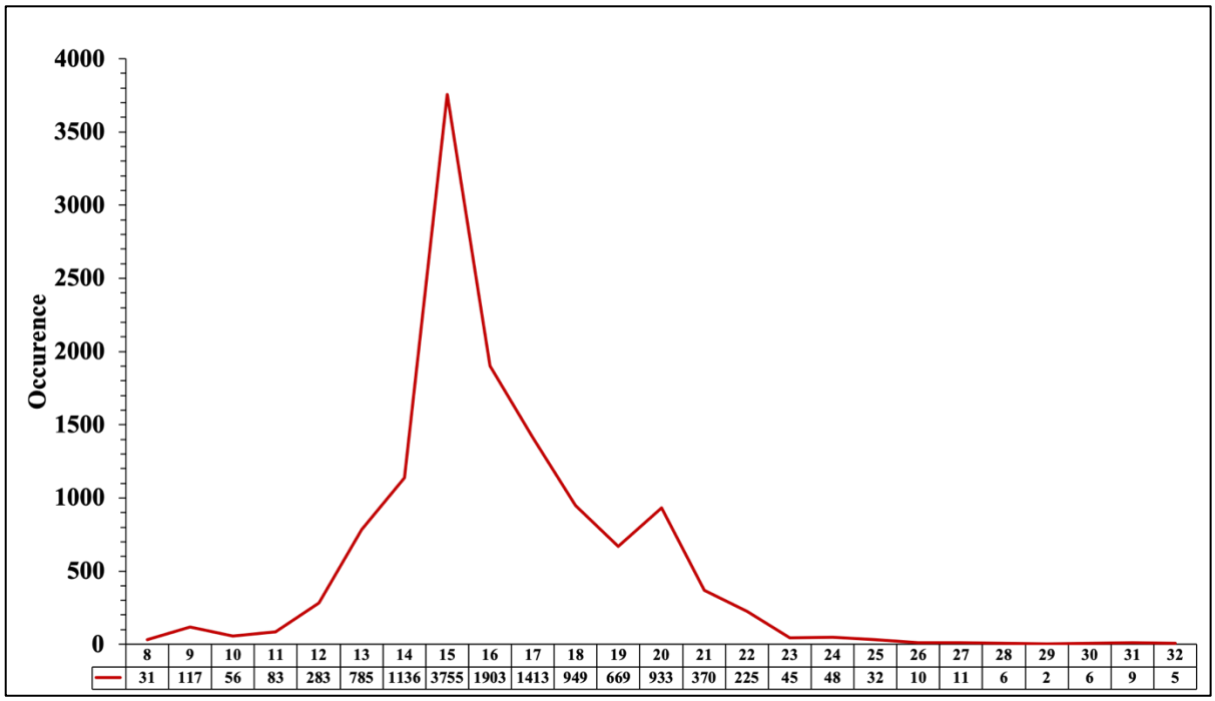


Figure 4: Length-wise distribution of HLA-DRB1*04:01 binding peptides

Finally, we have generated three different datasets to train and evaluate the models and named them balanced dataset, alternate dataset, and realistic dataset. Where, balanced dataset

comprises of 12676 HLA-DRB1*04:01 binding and 12676 non-binding peptides derived from IEDB; alternate dataset consists of 12676 binding and non-binding peptides randomly generated using Swiss-Prot database; and realistic dataset contains 12676 HLA-DRB1*04:01 binding and 86300 non-binding peptides derived using IEDB. Each dataset was further divided into training and validation dataset, where 80% of the data constitute training and the remaining 20% make validation dataset. To avoid the biasness in the length distribution in training and validation dataset, we have arranged all peptides as per their length and then transferred every fifth peptide into the validation dataset and rest constitutes training dataset.

3.2 Composition Analysis

To check the abundance of each amino acid in each dataset, we have calculated the composition of each amino acid using equation 1. We have implemented the amino acid composition module of Pfeature to calculate the composition of positive and negative set separately in each dataset.

$$CR_i = \frac{NR_i}{TR} \quad [1]$$

Where, CR_i represents composition of residue i; NR_i is total number of residues of type i; and TR stands for total number of residues.

3.3Position Conservation Analysis

To understand the position specific preference of residues, we have created the logos using two-sample logo (TSL) (74) webserver. In order to create the logo, it is prerequisite to fix the peptide length. Since, the minimum length of the considered peptide is 9, hence to achieve the fix length criteria we have taken the 9 residues from N-terminal and 9-residues from C-terminal. Finally, to create a fix length peptide with 18 residues we have joined both the regions. We have created the TSL for each dataset i.e. balanced, alternate, and realistic dataset.

3.4 Generation of Features

In this study to represent the sequence as a numerical vector, we have implemented the composition and binary profile module of Pfeature (75). By using Pfeature we have computed a wide range of features such as, composition- and binary profile-based features. Using composition module we have calculated fifteen different type of features such as amino acid

composition (AAC), dipeptide composition (DPC), atomic composition (ATC), bond composition (BTC), physico-chemical properties based composition (PCP), residue repeat information (RRI), distance distribution of residues (DDOR), Shannon entropy for all residues (SER), Shannon entropy based on physico-chemical properties (SPC), conjoint triad calculation (CTC), composition enhanced transition and distribution (CeTD), pseudo amino acid composition (PAAC), amphiphilic pseudo amino acid composition (APAAC), quasi-sequence order (QSO), and sequence order coupling number (SOCN). By implementing binary profile-based module, we have calculated four different features such as binary profile of first nine residues (N_9), binary profile of last nine residues (C_9), and combination of N_9 and C_9 binary profile (N_9C_9). In order to make it more clear, we have shown the example sequences of different length in Table 2, and highlighted the regions in the sequences which is designated as N_9, C_9 and N_9C_9, respectively.

Table 2: Generation of N_9, C_9, and N_9C_9 patterns from the original sequences with varying length

Original Sequences	N9	C9	N9C9
TQQKKADRY	**TQQKKADRY**	**YRDAKKQQT**	**TQQKKADRYYRDAKKQQT**
ISAYLLSKNNAI	**ISAYLLSKN**NAI	**IANNKSLLY**ASI	**ISAYLLSKNIANNKSLLY**
GTFQKWAAVVVPSGE	**GTFQKWAAV**VVPSGE	**EGSPVVVAA**WKQFTG	**GTFQKWAAVEGSPVVVAA**
SAIEYTIENVFESAPNPR	**SAIEYTIEN**VFESAPNPR	**RPNPASEFV**NEITYEIAS	**SAIEYTIENRPNPASEFV**
LPGDKSKAFDFLSEETEASLAS	**LPGDKSKAF**DFLSEETEASLAS	**SALSAETEE**SLFDFAKSKDGPL	**LPGDKSKAFSALSAETEE**

Similarly, binary profile for pattern size with twenty-two residues (NC_{22}) were also generated. The major challenge in calculating the binary profile for NC_{22} pattern was the varying length of the peptides. In order to tackle that situation, we have appended the dummy variable "X" in the sequences having length less than 22 as shown in Table 3.

Table 3: Generation of NC_{22} patterns from the original sequences with varying length

Original Sequences	Original Length	NC22
TQQKKADRY	**9**	**TQQKKADRYXXXXXXXXXXXXX**
ISAYLLSKNNAI	**12**	**ISAYLLSKNNAIXXXXXXXXXX**
GTFQKWAAVVVPSGE	**15**	**GTFQKWAAVVVPSGEXXXXXXX**
SAIEYTIENVFESAPNPR	**18**	**SAIEYTIENVFESAPNPRXXXX**
LPGDKSKAFDFLSEETEASLAS	**22**	**LPGDKSKAFDFLSEETEASLAS**

In Table 4, we have reported the length of the vector size generated by composition, and binary profile based features. As shown in the Table 4, feature NC_{22} generated highest number of features with vector size 462, whereas SOCN reports minimum number of features i.e. 2.

Table 4: Description of features calculated using Pfeature

Module	Type of Feature	Vector size
Composition	Amino acid composition	20
	Dipeptide composition	400
	Atomic composition	5
	Bond composition	4
	Physico-chemical properties based composition	30
	Residue repeat information	20
	Distance distribution of residues	20
	Shannon entropy for residues	20
	Shannon entropy based on physico-chemical properties	25
	Conjoint triad calculation	343
	Composition enhanced transition and distribution	189
	Pseudo amino acid composition	21
	Amphiphilic pseudo amino acid composition	23
	Quasi-sequence order	42
	Sequence order coupling number	2
Binary	N9	189
	C9	189
	N9C9	378
	NC22	462
	Combined	2382

3.5 Model Development

In order to train and develop prediction models, we have used various classifiers such as decision tree (DT), random forest (RF), logistic regression (LR), extreme gradient boosting (XGB), k-nearest neighbor (KNN), gaussian naïve Bayes (GNB), extremely randomized tree (ET), and support vector classifier (SVC) using scikit-learn (76) library of python. The description of each classifier is mentioned below in details.

DT is a rule-based supervised machine learning algorithm, in which the decision i.e. the assignment of a class is the end product of the set of rules. These set of rules are defined using the training set that is used to train the final model. This method leads to the development of a tree in which each node represents the dataset feature which is used to split the data, this process

continues till all the data points belong to a particular class get isolated. DT algorithm can be used to achieve classification as well as regression tasks.

RF is an ensemble-based approach which is also a supervised machine learning approach. As the name exhibits, RF contains a forest or a huge number of individual decision trees on various subsets of samples in the dataset, where each tree provides a particular output or class and by using the voting approach a single class would be predicted as the model class. Moreover, this meta classifier also applies mean based approach to enhance the accuracy of the model and avoid over-fitting.

LR is a statistical approach which implements the logistic function to model the probability of the binary/discrete output by using the independent variables. It is also a very powerful supervised machine learning algorithm which shares the resembles with the multiple linear regression with the exception that the response variable is a binomial.

XGB is also a tree-based approach which lies under the shadow of supervised machine learning techniques. It is an efficient, portable, and flexible gradient boosting algorithm. It implements iterative approach in which ensembles of decision trees are created where one tree is added at a time and fit to reduce the errors in the predictions resulted due to previous models. XGB provides the parallel tree boosting which make it a fast and accurate method. The difference between XGB and gradient boosting lies in the metric used to identify the best split for a tree.

KNN works on the ideology of the proximity and predicts the class of an unknown variable based on the closeness of its data points to the trained dataset. The learning process of this approach is occurrence-based, lazy and non-parametric. Instead of learning weights for features from the training dataset, it uses the entire dataset to make predictions for the unseen data.

GNB is a probabilistic approach based on the bayes theorem. It is assumed that the features involved in the training of a model, follows gaussian distribution, are independent from each other and makes an equal contribution to the prediction. The primary task of this algorithm is to create a prediction model that results in the sample probabilities to belong to a particular label.

ET is also an ensemble-based technique which considers the predictions from a huge number of de-correlated decision trees to make the overall prediction. It is quite similar to the RF classifier with the difference in the construction of the decision trees and selection of the threshold to split the split the nodes. Moreover, it is faster as compare to RF as it chooses threshold randomly then finding optimal cut point.

SVC is a supervised machine learning algorithm which finds the extreme data points that aids in the creation of hyperplanes, which further separates the n-dimensional space into different classes. Further, the generated model can be used to assign the classes to unseen data points. Support vector machines can be used for either classification as well as for regression tasks.

3.6 Cross-Validation

To avoid the overfitting and biasness of the generated model, we have implemented the five-fold cross-validation technique. Moreover, it also allows to assess the efficiency of the prediction models. Other advantages of cross-validation are highly accurate measures for out-of-sample accuracy and highly effective use of data. As per the standard norms, we have implemented the five-fold cross validation technique on training dataset and kept the validation dataset untouched. As depicted in Figure 5, in this technique the entire dataset is divided into five parts, where four parts are used to train the model and tested on the remaining fifth one. The same process is iterated five times in such a way that each set/part gets the chance to act as testing dataset. Finally, the overall performance is the mean of the performances of five iterations.

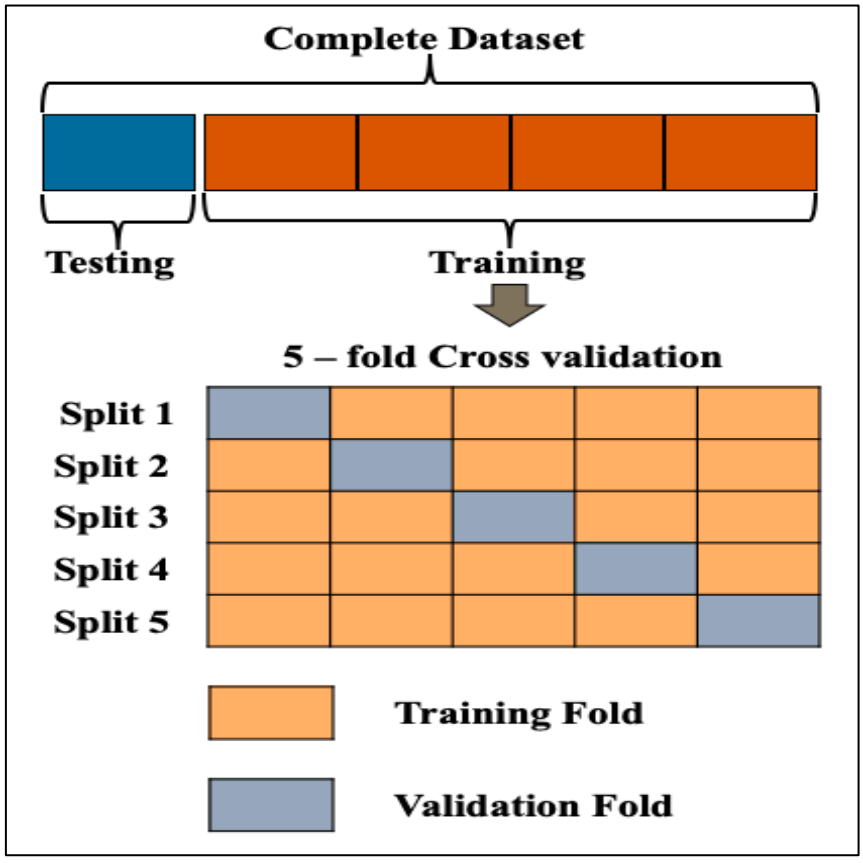


Figure 5: Graphical representation of five-fold cross validation

3.7 Evaluation of Parameters

In order to evaluate the efficiency of the different generated models developed using various classifiers, we have used well established evaluation parameters. In this method, we have used both threshold-dependent and -independent parameters. In threshold-dependent parameters we have used sensitivity which exhibits the percentage of correctly predicted binders, specificity defines the percentage of correctly predicted non-binders, accuracy denotes the percentage of correct prediction, F1-score sums up the predictive performance of the models, kappa measures the reliability between predicted and observed values, and Mathews correlation coefficient (MCC) represents the correlation between observed and predicted values, but these are the threshold dependent parameters which vary with threshold. On the other hand, area under the receiver operating characteristics curve (AUROC) is the measure of separability and it signifies how well the model is capable of distinguishing between the classes. Threshold dependent parameters were calculated using the following equations:

$$Sensitivity = \frac{TP}{TP + FN} \quad [2]$$

$$Specificity = \frac{TN}{TN + FP} \quad [3]$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad [4]$$

$$F1 - Score = \frac{2TP}{2TP + FP + FN} \quad [5]$$

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad [6]$$

$$K = \frac{(TP + TN + FP + FN)(TP + TN) - [(TP + FP)(TP + FN) + (FN + TN)(FP + TN)]}{(TP + TN + FP + FN)^2 - [(TP + FP)(TP + FN) + (FN + TN)(FP + TN)]} \quad [7]$$

Where, TP stands true positive; TN stands for true negative; FP stands for false positive; FN stands for false negative

3.8 Model Optimization

We evaluated eight different machine learning algorithms, and tuned the hyper parameters according to the training dataset. For this purpose, we used GridSearchCV to find the best performing model for each of our machine learning classifiers and optimised them by maximizing the AUROC.

3.9 Similarity Search

In order to predict if the query peptide is a binder of a HLA-DRB1*04:01 using similarity search, we have implemented the Basic Local Alignment Search Tool (BLAST) (77) using the NCBI-blast executable version 2.13.0. We have created the custom database using our dataset by implementing "makeblastdb" module of NCBI-blast. Then, to make the prediction for query sequences we have implemented the "blastp" module with "blastp-short" as task since the peptide length are small. Top-hit against the query sequences were considered to assign the classes.

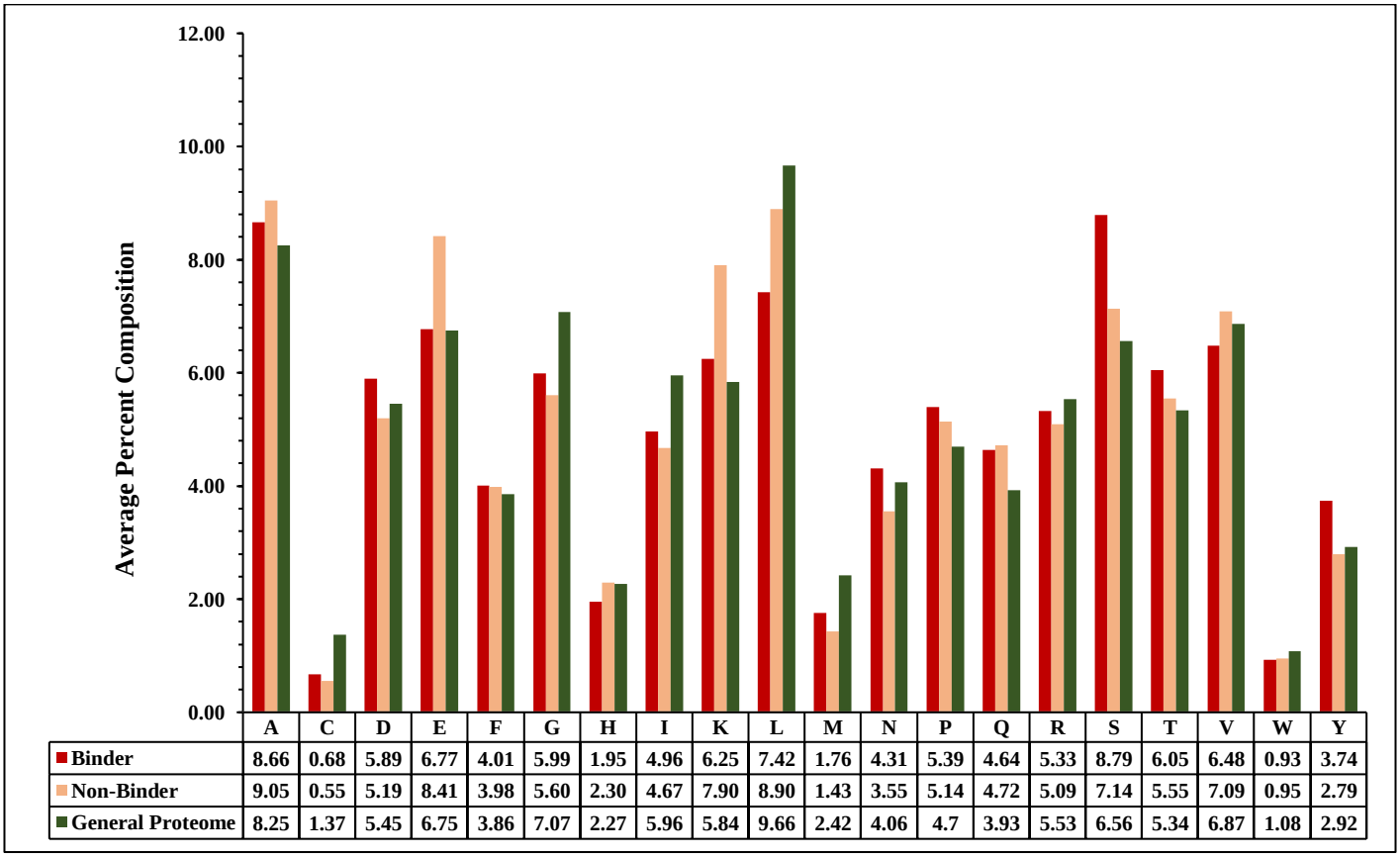
3.10 Motif Analysis

To make the predictions using the small regions which are shared by all the sequences of a particular class also called motifs, we have implemented the Motif – Emerging and with Classes - Identification (MERCI) tool (78) with default the parameters. We have identified the motifs which are specific to the HLA-DRB1*04:01 binders and used them to assign the class as binder to the query/unseen data if the particular motif is found else assigned them as non-binders.

Chapter 4
Results

4.1 Composition Analysis

In this study, we calculated the average composition of each residue in HLA-DRB1*04:01 binders and non-binders in balanced, alternate, and realistic dataset. The amino acid composition is calculated using Pfeature (75). The average residue composition for each dataset is provided in Figure 6, and it exhibits that serine residue is abundant in HLA-DRB1*04:01 binding peptides in comparison to the non-binding peptides. Moreover, the similarity in the trends of negative dataset generated randomly using Swiss-Prot (73) database and general proteome signifies that the negative dataset is not biased towards a particular amino acid or nature of amino acids.



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Binder	8.66	0.68	5.89	6.77	4.01	5.99	1.95	4.96	6.25	7.42	1.76	4.31	5.39	4.64	5.33	8.79	6.05	6.48	0.93	3.74
Non-Binder	9.05	0.55	5.19	8.41	3.98	5.60	2.30	4.67	7.90	8.90	1.43	3.55	5.14	4.72	5.09	7.14	5.55	7.09	0.95	2.79
General Proteome	8.25	1.37	5.45	6.75	3.86	7.07	2.27	5.96	5.84	9.66	2.42	4.06	4.7	3.93	5.53	6.56	5.34	6.87	1.08	2.92

Figure 6: Average percent amino acid composition of HLA-DRB1*04:01 binder, non-binders and general proteome

4.2Position Preference Analysis

In this study, preference of particular residues at a specific positions in a peptide was studied by creating the TSL for HLA-DRB1*04:01 binders and non-binders in the balanced, alternate, and realistic dataset as shown in Figure 7. The TSL for each dataset is of length 18, where first nine position represents nine residues for N-terminal and position 10-18 represents the nine residues from C-terminal.

In case of realistic dataset, positions 4, 5, and 6 are preferred by hydrophobic residues 'L/F/Y/I/V' ; where position 9 is covered by polar and uncharged amino acids 'S/T/A'; position 10 is preferred by positive charged amino acid residues 'K/R'; 'S/A' amino acids are favoured in positions 13-15; positions 16 and 17 are preferred by polar amino acids 'S/T', and 'D' residue is found to be most abundant at position 18 in HLA-DRB1*04:01 binding peptides. On the other hand, in case of HLA-DRB1*04:01 non-binding peptides, 'P' is preferred at position 2; positions 4, and 5, are most preferred by positive amino acid 'K' ; position 10 also preferred by positive charged residues 'K/R'; and positions 14-19 showed abundance for residues 'A/L'.

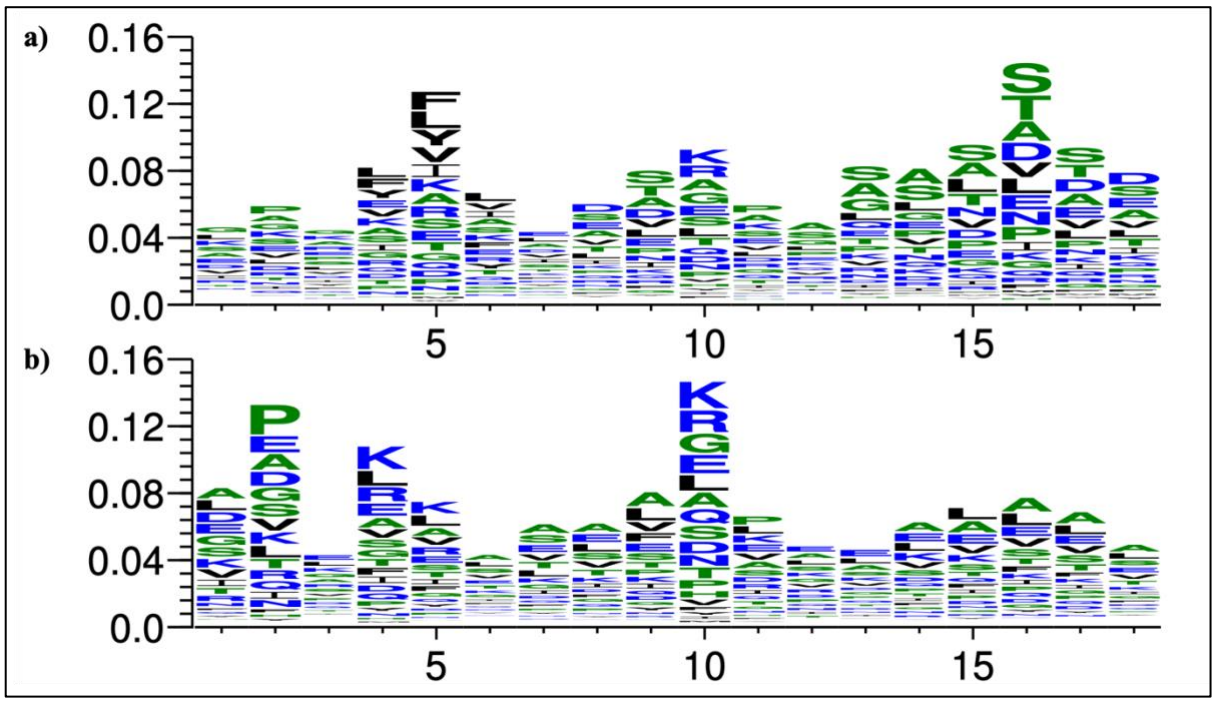


Figure 7: Positional preference representation using weblogo in a) HLA-DRB1*04:01 binders, b) HLA-DRB1*04:01 non-binders

4.3 Performance of models on composition based module

We have calculated fifteen different types of features using composition based module of Pfeature (75) and used them to develop the prediction models using eight different classifiers from sklearn (76) library of python. The models were developed by implementing classifiers like DT, RF, LR, KNN, XGB, GNB, ET, and SVC. The models were trained on the training dataset and external validated on the testing dataset of balanced, alternate, and realistic dataset. Table 5 exhibits the performance of best performing model developed on training and validation dataset using different types of features is reported in terms of sensitivity, specificity, accuracy, AUROC, F1-score, kappa , and MCC.

As shown in Table 5, extra-tree classifier based model developed on DPC features outperformed all the other models developed on other features, with AUROC of 0.92 on training and validation data of balanced dataset; 0.90 AUROC on training and validation data of alternate dataset; and AUROC of 0.96 on training and validation data of realistic dataset. CTC based model performed second best with AUROC of 0.90 on training and validation dataset respectively, for balanced dataset; alternated dataset was able to achieve AUROC of 0.87 on training and validation dataset; and realistic dataset attains AUROC of 0.94 and 0.93 on training and validation dataset, respectively.

Table 5: Performance measures for best performing model developed using fifteen different types composition based features calculated using Pfeature for balanced, alternate, and realistic dataset

Features	Dataset	Balanced Dataset							Alternate Dataset							Realistic Dataset						
		Sens	Spec	Acc	AUC	F1	K	MCC	Sens	Spec	Acc	AUC	F1	K	MCC	Sens	Spec	Acc	AUC	F1	K	MCC
AAC	**Train**	79.43	77.91	78.67	0.88	0.79	0.57	0.57	77.37	76.65	77.01	0.86	0.77	0.54	0.54	83.53	83.98	83.93	0.92	0.57	0.48	0.52
	Test	81.07	75.36	78.21	0.88	0.79	0.56	0.57	77.04	76.77	76.91	0.85	0.77	0.54	0.54	84.26	84.12	84.14	0.92	0.58	0.49	0.53
DPC	**Train**	**83.18**	**83.68**	**83.43**	**0.92**	**0.83**	**0.67**	**0.67**	**81.91**	**82.18**	**82.04**	**0.90**	**0.82**	**0.64**	**0.64**	**88.71**	**89.30**	**89.22**	**0.96**	**0.68**	**0.62**	**0.64**
	Test	**83.75**	**82.02**	**82.88**	**0.92**	**0.83**	**0.66**	**0.66**	**81.14**	**82.89**	**82.02**	**0.90**	**0.82**	**0.64**	**0.64**	**89.19**	**89.50**	**89.46**	**0.96**	**0.68**	**0.63**	**0.65**
ATC	**Train**	56.78	55.61	56.20	0.59	0.57	0.12	0.12	59.13	61.10	60.11	0.64	0.60	0.20	0.20	55.68	60.64	60.00	0.62	0.26	0.08	0.11
	Test	55.98	54.26	55.12	0.58	0.56	0.10	0.10	58.54	63.29	60.92	0.64	0.60	0.22	0.22	58.42	59.91	59.72	0.62	0.27	0.09	0.12
BTC	**Train**	54.43	54.34	54.39	0.56	0.54	0.09	0.09	55.43	57.17	56.30	0.59	0.56	0.13	0.13	57.43	54.90	55.23	0.59	0.25	0.06	0.08
	Test	55.58	53.98	54.78	0.56	0.55	0.10	0.10	55.35	56.31	55.83	0.58	0.56	0.12	0.12	59.72	54.38	55.07	0.60	0.25	0.07	0.09
PCP	**Train**	71.75	72.85	72.30	0.80	0.72	0.45	0.45	74.12	72.46	73.29	0.81	0.74	0.47	0.47	76.11	77.00	76.89	0.85	0.46	0.34	0.39
	Test	73.22	70.86	72.04	0.80	0.72	0.44	0.44	71.95	71.06	71.51	0.80	0.72	0.43	0.43	77.79	77.16	77.24	0.85	0.47	0.35	0.40
RRI	**Train**	77.41	77.31	77.36	0.86	0.77	0.55	0.55	75.41	74.76	75.09	0.83	0.75	0.50	0.50	81.60	80.64	80.77	0.89	0.52	0.42	0.47
	Test	77.99	75.24	76.61	0.86	0.77	0.53	0.53	75.19	73.07	74.13	0.83	0.74	0.48	0.48	81.50	80.66	80.77	0.89	0.52	0.42	0.47
DDOR	**Train**	79.14	77.55	78.35	0.88	0.79	0.57	0.57	75.36	76.41	75.88	0.85	0.76	0.52	0.52	84.01	81.95	82.21	0.91	0.55	0.45	0.50
	Test	79.88	75.00	77.44	0.88	0.78	0.55	0.55	75.94	76.26	76.10	0.84	0.76	0.52	0.52	83.24	82.25	82.38	0.92	0.55	0.45	0.50
SER	**Train**	78.01	79.58	78.79	0.88	0.79	0.58	0.58	77.12	77.54	77.33	0.86	0.77	0.55	0.55	83.09	84.74	84.53	0.92	0.58	0.50	0.53

	Test	79.76	77.13	78.45	0.87	0.79	0.57	0.57	77.48	76.26	76.87	0.85	0.77	0.54	0.54	84.77	84.00	84.10	0.92	0.58	0.49	0.53
SEP	**Train**	70.41	72.08	71.24	0.78	0.71	0.43	0.43	72.18	73.25	72.71	0.80	0.73	0.45	0.45	73.66	76.19	75.87	0.83	0.44	0.32	0.36
	Test	70.26	70.47	70.36	0.78	0.70	0.41	0.41	70.14	72.60	71.37	0.79	0.71	0.43	0.43	74.40	76.29	76.05	0.84	0.44	0.32	0.37
CTC	**Train**	80.99	80.94	80.96	0.90	0.81	0.62	0.62	78.67	79.13	78.90	0.87	0.79	0.58	0.58	86.45	87.65	87.50	0.94	0.64	0.57	0.60
	Test	80.87	78.43	79.65	0.90	0.80	0.59	0.59	77.36	78.23	77.80	0.87	0.78	0.56	0.56	85.60	87.88	87.59	0.93	0.64	0.57	0.60
CeTD	**Train**	75.94	77.84	76.89	0.85	0.77	0.54	0.54	73.97	75.09	74.53	0.83	0.74	0.49	0.49	82.54	80.24	80.53	0.90	0.52	0.42	0.47
	Test	75.27	75.91	75.59	0.85	0.76	0.51	0.51	72.62	74.76	73.69	0.82	0.73	0.47	0.47	82.49	80.47	80.73	0.90	0.52	0.42	0.47
PAAC	**Train**	79.50	77.93	78.71	0.88	0.79	0.57	0.57	77.31	77.32	77.31	0.86	0.77	0.55	0.55	84.03	83.98	83.98	0.92	0.57	0.49	0.53
	Test	80.36	75.91	78.13	0.88	0.79	0.56	0.56	76.69	77.45	77.07	0.85	0.77	0.54	0.54	84.22	85.05	84.95	0.92	0.59	0.51	0.54
APAAC	**Train**	79.71	78.86	79.28	0.88	0.79	0.59	0.59	77.04	78.44	77.74	0.86	0.78	0.56	0.56	84.37	84.59	84.56	0.92	0.58	0.50	0.54
	Test	80.16	76.22	78.19	0.88	0.79	0.56	0.56	77.12	78.63	77.87	0.86	0.78	0.56	0.56	84.30	85.49	85.34	0.93	0.60	0.52	0.55
QSO	**Train**	78.28	78.83	78.55	0.88	0.79	0.57	0.57	76.85	76.19	76.52	0.86	0.77	0.53	0.53	84.23	83.28	83.40	0.92	0.57	0.48	0.52
	Test	79.92	76.66	78.29	0.88	0.79	0.57	0.57	77.12	74.69	75.90	0.85	0.76	0.52	0.52	83.08	83.83	83.73	0.92	0.57	0.48	0.52
SOCN	**Train**	50.18	54.79	52.49	0.54	0.51	0.05	0.05	55.94	51.37	53.66	0.55	0.55	0.07	0.07	58.43	46.56	48.08	0.54	0.22	0.02	0.03
	Test	52.43	50.63	51.53	0.52	0.52	0.03	0.03	52.98	52.88	52.93	0.55	0.53	0.06	0.06	52.07	50.95	51.09	0.52	0.21	0.01	0.02

***Sens: Sensitivity; Spec: Specificity; Acc: Accuracy; AUC: Area under the receiver operating characteristic curve; F1: F1 score; MCC: Matthews Correlation Coefficient; K: Cohen's Kappa**

4.4 Performance of models on binary profile based module

Similarly, we have generated the binary profiles for different patterns such as N_9, C_9, N_9C_9, and NC_{22}, to develop the prediction models with the ability to classify HLA-DRB1*04:01 binders. Table 6 represents the performance of best models developed by implementing classifiers for each pattern type. As shown in the Table 6, extra-tree classifier based model developed using pattern NC_{22} outperformed the other patterns with AUROC of 0.90 on training and validation dataset for balanced dataset, 0.87 on training and validation dataset for alternate dataset, and 0.94 on training and validation dataset for realistic dataset. Followed by models developed on pattern N_9C_9 , for balanced dataset it attains the maximum AUROC of 0.87 and 0.86 on training and validation dataset, 0.85 for alternate dataset, and 0.90 AUROC for the training and validation dataset of realistic dataset. Whereas, models developed on C_9 feature performed slightly better than models developed on N_9 feature, with AUROC of 0.86, 0.84, and 0.90 on the training and validation dataset of balanced, alternate, and realistic dataset. Finally, models developed on N_9 are the least performing with equal AUROC of 0.86 on training and validation dataset of balanced dataset, equal AUROC of 0.83 on training and validation dataset of alternate dataset, and equal AUROC of 0.90 on training and validation dataset of realistic dataset.

Table 6: Performance measures for best performing model developed using four different types binary profile based features calculated using Pfeature for balanced, alternate, and realistic dataset

Features	Dataset	Balanced Dataset							Alternate Dataset							Realistic Dataset						
		Sens	**Spec**	**Acc**	**AUC**	**F1**	**K**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**F1**	**K**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**F1**	**K**	**MCC**
N_9	Train	76.91	76.74	76.82	0.86	0.77	0.54	0.54	74.92	74.82	74.87	0.83	0.75	0.50	0.50	82.63	81.20	81.38	0.90	0.53	0.43	0.48
	Test	77.99	76.46	77.22	0.86	0.77	0.54	0.55	75.15	73.98	74.56	0.83	0.75	0.49	0.49	83.16	81.31	81.55	0.90	0.54	0.44	0.49
C_9	Train	76.81	77.22	77.01	0.86	0.77	0.54	0.54	75.75	75.70	75.73	0.84	0.76	0.52	0.52	81.78	80.19	80.39	0.90	0.52	0.41	0.46
	Test	77.08	75.79	76.44	0.86	0.77	0.53	0.53	75.62	74.57	75.09	0.84	0.75	0.50	0.50	82.05	80.86	81.01	0.90	0.53	0.43	0.47
N_9C_9	Train	77.79	79.50	78.65	0.87	0.79	0.57	0.57	76.98	76.75	76.86	0.85	0.77	0.54	0.54	81.05	82.82	82.59	0.90	0.54	0.45	0.49
	Test	77.12	78.43	77.78	0.86	0.78	0.56	0.56	78.15	77.21	77.68	0.85	0.78	0.55	0.55	79.68	84.38	83.78	0.90	0.56	0.47	0.50
NC_{22}	Train	**82.11**	**81.48**	**81.80**	**0.90**	**0.82**	**0.64**	**0.64**	**78.51**	**78.56**	**78.54**	**0.87**	**0.79**	**0.57**	**0.57**	**86.38**	**86.96**	**86.88**	**0.94**	**0.63**	**0.56**	**0.59**
	Test	**82.09**	**80.88**	**81.48**	**0.90**	**0.82**	**0.63**	**0.63**	**78.58**	**76.07**	**77.32**	**0.87**	**0.78**	**0.55**	**0.55**	**86.11**	**87.52**	**87.34**	**0.94**	**0.64**	**0.57**	**0.60**

***Sens: Sensitivity; Spec: Specificity; Acc: Accuracy; AUC: Area under the receiver operating characteristic curve; F1: F1 score; MCC: Matthews Correlation Coefficient; K: Cohen's Kappa**

4.5 Performance of models on combined features

Further, we have combined all the features to develop the vector of size 2832 for each peptide belong to different datasets and develop the prediction models using eight different classifiers by hyper-tuning the parameters to maximize the AUROC on the training dataset and validated on the testing dataset. Table 7 comprises the threshold-dependent and threshold-independent performance measure of all the classifiers trained and tested on balanced, alternate, and realistic dataset. ET-based model developed using combined features outperformed all the other classifiers by attaining the maximum AUROC of 0.91 and 0.90 on training and validation dataset of balanced dataset, 0.88 on both the training and validation dataset of alternate dataset, and AUROC of 0.94 and 0.95 on the training and validation dataset of realistic dataset. Similarly, RF-based model also performed better in comparison to the other classifiers, with AUROC of greater than 0.92 on realistic dataset, >0.87 on balanced dataset, and AUROC >0.86 on alternate dataset; followed by XGB based model with AUROC >0.89 on realistic, >0.86 on balanced, and >0.84 on alternate dataset.

Table 7: Performance measures for all model developed using all classifiers on combined features for balanced, alternate, and realistic dataset

Classifier	Dataset	Balanced Dataset							Alternate Dataset							Realistic Dataset						
		Sens	Spec	Acc	AUC	F1	Kappa	MCC	Sens	Spec	Acc	AUC	F1	Kappa	MCC	Sens	Spec	Acc	AUC	F1	Kappa	MCC
DT	**Train**	70.23	65.87	68.05	0.68	0.69	0.36	0.36	69.01	64.65	66.83	0.67	0.68	0.34	0.34	46.49	92.17	86.32	0.69	0.47	0.39	0.39
	Test	69.59	65.66	67.62	0.68	0.68	0.35	0.35	68.80	64.35	66.58	0.67	0.67	0.33	0.33	46.08	91.72	85.87	0.69	0.46	0.37	0.37
RF	**Train**	79.94	78.85	79.40	0.88	0.80	0.59	0.59	76.91	76.62	76.76	0.85	0.77	0.54	0.54	84.97	82.98	83.24	0.92	0.57	0.48	0.52
	Test	80.55	77.17	78.86	0.87	0.79	0.58	0.58	76.49	77.21	76.85	0.85	0.77	0.54	0.54	85.13	82.77	83.07	0.92	0.56	0.47	0.52
LR	**Train**	62.29	62.69	62.49	0.67	0.62	0.25	0.25	66.79	65.62	66.21	0.71	0.66	0.32	0.32	63.85	62.52	62.69	0.68	0.31	0.14	0.18
	Test	63.35	59.19	61.27	0.66	0.62	0.23	0.23	64.30	65.85	65.08	0.70	0.65	0.30	0.30	64.77	61.76	62.14	0.68	0.31	0.14	0.18
XGB	**Train**	77.54	77.75	77.64	0.86	0.78	0.55	0.55	76.86	76.71	76.78	0.85	0.77	0.54	0.54	82.12	80.81	80.98	0.90	0.53	0.43	0.47
	Test	79.01	78.00	78.51	0.86	0.79	0.57	0.57	76.29	76.22	76.26	0.84	0.76	0.53	0.53	80.99	80.29	80.38	0.89	0.51	0.41	0.46
KNN	**Train**	57.94	54.99	56.47	0.59	0.57	0.13	0.13	64.10	52.38	58.24	0.61	0.61	0.17	0.17	62.92	56.57	57.38	0.62	0.27	0.09	0.13
	Test	56.65	53.55	55.10	0.57	0.56	0.10	0.10	62.92	52.29	57.60	0.61	0.60	0.15	0.15	63.16	56.34	57.21	0.62	0.27	0.09	0.13
GNB	**Train**	63.41	63.33	63.37	0.68	0.63	0.27	0.27	63.09	71.52	67.30	0.70	0.66	0.35	0.35	64.52	64.58	64.57	0.69	0.32	0.16	0.20
	Test	65.09	61.87	63.48	0.68	0.64	0.27	0.27	62.68	71.77	67.23	0.70	0.66	0.34	0.35	63.35	64.49	64.34	0.69	0.31	0.15	0.19
ET	**Train**	**82.18**	**81.95**	**82.07**	**0.91**	**0.82**	**0.64**	**0.64**	**78.80**	**80.05**	**79.42**	**0.88**	**0.79**	**0.59**	**0.59**	**86.68**	**87.28**	**87.20**	**0.94**	**0.63**	**0.56**	**0.60**
	Test	**82.01**	**80.56**	**81.29**	**0.90**	**0.81**	**0.63**	**0.63**	**78.58**	**80.24**	**79.41**	**0.88**	**0.79**	**0.59**	**0.59**	**86.43**	**87.62**	**87.47**	**0.95**	**0.64**	**0.57**	**0.60**
SVC	**Train**	57.46	47.80	52.63	0.54	0.55	0.05	0.05	55.55	53.13	54.34	0.57	0.55	0.09	0.09	56.44	47.20	48.38	0.54	0.22	0.02	0.02
	Test	56.61	47.95	52.28	0.54	0.54	0.05	0.05	55.70	55.05	55.37	0.58	0.56	0.11	0.11	53.61	51.91	52.13	0.54	0.22	0.03	0.04

*** Sens: Sensitivity; Spec: Specificity; Acc: Accuracy; AUC: Area under the receiver operating characteristic curve; F1: F1 score; MCC: Matthews Correlation Coefficient; K: Cohen's Kappa**

4.6 Performance of models on selected features

In the machine learning based prediction methods selecting the appropriate number of features is the key step in training a model, as it reduces the time for training exponentially and moreover avoid the curse of over-fitting and large dimensions of features. Hence, it is a very crucial step. There are ample of feature selection method exists in the literature. We have implemented the SVC-L1 based feature selection technique. SVC-L1 stands for support vector classifier with linear kernel and L1 penalty regularization. The reason of using this method is its algorithm in which it applies various methods to select the best performing features, moreover, it's processing is quite faster than other methods. It minimizes the objective function which is the result of the loss function and regularization.

Based on this technique, we left with 125 features in case of balanced dataset, 131 features for alternate, and 258 features were selected in realistic dataset. Table 8 contains the performance measures such as sensitivity, specificity, accuracy, F1, kappa, MCC and AUROC for the models developed using eight classifiers for each dataset. Among all the classifiers, ET-based model has outperformed in each dataset with AUROC 0.90 on training as well as on validation dataset of the balanced dataset, AUROC 0f 0.88 and 0.87 on training and validation of alternate dataset, and AUROC of 0.95 on both training and validation dataset of the realistic dataset. RF-based model also performed equivalently with AUROC of 0.87 in balanced dataset, >0.84 in alternate dataset, and 0.92 in the realistic dataset. Whereas, models developed using classifiers DT, LR, KNN, GNB, and SVC performed poorly in terms of all the considered parameters.

Table 8: Performance of various classifiers after reducing the features using SVC-L1

Classifier	Dataset	Balanced Dataset							Alternate Dataset							Realistic Dataset						
		Sens	Spec	Acc	AUC	F1	Kappa	MCC	Sens	Spec	Acc	AUC	F1	Kappa	MCC	Sens	Spec	Acc	AUC	F1	Kappa	MCC
DT	**Train**	68.40	65.51	66.95	0.67	0.67	0.34	0.34	67.04	62.81	64.92	0.65	0.66	0.30	0.30	45.85	91.91	86.01	0.69	0.46	0.38	0.38
	Test	68.21	64.71	66.46	0.67	0.67	0.33	0.33	66.51	63.49	65.00	0.65	0.66	0.30	0.30	45.84	92.02	86.11	0.69	0.46	0.38	0.38
RF	**Train**	79.49	77.98	78.73	0.87	0.79	0.58	0.58	76.29	75.56	75.93	0.85	0.76	0.52	0.52	85.41	83.31	83.58	0.92	0.57	0.48	0.53
	Test	79.76	76.77	78.27	0.87	0.79	0.57	0.57	76.25	75.91	76.08	0.84	0.76	0.52	0.52	85.37	83.54	83.77	0.92	0.57	0.49	0.53
LR	**Train**	63.53	63.94	63.74	0.69	0.64	0.28	0.28	64.83	66.63	65.73	0.71	0.65	0.32	0.32	67.43	67.89	67.83	0.74	0.35	0.20	0.25
	Test	62.72	61.95	62.34	0.67	0.63	0.25	0.25	63.83	66.84	65.33	0.71	0.65	0.31	0.31	65.44	67.43	67.18	0.73	0.34	0.18	0.23
XGB	**Train**	74.79	75.65	75.22	0.83	0.75	0.50	0.50	74.42	74.69	74.56	0.82	0.75	0.49	0.49	79.93	79.97	79.96	0.88	0.51	0.40	0.45
	Test	74.28	74.61	74.44	0.82	0.74	0.49	0.49	73.96	74.84	74.40	0.82	0.74	0.49	0.49	77.91	79.54	79.33	0.87	0.49	0.38	0.43
KNN	**Train**	56.53	55.63	56.08	0.58	0.56	0.12	0.12	63.06	51.78	57.42	0.60	0.60	0.15	0.15	80.67	70.60	71.89	0.82	0.42	0.29	0.36
	Test	55.66	54.97	55.32	0.57	0.56	0.11	0.11	62.64	53.98	58.31	0.61	0.60	0.17	0.17	80.51	70.36	71.66	0.81	0.42	0.29	0.35
GNB	**Train**	64.43	64.29	64.36	0.70	0.64	0.29	0.29	67.14	67.08	67.11	0.72	0.67	0.34	0.34	65.62	65.52	65.53	0.71	0.33	0.17	0.21
	Test	64.93	62.19	63.56	0.68	0.64	0.27	0.27	65.01	66.88	65.94	0.71	0.66	0.32	0.32	65.52	64.88	64.96	0.70	0.32	0.16	0.21
ET	**Train**	**81.75**	**80.51**	**81.13**	**0.90**	**0.81**	**0.62**	**0.62**	**78.53**	**79.43**	**78.98**	**0.88**	**0.79**	**0.58**	**0.58**	**87.24**	**87.37**	**87.36**	**0.95**	**0.64**	**0.57**	**0.60**
	Test	**82.17**	**78.55**	**80.36**	**0.90**	**0.81**	**0.61**	**0.61**	**77.95**	**78.75**	**78.35**	**0.87**	**0.78**	**0.57**	**0.57**	**87.57**	**88.11**	**88.04**	**0.95**	**0.65**	**0.59**	**0.62**
SVC	**Train**	57.37	47.16	52.27	0.54	0.55	0.05	0.05	54.91	52.99	53.95	0.56	0.54	0.08	0.08	73.21	74.34	74.19	0.82	0.42	0.29	0.34
	Test	56.37	47.71	52.04	0.53	0.54	0.04	0.04	55.62	54.81	55.22	0.57	0.55	0.10	0.10	72.27	75.69	75.25	0.82	0.43	0.30	0.35

*** Sens: Sensitivity; Spec: Specificity; Acc: Accuracy; AUC: Area under the receiver operating characteristic curve; F1: F1 score; MCC: Matthews Correlation Coefficient; K: Cohen's Kappa**

4.7 Performance of hybrid model

On observing the results of various machine learning classifiers on different type of features, it was found that models developed on realistic dataset has surpassed the performance of other generated dataset. NC_{22} feature based ET classifier model performed with the AUROC of 0.94 on both training and validation dataset; the AUROC increased to 0.95 on validation dataset when all the features were combined; on selecting the relevant features using SVC-L1 the AUROC on training dataset reaches 0.95. On observing all the results, it is found that ET-based model developed on DPC features outperformed all the other features with AUROC of 0.96 on training as well as on validation dataset. Hence, we combined ET-based model of DPC with similarity search using BLAST (77), and called it as hybrid model, in order to improve the performance.

We have implemented BLAST with varying e-value in order to find the optimal value at which we can achieve the maximum AUROC. Table 9 captures the results for each dataset at different e-values for training as well as validation dataset. We varied the e-value from 1.00e-06 to 1.00e+02 and attained the maximum AUROC of 0.98 and 0.99 on training and validation dataset, respectively at e-value 1.00e+00 on the realistic dataset, followed by AUROC of 0.93 on balanced dataset, and alternate dataset is able to achieve AUROC>0.92 in training as well as validation dataset. There is significant amount of improvement in all the performance measure after combining machine learning model with similarity search. The same model has been implemented at the backend of the server and respective standalone packages.

Table 9: Performance of hybrid model at different e-values on training and testing dataset

E-value	Dataset	Balanced Dataset							Alternate Dataset							Realistic Dataset						
		Sens	Spec	Acc	AUC	F1	K	MCC	Sens	Spec	Acc	AUC	F1	K	MCC	Sens	Spec	Acc	AUC	F1	K	MCC
1.00E-06	Train	83.41	82.56	82.99	0.92	0.83	0.66	0.66	81.53	82.85	82.19	0.90	0.82	0.64	0.64	88.47	89.07	88.99	0.95	0.67	0.61	0.64
	Test	85.05	81.06	83.06	0.92	0.83	0.66	0.66	82.92	80.67	81.79	0.90	0.82	0.64	0.64	88.32	89.77	89.58	0.95	0.68	0.63	0.65
1.00E-05	Train	83.34	82.68	83.01	0.92	0.83	0.66	0.66	81.63	82.87	82.25	0.90	0.82	0.65	0.65	88.94	88.27	88.35	0.95	0.66	0.60	0.63
	Test	85.13	81.41	83.27	0.92	0.84	0.67	0.67	83.35	80.67	82.01	0.91	0.82	0.64	0.64	89.56	89.33	89.36	0.95	0.68	0.62	0.65
1.00E-04	Train	83.46	82.76	83.11	0.92	0.83	0.66	0.66	81.86	82.89	82.38	0.90	0.82	0.65	0.65	88.21	88.95	88.86	0.95	0.67	0.61	0.63
	Test	85.09	82.03	83.56	0.93	0.84	0.67	0.67	84.01	80.71	82.36	0.91	0.83	0.65	0.65	88.90	90.32	90.14	0.95	0.70	0.64	0.66
1.00E-03	Train	83.85	82.87	83.36	0.92	0.83	0.67	0.67	82.42	82.95	82.68	0.91	0.83	0.65	0.65	88.89	88.60	88.64	0.95	0.67	0.60	0.63
	Test	85.40	81.99	83.70	0.93	0.84	0.67	0.67	84.47	80.86	82.67	0.91	0.83	0.65	0.65	89.33	90.25	90.13	0.96	0.70	0.64	0.67
1.00E-02	Train	84.34	83.05	83.69	0.93	0.84	0.67	0.67	83.46	82.97	83.22	0.91	0.83	0.66	0.66	89.33	88.46	88.57	0.96	0.67	0.60	0.63
	Test	86.18	82.45	84.32	0.93	0.85	0.69	0.69	85.68	80.75	83.21	0.92	0.84	0.66	0.67	89.67	90.24	90.17	0.96	0.70	0.64	0.67
1.00E-01	Train	84.28	84.50	84.39	0.93	0.84	0.69	0.69	84.36	84.79	84.57	0.92	0.85	0.69	0.69	89.31	88.26	88.40	0.96	0.66	0.60	0.63
	Test	86.37	83.50	84.94	0.94	0.85	0.70	0.70	87.03	83.00	85.02	0.93	0.85	0.70	0.70	89.64	90.21	90.14	0.96	0.70	0.64	0.67
1.00E+00	Train	85.01	84.35	84.68	0.93	0.85	0.69	0.69	85.52	84.02	84.77	0.92	0.85	0.70	0.70	89.31	88.26	88.40	0.98	0.66	0.60	0.63
	Test	87.19	83.77	85.48	0.94	0.86	0.71	0.71	88.32	82.30	85.31	0.93	0.86	0.71	0.71	89.64	90.21	90.14	0.99	0.70	0.64	0.67
1.00E+01	Train	84.81	84.74	84.78	0.93	0.85	0.70	0.70	85.05	84.60	84.82	0.92	0.85	0.70	0.70	88.57	88.42	88.44	0.96	0.66	0.60	0.63
	Test	86.65	85.05	85.85	0.94	0.86	0.72	0.72	87.62	84.67	86.14	0.93	0.86	0.72	0.72	90.02	89.33	89.42	0.96	0.69	0.63	0.65
1.00E+02	Train	83.95	85.70	84.82	0.93	0.85	0.70	0.70	84.33	84.20	84.26	0.92	0.84	0.69	0.69	89.82	89.56	89.59	0.96	0.69	0.63	0.66
	Test	85.37	85.56	85.46	0.94	0.85	0.71	0.71	87.15	82.84	85.00	0.93	0.85	0.70	0.70	90.06	89.80	89.83	0.96	0.69	0.64	0.66

*** Sens: Sensitivity; Spec: Specificity; Acc: Accuracy; AUC: Area under the receiver operating characteristic curve; F1: F1 score; MCC: Matthews Correlation Coefficient; K: Cohen's kappa**

4.8 Motif analysis

In this study, we have implemented MERCI software with default parameters to obtain the specific regions i.e. motifs from the realistic dataset which are highly specific to HLA-DRB1*04:01 binders but absent in non-binder, similar procedure was repeated for non-binders where we searched for non-binder specific motifs which are exclusively present in the binders and absent in the binder sequences. In Table 10, we have reported motifs specific to binders and non-binders along with their coverage in the positive and negative dataset. Residue T, V, F, P, Q, and T are dominant in binders, where residue D, Y, and K covers the most of the motifs.

Table 10: Exclusive motifs specific to HLA-DRB1*04:01 binder and non-binders

HLA-DRB1*04:01 Binders			HLA-DRB1*04:01 Non-binders		
Motif	**# Sequences**	**Coverage**	**Motif**	**# Sequences**	**Coverage**
A-F-V-K-D	56	56	Y-D-G-K-D	335	335
V-A-F-V-K-D	53	109	Y-D-G-K-D-Y	315	650
D-V-A-F-V-K-D	49	158	A-Y-D-G-K-D	293	943
F-T-P-E-T	46	204	R-K-W-E-A	276	1219
F-T-P-E-T-N	46	250	A-Y-D-G-K-D-Y	274	1493
F-T-P-E-T-N-P	46	296	R-K-W-E-A-A	251	1774
T-P-E-T-N	46	342	S-D-H-E	249	1993
T-P-E-T-N-P	46	388	Y-D-G-K-D-Y-I	243	2236
D-Q-T-V-I	45	433	Y-D-G-K-D-Y-I-A	236	2472
D-Q-T-V-I-Q	45	478	Y-D-G-K-D-Y-I-A-L	236	2708
F-V-K-D-Q	45	523			
F-V-K-D-Q-T	45	568			
F-V-K-D-Q-T-V	45	613			
F-V-K-D-Q-T-V-I	45	658			
F-V-K-D-Q-T-V-I-Q	45	703			
I-F-T-P-E	45	748			
I-F-T-P-E-T	45	793			
I-F-T-P-E-T-N	45	838			
I-F-T-P-E-T-N-P	45	883			
K-D-Q-T-V-I	45	928			
K-D-Q-T-V-I-Q	45	973			
S-I-F-T-P	45	1018			
S-I-F-T-P-E	45	1063			
S-I-F-T-P-E-T	45	1108			
S-I-F-T-P-E-T-N	45	1153			
S-I-F-T-P-E-T-N-P	45	1198			

V-K-D-Q-T	45	1243			
V-K-D-Q-T-V	45	1288			
V-K-D-Q-T-V-I	45	1333			
V-K-D-Q-T-V-I-Q	45	1378			

4.9 Comparison with the existing methods

In order to understand the pros and cons of a newly developed method, it is of utter importance to compare its performance with the existing methods. Since, HLADR4Pred2 is an update of HLADR4Pred (37), hence its comprehensive comparison is required to understand the advantages of the newer version over older versions. Table 11 accumulates differences in the HLADR4Pred and HLADR4Pred2 at the level of dataset, implemented features, prediction approach, webserver and standalone. In the newer version, we have used the dataset with varying length i.e. 9-22 amino acids, whereas the older version was developed on the peptides having length of 9 residues. In terms of dataset size, HLADR4Pred was developed using 567 HLA-DRB1*04:01 binding peptides, on the other hand, HLA-DR4Pred2.0 is developed using 12676 HLA-DRB1*04:01 binders i.e. 22 times more data was used in the newer version. Moreover, older version was developed by using binary profile as the input feature where we have used dipeptide composition. We have developed the hybrid model by combining machine learning and similarity approach using BLAST, but older version was developed using machine learning algorithms only, i.e. support vector machine (SVM) and artificial neural network (ANN). We have also provided the option make the prediction on similarity search only, where an uncharacterized peptide can be assigned as HLA-DRB1*04:01 binder if a hit in found in our customized database else will be assigned as non-binder if no hit is found.

While considering the webserver, older version is not compatible with the smart devices of today's world while the new version is compatible with all the modern devices. Other than that we extracted the motifs using MERCI software which are specific to the HLA-DRB1*04:01 binders and used them for making prediction for the unseen data. We have increased the services to the community too, as both the versions have the predict module but along with that in the newer version we have also provided the facilities like scanning of the proteins to search binding regions, designing of binders, prediction using BLAST and motif search. Moreover, we have given the Perl and python-based standalone which can be used in the absence of internet or for the bulk dataset that may take longer time on the webserver. We have compiled the comparison between the older and newer version of HLADR4Pred in Table

11. In a nutshell, HLADR4Pred2 has number of novel features in terms of facilities as well as algorithm.

Table 11: Extensive comparison between HLADR4Pred and HLADR4Pred2

S.No.	HLADR4Pred	HLADR4Pred2
Dataset		
1	567 HLA-DRB1*04:01 binders and 567 non-binders	12676 HLA-DRB1*04:01 binders and 86300 non-binders
2	Peptides with length 9 amino acids	Peptides with length 9-22 amino acids
Features		
3	Binary profile	Dipeptide composition
4	No similarity search was performed	Similarity search was performed
Algorithm		
5	SVM and ANN based model	Extra-tree classifier was implemented
6	Only ML based model was implemented	Hybrid model with ML + BLAST was implemented
Webserver		
7	Non-responsive template	Responsive template
8	Not compatible with modern devices	Compatible with all modern day devices
9	No facility of scanning or designing	Options of scanning and designing are provided
10	No similarity search option	BLAST search against database made up of HLA-DRB1*04:01 binding peptides
11	No motif search was performed	Facility with motif scan based prediction is available
Standalone		
12	No standalone is available	Python- and Perl-based standalone is available
13	No GitHub repository is provided	GitHub repository is available with standalone
14	No docker based distribution is available	Docker based option is available via GPSRdocker

Other than HLADR4Pred, there are number of other methods with the ability to predict the binders for HLA-class II alleles. Hence, it is crucial to benchmark the performance of the other existing methods with HLADR4Pred2. For that purpose, we have taken out the validation dataset and tested the performance of the existing methods on the same. Propred is able to predict the HLA-DR binding sites and able to achieve 55.26% accuracy with AUROC 0.74, where NetMHCIIpan 4.0 achieved accuracy 65.82% with AUROC 0.72, followed by TEPITOPE with accuracy of 67.75% with balanced sensitivity and specificity, SMM-align

predicts the MHC class II binding affinity using stabilization matrix alignment method achieved accuracy of 67.95%. Artificial neural network based method i.e. NNAlign develop the model on sequence motifs detected in the training data, attained the accuracy of 68.64%, followed by consensus IEDB method with uses the consensus of SMM-Align, NNAlign, and Sturniolo method to calculate the adjusted rank based on which the predictions are made and it attained the accuracy of 69.41% on the independent dataset. Finally, older version of HLADR4Pred2 achieved the accuracy of 75.04 with AUROC 0.69, but the difference between sensitivity and specificity is significant. Our new approach has outperformed all the existing with methods with AUROC of 0.961 and accuracy 87.39%. These results showed that HLADR4Pred2 is an reliable method which has outperformed the other methods on the independent dataset which was not used while training or testing the model.

Table 12: Comparison of HLADR4Pred2 approach with the existing methods

Methods	Sensitivity	Specificity	Accuracy	AUROC	F1-score	Kappa	MCC
Propred	78.378	44.156	55.263	0.735	0.532	0.181	0.219
NetMHCIIpan 4.0	65.249	66.253	65.819	0.717	0.623	0.311	0.313
TEPITOPE	68.278	67.336	67.747	NA	0.297	0.347	0.353
SMM-Align	68.535	67.495	67.948	NA	0.292	0.357	0.358
NNAlign	68.946	68.410	68.643	NA	0.288	0.367	0.371
Consensus IEDB	69.409	69.404	69.406	NA	0.283	0.380	0.385
Hladr4pred	54.098	79.447	75.036	0.690	0.430	0.279	0.289
HLADR4Pred2	89.640	90.213	90.143	0.988	0.859	0.745	0.746

4.10 Case Study: HLA-DRB1*04:01-binders in COVID-19 variants

Recent studies report that HLA-DRB1*04:01 binding sites are associated with the severity of COVID-19 patients (79–81). The mutations associated with spike protein in COVID-19 variants can alter the binding of peptides (82,83). In order to understand the effect of mutations in different variants of COVID-19 with the HLA-binding peptides, we utilized "SCAN" module of our HLA-DR4Pred 2.0 server (https://webs.iiitd.edu.in/raghava/hladr4pred2/scan.php). First we created mutated proteins of COVID-19 variants using the reference spike protein sequence. As reported in Centres for Disease Control and Prevention (CDC portal) [https://www.cdc.gov/hai/data/portal/], the alpha variant possess seven mutation named as N501Y, A570D, D613G, P681H, T716I, D981A and D1118H, whereas beta variant in corporate D80A, D215G, K417N, E484K, N501Y, D614G, A701V, L18F and R246I mutations. Similarly, spike protein of delta variant incorporates

T19R, T95I, G142D, R158G, L452R, T478K, D614G, L681R and D950N mutations. Recently, reported COVID-19 variant Omicron possess highest number of mutations i.e., 30 mutations in spike protein A67V, del 69-70, T95I, G142D, del 143-145, del 211, L212I, ins214EPE, G339D, S371L, S373P, S375F, K417N, N440K, G446S, S477N, T478K, E484A, Q493K, G496S, Q498R, N501Y, Y505H, T547K, D614G, H655Y , N679K, P681H, N764K, D796Y , N856K, Q954H, N969K, L981F. Currently, we created the mutated proteins of different variants of COVID-19 and predict the binding peptides and effect of mutation on bindings in different COVID variants. We observed that in alpha variant (D981A and D613G), beta variant (D80A), gamma variant (D137Y), delta variant (G142D, L681R) and omicron associated mutations alter the nature of HLA-binding peptides to non-binders or vice versa, as shown in Table 13.

Table 13: Alterations in the binding peptides of HLA-DRB1*04:01 by mutations in Spike protein of SARS-CoV-2 variants

COVID-19 Variants	Mutation	Reference peptide	Mutated Peptide	Prediction (Binder/Non-Binder)	
				Reference	Mutated
Alpha	D981A	SGTNGTKRFDNPVLP	SGTNGTKRFANPVLP	Binder	Non-Binder
		GTNGTKRFDNPVLPF	GTNGTKRFANPVLPF	Binder	Non-Binder
		TNGTKRFDNPVLPFN	TNGTKRFANPVLPFN	Binder	Non-Binder
		RFDNPVLPFNDGVYF	RFANPVLPFNDGVYF	Non-Binder	Binder
		FDNPVLPFNDGVYFA	FANPVLPFNDGVYFA	Non-Binder	Binder
	D614G	RDLPQGFSALEPLVD	RGLPQGFSALEPLVD	Non-Binder	Binder
Beta	D80A	SGTNGTKRFDNPVLP	SGTNGTKRFANPVLP	Binder	Non-Binder
		GTNGTKRFDNPVLPF	GTNGTKRFANPVLPF	Binder	Non-Binder
		TNGTKRFDNPVLPFN	TNGTKRFANPVLPFN	Binder	Non-Binder
		RFDNPVLPFNDGVYF	RFANPVLPFNDGVYF	Non-Binder	Binder
		FDNPVLPFNDGVYFA	FANPVLPFNDGVYFA	Non-Binder	Binder
Gamma	D137Y	VIKVCEFQFCNDPFL	VIKVCEFQFCNYPFL	Binder	Non-Binder
		IKVCEFQFCNDPFLG	IKVCEFQFCNYPFLG	Binder	Non-Binder
		CNDPFLGVYYHKNNK	CNYPFLGVYYHKNNK	Binder	Non-Binder
		NDPFLGVYYHKNNKS	NYPFLGVYYHKNNKS	Binder	Non-Binder
Delta	G142D	NDPFLGVYYHKNNKS	NDPFLDVYYHKNNKS	Non-Binder	Binder
	L681R	YLYRLFRKSNLKPFE	YRYRLFRKSNLKPFE	Non-Binder	Binder
Omicron	Del 68-69, 141, 142, 144, 210	GTNGTKRFDNPVLPF	NGTKRFDNPVLPFND	Binder	Non-Binder
		TNGTKRFDNPVLPFN	GTKRFDNPVLPFNDG	Binder	Non-Binder
		GTKRFDNPVLPFNDG	KRFDNPVLPFNDGVY	Non-Binder	Binder
		KRFDNPVLPFNDGVY	FDNPVLPFNDGVYFA	Binder	Non-Binder
		RFDNPVLPFNDGVYF	DNPVLPFNDGVYFAS	Non-Binder	Binder
		FDNPVLPFNDGVYFA	NPVLPFNDGVYFAST	Non-Binder	Binder

Chapter 5
Scientific Service

5.1 Webserver Architecture

We have developed the user-friendly updated version of our old webserver HLADR4Pred, and named it as HLADR4Pred 2.0 to predict, scan, and design the HLA-DRB1*04:01 binding peptides. The front-end of the webserver was developed using HTML (v5), PHP (v7), CSS (v3), and JavaScript (v 1.8). The backend of the server uses Perl and python 3.6. The server is developed on a Linux (Ubuntu v14.04.6) and based on responsive template that is the resolution gets adjusted as per the screen size. The compatibility of the server is tested and found out to be compatible with all the modern devices like mobile, tablet, laptop, iMac, and desktop. The server incorporates six major modules such as predict, scan, design, blast, motif-scan, and standalone.

5.2Webserver Implementation

In order to serve the scientific community, we have developed an easy-to-use webserver using HTML5, CSS3, PHP7, and JavaScript and named it HLA-DR4Pred 2.0 which is available at https://webs.iiitd.edu.in/raghava/hladr4pred2/. There are six major modules in the server such as, “PREDICT”, “SCAN”, “DESIGN”, “BLAST”, “MOTIF-SCAN”, and “STANDALONE”. The description of each module in provided below.

a) PREDICT: This module allows users to predict the potential of an uncharacterized peptide to a HLA-DRB1*04:01 binder. It allows to provide either paste or upload a file containing single or multiple peptide sequences in FASTA format with length between 9 - 22. This module also permits to choose a desired threshold for prediction along with physicochemical properties to be displayed. The resulting page exhibits the score(s) and prediction which can be downloadable in the .csv format.

b) SCAN: This module allows user to provide sequences with length more than 22, which is a constraint in the predict module. In this module, users are allowed to paste or upload a sequence file in FASTA format. Users are asked to choose a desired window size on which the overlapping patterns are generated from the input sequence(s) and used them to make predictions. The users are allowed to choose the output format as graphical or tabular. The graphical output page highlights the binders in the submitted sequences. The tabular output page provides the start and end position of the generated patterns along with score and prediction as binder or non-binder based on the selected threshold. Users can download that results in the .csv format.

c) DESIGN: The design method permits users to generate all the possible mutants of an input sequence by mutating each residue at a time and use the same to predict if the mutated pattern is a binder or not. This module comes with the restriction of length between 9-22. Users are allowed to submit sequences in the FATSA format only in the text or file form which can be uploaded. The result will exhibit the occurred mutation with wildtype to mutant residue along with the position at which it occurred. The output page is downloadable in the .csv format. Figure 8 exhibits the function of three major modules such as “predict”, “scan” and “design”. It exhibits the input as well as the output page of each module.

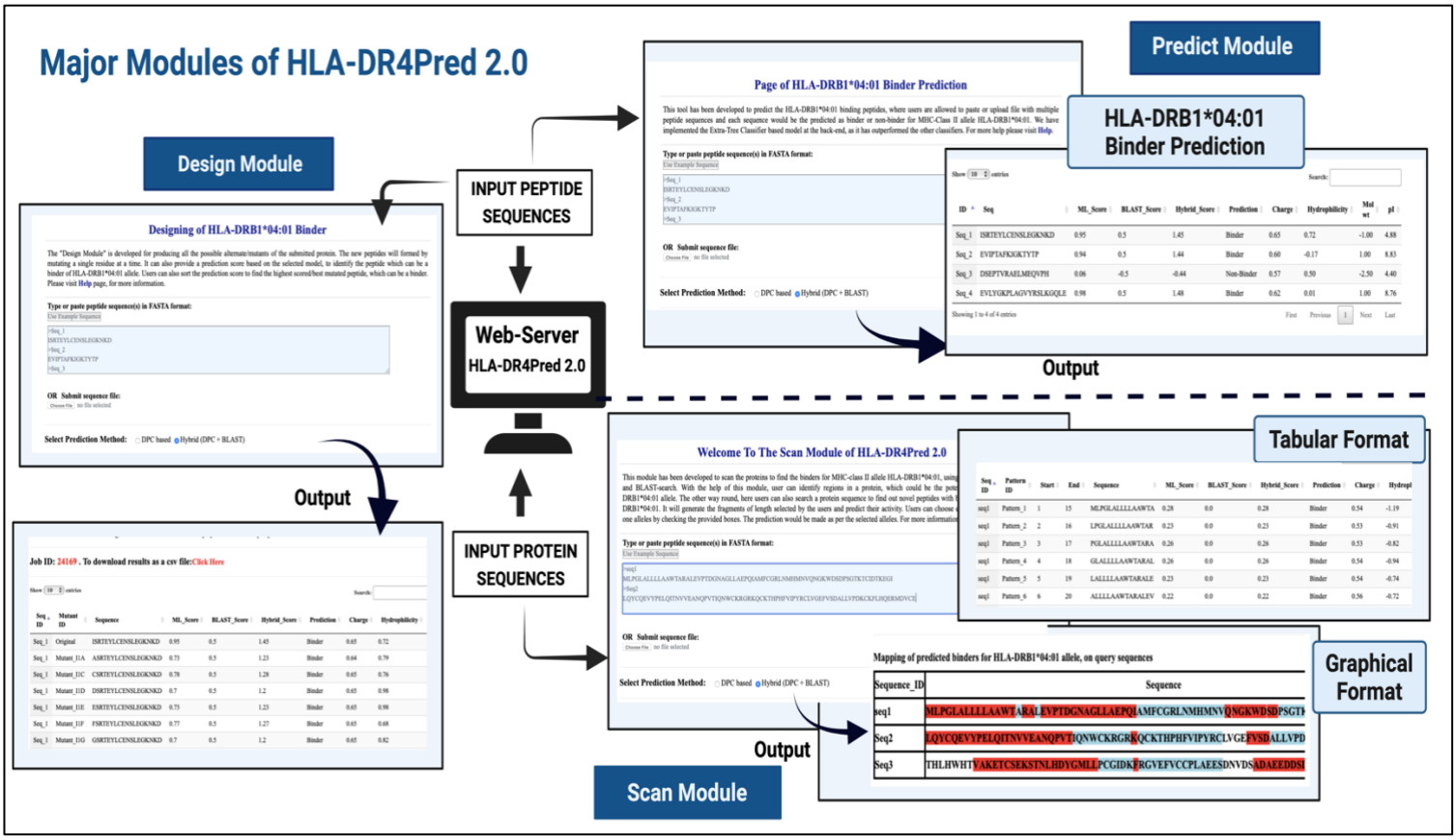


Figure 8: Usage of predict, scan, and design module of HLADR4Pred 2.0

d) BLAST: In the present module, user can make the predictions if a submitted sequence(s) is a binder or non-binder by performing similarity search using BLAST. The page permits to choose a desired e-value on which the prediction will be made as binder if a hit is found in the custom database else predicted as non-binder. This module takes the input sequence(s) in the FASTA format and the output page is downloadable in the .csv format.

e) MOTIF-SCAN: In this approach HLA-DRB1*04:01 binding motifs are searched in the input sequences and the predicted as binder if the motifs is found else designated as non-binder. This module also allows to choose ten different physicochemical properties to be calculated as displayed. The result page provides the prediction for each submitted sequence, and it permits to download the results in the .csv format. Figure 9 shows the usage of modules based on similarity and motif search such as BLAST AND MOTIF-SCAN, respectively. It exhibits the input and output page of each module.

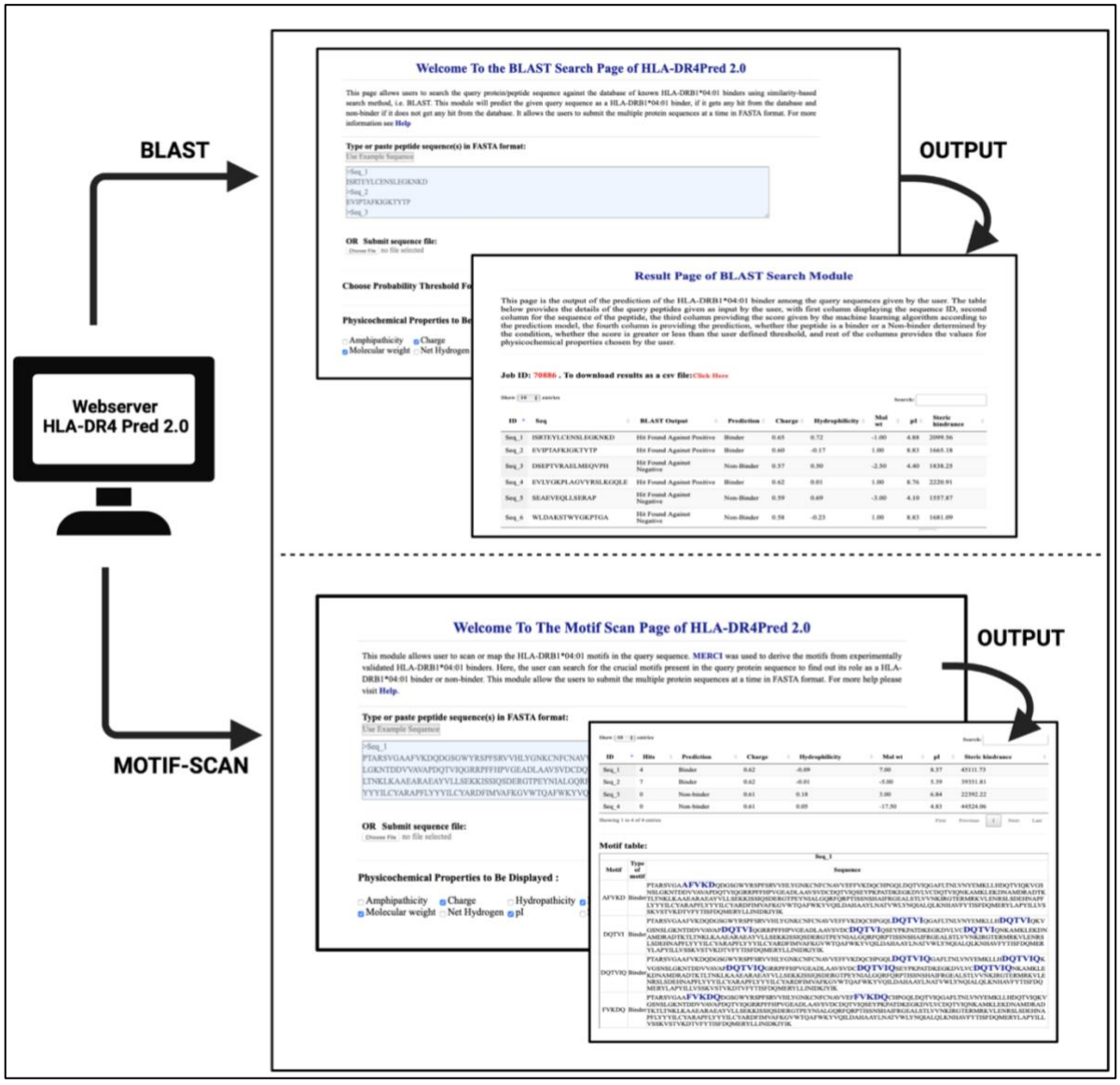


Figure 9: Usage of BLAST and Motif-scan module of HLADR4Pred 2.0

f) STANDALONE: We have developed python- and Perl based standalone which users can use in their local machines in the absence of the internet or for large number of sequences which may take large amount of time on webserver. The standalone is available for download at webserver(https://webs.iiitd.edu.in/raghava/hladr4pred2/standalone.php) and at GitHub(https://github.com/raghavagps/hladr4pred2). The same standalone is also available through docker facility via GPSRdocker package (84).

5.3 Standalone Development and Implementation

The entire data analysis and prediction pipeline was coded in Python 3.8.5 using scikit-learn (76) library for the development of prediction models using different classifiers, 'pandas' library was used to load and pre-process the data. Webserver comes with some restrictions due to space and computational power issues, make it not suitable to run or submit huge datasets which may take longer computational time. Moreover, the availability of internet is necessary to use the facilities on the webserver. To handle such challenges, we have developed standalone in two different programming languages such as Python and Perl, which are checked on different operating systems such as windows, Linux, and macOS. Moreover, to save users from the installation of complex libraries/dependencies, we have incorporate the same package in docker facility which can be used via GPSRdocker (84).

The standalone version of HLADR4Pred2 is easy-to-use, which takes the fasta file comprising peptide(s) as the input and provides the output in the .csv format. Figure 10 represents the usage of python-based standalone, in which user can get the complete help using command "python hladr4pred2.py -h". The only required argument in the standalone is the input file comprising of peptide sequences, where rest of the arguments are optional. It takes input file with '-i' tag, '-o' tag to define the output file name, if -o tag is not given, it stores the output with filename outfile.csv. Moreover, there are three types of jobs a user can give such as "predict", "scan", and "design" which works exactly same as in the webserver. This method also allow users to set a threshold using "-t" tag as per their wish, if not given it takes the default value of 0.16. Python-based standalone also takes window length of the peptide as an input under "-w" tag, it is an essential argument while using performing the job of scanning in longer sequences. Finally, "-d" tag is responsible for the display function and it takes input as 1 or 2, where 1 only stores/display binders sequences among the submitted sequences and option 2 provides the prediction of each sequence submitted as the query.

```
(base) [einstein@Sumeets-MacBook-Air hladr4pred2]$ python3 hladr4pred2.py -h
usage: hladr4pred2.py [-h] -i INPUT [-o OUTPUT] [-j {1,2,3}] [-t THRESHOLD]
                      [-w {9,10,11,12,13,14,15,16,17,18,19,20,21,22}]
                      [-d {1,2}]

Please provide following arguments

optional arguments:
  -h, --help            show this help message and exit
  -i INPUT, --input INPUT
                        Input: protein or peptide sequence(s) in FASTA format
                        or single sequence per line in single letter code
  -o OUTPUT, --output OUTPUT
                        Output: File for saving results by default outfile.csv
  -j {1,2,3}, --job {1,2,3}
                        Job Type: 1:Predict, 2: Design, 3:Scan, by default 1
  -t THRESHOLD, --threshold THRESHOLD
                        Threshold: Value between 0 to 1 by default 0.16
  -w {9,10,11,12,13,14,15,16,17,18,19,20,21,22}, --winleng {9,10,11,12,13,14,15,16,17,18,19,20,
21,22}
                        Window Length: 9 to 20 (scan mode only), by default 9
  -d {1,2}, --display {1,2}
                        Display: 1:Binders only, 2: All peptides, by default 1
```

Figure 10: Usage of python-based standalone of HLA-DR4Pred2.0

Similarly, Perl-based standalone also works with tag like '-i', '-o', '-t', '-m' and '-s' as shown in Figure 11 which displays the entire usage. The '-i' defines the input file in fasta format, '-o' is for defining the output filename in which the results will be stored, '-t' defines the threshold, '-m' is to provides the method such as 1 for prediction, 2 for scanning, and 3 for designing the binders, and '-s' defines the scan length used in the scanning module using which the overlapping patterns of a longer peptide/protein will be generated to make the predictions. Figure 11 also shows the example usage which as user can use to run the example provided along with the standalone. Similar options are given when user wants to use the docker version of HLADR4Pred2 by pulling the image of GPSRdocker (84).

```
(base) sumeet@gpsr:~/standalone/hladr4pred2$ perl hladr4pred2.pl
USAGE: hladr4pred2.pl -i <fasta format sequences> -o <output file name> -t <threshold> -m <method> -s <scanning length>

Example Command: ./hladr4pred2.pl -i /gpsr/examples/example_hladr4pred2.fasta -o out -t 0.16 -m 1 -s 9

-i      Sequence in FASTA or single line format
-o      output file
-t      threshold
-m      Methods:
        Choose methods from following options, and provide option as 1,2 or 3:
        1 for Prediction
        2 for Scan
        3 for Design
        By Default its 1
-s      Length of the overlapping patterns to be generated from the submitted sequences
```

Figure 11: Usage of Perl-based standalone of HLA-DR4Pred2.0

Chapter 6
Discussion & Conclusion

HLA system is the major histocompatibility complex in humans and is the most import part of our immune system (6). HLA genes regulate the immune responses while infectious diseases and viral/pathogenic attack and provide protections (1,85–87). Due to high polymorphism, thousands of HLA alleles are reported in IMGT/HLA database (5), out of which few were associated with number of diseases. In the past few decades researches proves that one of the HLA-DR4 family allele HLA-DRB1*04 play major role in the regulation of immune responses and associated with several autoimmune disorders and COVID-19 severity (79,80,85,88). Therefore the identification of HLA-DRB1*04-binding peptides is very crucial for understanding the severity of autoimmune diseases (87,89,90). Therefore in the past a number of computational tools have been developed for the identification of HLA-binding peptides (64,91–94). These tools predict the binders against different HLA-alleles. In order to strengthen the previous studies and to improve the accuracy of prediction models we have developed a highly accurate method named "HLA-DR4Pred 2.0".

In the current study, we have extracted the experimentally validated HLA-DRB1*04:01 binding and non-binding peptides from IEDB. A total of 12676 binders (i.e., positive dataset) and 86300 non-binders (i.e., negative dataset) collected for the development of prediction models. From amino acid composition we observed that, serine amino acid is highly prominent in the HLA-DRB1*04:01 binding peptides in comparison with non-binders. Positional analysis also revealed that Serine residue is predominantly located at 9th, 13th, 14th , 15th and 16th positions in positive dataset, whereas leucine highly conserved in negative datasets. Firstly we have computed various composition-based features (AAC, DPC, ATC, BTC, PCP, RRI, SER, DDOR, SEP, CTC, CeTD, PAAC, APAAC, QSO, SOCN) and binary profile based features using Pfeature standalone package. We have developed various machine learning models using eight different classifiers such as SVC, DT, RF, XGB, KNN, LR, ET, and GNB. As shown in most of results developed on different feature in datasets, ET based models outperform the other classifiers. The performance on the realistic dataset has outperform the other datasets (Table 5,6,7,8,9). Similarly, in case of binary profile based features NC_{22} based models outperform the other patterns as shown in Table 6.

The performance is in the terms of AUROC is 0.94 on training and validation datasets. While the performance on combined features in around 0.94 AUROC on training and validation realistic dataset. After selecting the best features we obtained highest AUROC of 0.95 on training and validation dataset. We have observed that the DPC based models achieved the maximum AUROC of 0.96, accuracy is more than 89% on training and validation datasets

(Table 5). In order to achieve the maximum performance we have merged machine learning technique with similarity search using BLAST and attained 0.98 AUROC on training and 0.99 on validation datasets. AUROC plots for best performing features in each feature type is provided in Figure 12.

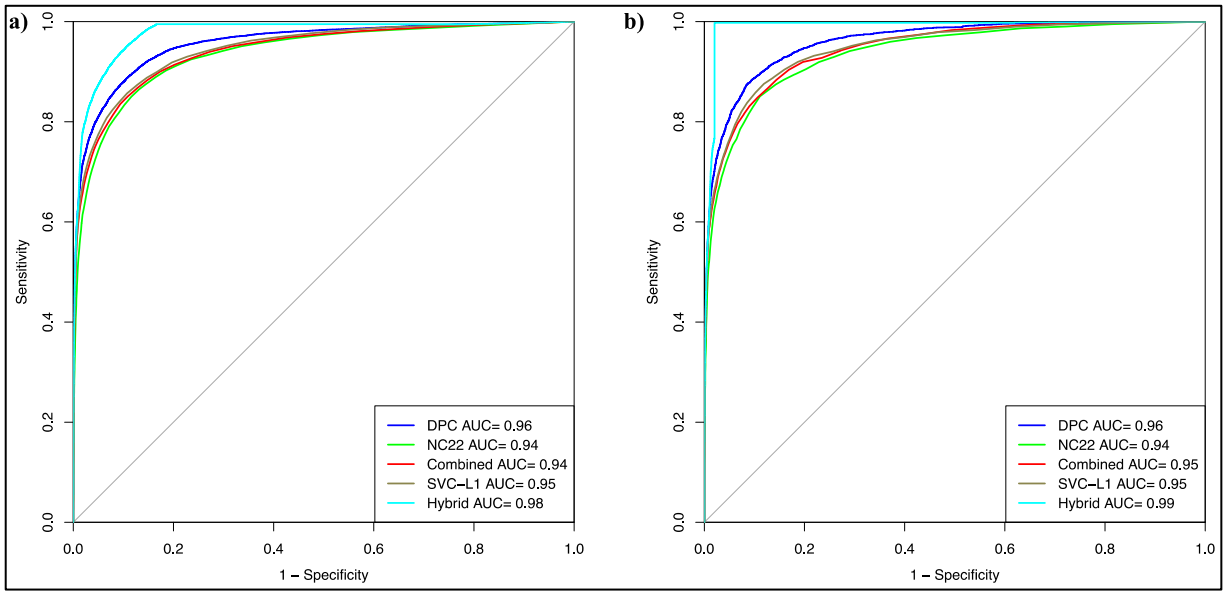


Figure 12: Comparison between the best performing features in each feature type

To compare our performance of our model with the existing methods, we have considered seven different methods as shown in Table 12, and found out that HLA-DR4Pred2.0 has outperformed all the other methods with AUROC of 0.961. In order to serve the scientific community we have developed a webserver and standalone package using the best features and classifiers. HLA-DR4Pred 2.0 incorporates five modules such as PREDICT, SCAN, DESIGN, BLAST, and MOTIF-SCAN. HLA-DR4Pred 2.0 tool predict the binding or non-binding peptides for MHC-Class II allele HLA-DRB1*04:01. Our webserver is freely accessible at https://webs.iiitd.edu.in/raghava/hladr4pred2/ and standalone package is available at https://webs.iiitd.edu.in/raghava/hladr4pred2/standalone.php. Detailed workflow of this study is represented in Figure 13.

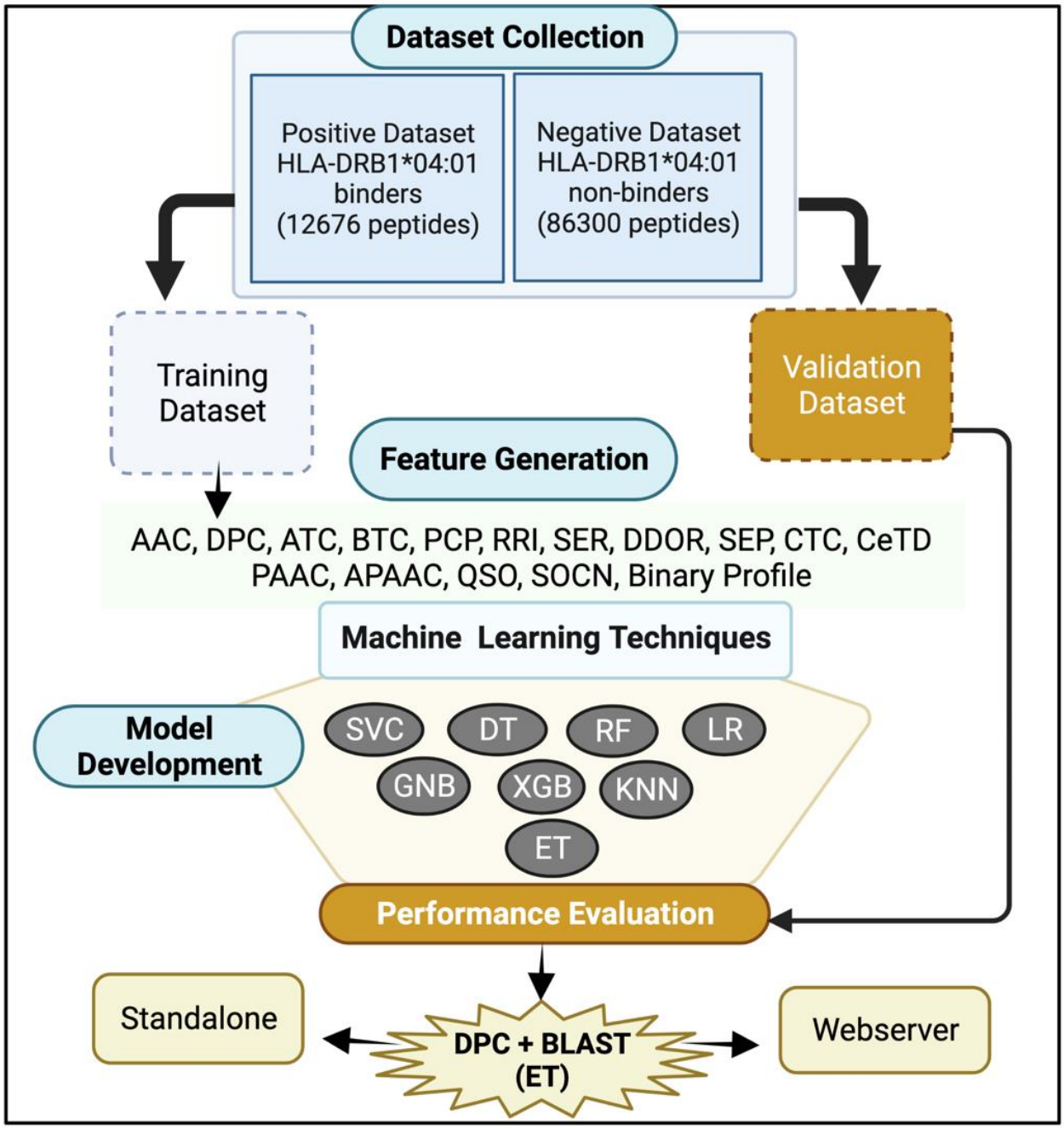


Figure 13: Overall workflow of the study

Chapter 7
Bibliography

1. Crux NB, Elahi S. Human Leukocyte Antigen (HLA) and Immune Regulation: How Do Classical and Non-Classical HLA Alleles Modulate Immune Response to Human Immunodeficiency Virus and Hepatitis C Virus Infections? Front Immunol. 2017;8:832.

2. Shiina T, Hosomichi K, Inoko H, Kulski JK. The HLA genomic loci map: expression, interaction, diversity and disease. J Hum Genet. 2009 Jan;54(1):15–39.

3. Choo SY. The HLA system: genetics, immunology, clinical testing, and clinical implications. Yonsei Med J. 2007 Feb;48(1):11–23.

4. Wang M, Claesson MH. Classification of human leukocyte antigen (HLA) supertypes. Methods Mol Biol. 2014;1184:309–17.

5. Robinson J, Barker DJ, Georgiou X, Cooper MA, Flicek P, Marsh SGE. IPD-IMGT/HLA Database. Nucleic Acids Res. 2020 Jan;48(D1):D948–55.

6. Mosaad YM. Clinical Role of Human Leukocyte Antigen in Health and Disease. Scand J Immunol [Internet]. 2015 Oct [cited 2018 Aug 1];82(4):283–306. Available from: http://doi.wiley.com/10.1111/sji.12329

7. Zheng D, Liwinski T, Elinav E. Interaction between microbiota and immunity in health and disease. Cell Res. 2020 Jun;30(6):492–506.

8. Leone P, Shin E-C, Perosa F, Vacca A, Dammacco F, Racanelli V. MHC class I antigen processing and presenting machinery: organization, function, and defects in tumor cells. J Natl Cancer Inst. 2013 Aug;105(16):1172–87.

9. Adler LN, Jiang W, Bhamidipati K, Millican M, Macaubas C, Hung S-C, et al. The Other Function: Class II-Restricted Antigen Presentation by B Cells. Front Immunol. 2017;8:319.

10. Sanchez-Trincado JL, Gomez-Perosanz M, Reche PA. Fundamentals and Methods for T- and B-Cell Epitope Prediction. J Immunol Res. 2017;2017:2680160.

11. Holland CJ, Cole DK, Godkin A. Re-Directing CD4(+) T Cell Responses with the Flanking Residues of MHC Class II-Bound Peptides: The Core is Not Enough. Front Immunol. 2013;4:172.

12. Wieczorek M, Abualrous ET, Sticht J, Alvaro-Benito M, Stolzenberg S, Noe F, et al. Major Histocompatibility Complex (MHC) Class I and MHC Class II Proteins: Conformational Plasticity in Antigen Presentation. Front Immunol. 2017;8:292.

13. Nielsen M, Lund O, Buus S, Lundegaard C. MHC class II epitope predictive algorithms. Immunology. 2010 Jul;130(3):319–28.

14. Rock KL, Reits E, Neefjes J. Present Yourself! By MHC Class I and MHC Class II Molecules. Trends Immunol. 2016 Nov;37(11):724–37.

15. Rammensee HG, Falk K, Rotzschke O. Peptides naturally presented by MHC class I molecules. Annu Rev Immunol. 1993;11:213–44.

16. Dunston GM, Halder RM. Vitiligo is associated with HLA-DR4 in black patients. A preliminary report. Arch Dermatol. 1990 Jan;126(1):56–60.

17. Taurog JD. HLA-DR4 and the spondyloarthropathies. Ann Rheum Dis. 2002 Mar;61(3):193–4.

18. Shi T, Lv W, Zhang L, Chen J, Chen H. Association of HLA-DR4/HLA-DRB1*04 with Vogt-Koyanagi-Harada disease: a systematic review and meta-analysis. Sci Rep. 2014 Nov;4:6887.

19. Stastny P, Ball EJ, Khan MA, Olsen NJ, Pincus T, Gao X. HLA-DR4 and other genetic markers in rheumatoid arthritis. Br J Rheumatol. 1988;27 Suppl 2:132–8.

20. Brassat D, Salemi G, Barcellos LF, McNeill G, Proia P, Hauser SL, et al. The HLA locus and multiple sclerosis in Sicily. Neurology. 2005 Jan;64(2):361–3.

21. Hoffmann S, Cepok S, Grummel V, Lehmann-Horn K, Hackermuller J, Stadler PF, et al. HLA-DRB1*0401 and HLA-DRB1*0408 are strongly associated with the development of antibodies against interferon-beta therapy in multiple sclerosis. Am J Hum Genet. 2008 Aug;83(2):219–27.

22. Muniz-Castrillo S, Vogrig A, Honnorat J. Associations between HLA and autoimmune neurological diseases with autoantibodies. Auto- Immun highlights. 2020 Jan;11(1):2.

23. Larsen CE, Alper CA. The genetics of HLA-associated disease. Curr Opin Immunol. 2004 Oct;16(5):660–7.

24. Kovalchuka L, Eglite J, Lucenko I, Zalite M, Viksna L, Krumina A. Associations of HLA DR and DQ molecules with Lyme borreliosis in Latvian patients. BMC Res Notes. 2012 Aug;5:438.

25. Newton JL, Harney SMJ, Wordsworth BP, Brown MA. A review of the MHC genetics of rheumatoid arthritis. Genes Immun. 2004 May;5(3):151–7.

26. Yamout BI, Alroughani R. Multiple Sclerosis. Semin Neurol. 2018 Apr;38(2):212–25.

27. Maahs DM, West NA, Lawrence JM, Mayer-Davis EJ. Epidemiology of type 1 diabetes. Endocrinol Metab Clin North Am. 2010 Sep;39(3):481–97.

28. Gillespie KM. Type 1 diabetes: pathogenesis and prevention. CMAJ. 2006 Jul;175(2):165–70.

29. McIver B, Morris JC. The pathogenesis of Graves' disease. Endocrinol Metab Clin North Am. 1998 Mar;27(1):73–89.

30. Khan H, Sureda A, Belwal T, Cetinkaya S, Suntar I, Tejada S, et al. Polyphenols in the treatment of autoimmune diseases. Autoimmun Rev. 2019 Jul;18(7):647–57.

31. Lundegaard C, Lund O, Buus S, Nielsen M. Major histocompatibility complex class I binding predictions as a tool in epitope discovery. Immunology. 2010 Jul;130(3):309–

18.

32. Nielsen M, Lundegaard C, Blicher T, Peters B, Sette A, Justesen S, et al. Quantitative predictions of peptide binding to any HLA-DR molecule of known sequence: NetMHCIIpan. PLoS Comput Biol. 2008 Jul;4(7):e1000107.

33. Lin HH, Zhang GL, Tongchusak S, Reinherz EL, Brusic V. Evaluation of MHC-II peptide binding prediction servers: applications for vaccine research. BMC Bioinformatics. 2008 Dec;9 Suppl 12:S22.

34. Wang P, Sidney J, Kim Y, Sette A, Lund O, Nielsen M, et al. Peptide binding predictions for HLA DR, DP and DQ molecules. BMC Bioinformatics. 2010 Nov;11:568.

35. Antunes DA, Devaurs D, Moll M, Lizee G, Kavraki LE. General Prediction of Peptide-MHC Binding Modes Using Incremental Docking: A Proof of Concept. Sci Rep. 2018 Mar;8(1):4327.

36. Wu J, Wang W, Zhang J, Zhou B, Zhao W, Su Z, et al. DeepHLApan: A Deep Learning Approach for Neoantigen Prediction Considering Both HLA-Peptide Binding and Immunogenicity. Front Immunol. 2019;10:2559.

37. Bhasin M, Raghava GPS. SVM based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence. Bioinformatics. 2004 Feb;20(3):421–3.

38. Singh H, Raghava GP. ProPred: prediction of HLA-DR binding sites. Bioinformatics [Internet]. 2001 Dec [cited 2018 Jun 27];17(12):1236–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11751237

39. Nielsen M, Lundegaard C, Lund O. Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. BMC Bioinformatics. 2007 Jul;8:238.

40. Bui H-H, Sidney J, Peters B, Sathiamurthy M, Sinichi A, Purton K-A, et al. Automated generation and evaluation of specific MHC binding predictive tools: ARB matrix applications. Immunogenetics. 2005 Jun;57(5):304–14.

41. Nielsen M, Andreatta M. NNAlign: a platform to construct and evaluate artificial neural network models of receptor-ligand interactions. Nucleic Acids Res. 2017 Jul;45(W1):W344–9.

42. Reynisson B, Alvarez B, Paul S, Peters B, Nielsen M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. Nucleic Acids Res. 2020 Jul;48(W1):W449–54.

43. Karosiene E, Rasmussen M, Blicher T, Lund O, Buus S, Nielsen M. NetMHCIIpan-3.0, a common pan-specific MHC class II prediction method including all three human MHC class II isotypes, HLA-DR, HLA-DP and HLA- DQ. Immunogenetics. 2013 Oct;65(10):711–24.

44. Abelin JG, Harjanto D, Malloy M, Suri P, Colson T, Goulding SP, et al. Defining HLA-II ligand processing and binding rules with mass spectrometry enhances cancer epitope prediction. Vol. 54, Immunity. United States; 2021. p. 388.

45. Sidney J, Southwood S, Moore C, Oseroff C, Pinilla C, Grey HM, et al. Measurement of MHC/peptide interactions by gel filtration or monoclonal antibody capture. Curr Protoc Immunol. 2013 Feb;Chapter 18:Unit 18.3.

46. Bassani-Sternberg M, Coukos G. Mass spectrometry-based antigen discovery for cancer immunotherapy. Curr Opin Immunol. 2016 Aug;41:9–17.

47. Caron E, Kowalewski DJ, Chiek Koh C, Sturm T, Schuster H, Aebersold R. Analysis of Major Histocompatibility Complex (MHC) Immunopeptidomes Using Mass Spectrometry. Mol Cell Proteomics. 2015 Dec;14(12):3105–17.

48. Wendorff M, Garcia Alvarez HM, Osterbye T, ElAbd H, Rosati E, Degenhardt F, et al. Unbiased Characterization of Peptide-HLA Class II Interactions Based on Large-Scale Peptide Microarrays; Assessment of the Impact on HLA Class II Ligand and Epitope Prediction. Front Immunol. 2020;11:1705.

49. Salvat R, Moise L, Bailey-Kellogg C, Griswold KE. A high throughput MHC II binding assay for quantitative analysis of peptide epitopes. J Vis Exp. 2014 Mar;(85).

50. Steere AC, Klitz W, Drouin EE, Falk BA, Kwok WW, Nepom GT, et al. Antibiotic-refractory Lyme arthritis is associated with HLA-DR molecules that bind a Borrelia burgdorferi peptide. J Exp Med. 2006 Apr;203(4):961–71.

51. Raddrizzani L, Sturniolo T, Guenot J, Bono E, Gallazzi F, Nagy ZA, et al. Different modes of peptide interaction enable HLA-DQ and HLA-DR molecules to bind diverse peptide repertoires. J Immunol. 1997 Jul;159(2):703–11.

52. Stern LJ, Brown JH, Jardetzky TS, Gorga JC, Urban RG, Strominger JL, et al. Crystal structure of the human class II MHC protein HLA-DR1 complexed with an influenza virus peptide. Nature. 1994 Mar;368(6468):215–21.

53. Zhu Y, Rudensky AY, Corper AL, Teyton L, Wilson IA. Crystal structure of MHC class II I-Ab in complex with a human CLIP peptide: prediction of an I-Ab peptide-binding motif. J Mol Biol. 2003 Feb;326(4):1157–74.

54. Rammensee HG, Friede T, Stevanoviic S. MHC ligands and peptide motifs: first listing. Immunogenetics. 1995;41(4):178–228.

55. Sturniolo T, Bono E, Ding J, Raddrizzani L, Tuereci O, Sahin U, et al. Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. Nat Biotechnol. 1999 Jun;17(6):555–61.

56. Nielsen M, Lund O. NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. BMC Bioinformatics. 2009 Sep;10:296.

57. Dimitrov I, Garnev P, Flower DR, Doytchinova I. EpiTOP--a proteochemometric tool for MHC class II binding prediction. Bioinformatics. 2010 Aug;26(16):2066–8.

58. Brusic V, Rudy G, Honeyman G, Hammer J, Harrison L. Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network. Bioinformatics. 1998;14(2):121–30.

59. Zhang GL, Khan AM, Srinivasan KN, August JT, Brusic V. MULTIPRED: a computational system for prediction of promiscuous HLA binding peptides. Nucleic Acids Res. 2005 Jul;33(Web Server issue):W172-9.

60. Jensen KK, Andreatta M, Marcatili P, Buus S, Greenbaum JA, Yan Z, et al. Improved methods for predicting peptide binding affinity to MHC class II molecules. Immunology. 2018 Jul;154(3):394–406.

61. Wang P, Sidney J, Dow C, Mothe B, Sette A, Peters B. A systematic assessment of MHC class II peptide binding predictions and evaluation of a consensus approach. PLoS Comput Biol. 2008 Apr;4(4):e1000048.

62. Zhang L, Chen Y, Wong H-S, Zhou S, Mamitsuka H, Zhu S. TEPITOPEpan: extending TEPITOPE for peptide binding prediction covering over 700 HLA-DR molecules. PLoS One. 2012;7(2):e30483.

63. Atanasova M, Patronov A, Dimitrov I, Flower DR, Doytchinova I. EpiDOCK: a molecular docking-based tool for MHC class II binding prediction. Protein Eng Des Sel. 2013 Oct;26(10):631–4.

64. Laimer J, Lackner P. MHCII3D-Robust Structure Based Prediction of MHC II Binding Peptides. Int J Mol Sci. 2020 Dec;22(1).

65. Sette A, Sidney J, Oseroff C, del Guercio MF, Southwood S, Arrhenius T, et al. HLA DR4w4-binding motifs illustrate the biochemical basis of degeneracy and specificity in peptide-DR interactions. J Immunol. 1993 Sep;151(6):3163–70.

66. Hammer J, Valsasnini P, Tolba K, Bolin D, Higelin J, Takacs B, et al. Promiscuous and allele-specific anchors in HLA-DR-binding peptides. Cell. 1993 Jul;74(1):197–203.

67. Max H, Halder T, Kropshofer H, Kalbus M, Muller CA, Kalbacher H. Characterization of peptides bound to extracellular and intracellular HLA-DR1 molecules. Hum Immunol. 1993 Nov;38(3):193–200.

68. Chicz RM, Urban RG, Gorga JC, Vignali DA, Lane WS, Strominger JL. Specificity and promiscuity among naturally processed peptides bound to HLA-DR alleles. J Exp Med. 1993 Jul;178(1):27–47.

69. Marshall KW, Wilson KJ, Liang J, Woods A, Zaller D, Rothbard JB. Prediction of peptide affinity to HLA DRB1*0401. J Immunol. 1995 Jun;154(11):5927–33.

70. Southwood S, Sidney J, Kondo A, del Guercio MF, Appella E, Hoffman S, et al. Several

common HLA-DR types share largely overlapping peptide binding repertoires. J Immunol. 1998 Apr;160(7):3363–73.

71. Borras-Cuesta F, Golvano J, Garcia-Granero M, Sarobe P, Riezu-Boj J, Huarte E, et al. Specific and general HLA-DR binding motifs: comparison of algorithms. Hum Immunol. 2000 Mar;61(3):266–78.

72. Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JR, et al. The Immune Epitope Database (IEDB): 2018 update. Nucleic Acids Res. 2019 Jan;47(D1):D339–43.

73. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res [Internet]. 2000 Jan 1 [cited 2019 Jul 23];28(1):45–8. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/28.1.45

74. Vacic V, Iakoucheva LM, Radivojac P. Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. Bioinformatics. 2006 Jun;22(12):1536–7.

75. Pande A, Patiyal S, Lathwal A, Arora C, Kaur D, Dhall A, et al. Computing wide range of protein/peptide features from their sequence and structure. bioRxiv [Internet]. 2019 Apr 4 [cited 2020 Jun 14];599126. Available from: https://www.biorxiv.org/content/10.1101/599126v1

76. Pedregosa FABIANPEDREGOSA F, Michel V, Grisel OLIVIERGRISEL O, Blondel M, Prettenhofer P, Weiss R, et al. Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot [Internet]. Vol. 12, Journal of Machine Learning Research. 2011 [cited 2020 Jun 14]. Available from: http://scikit-learn.sourceforge.net.

77. McGinnis S, Madden TL. BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res. 2004 Jul;32(Web Server issue):W20-5.

78. Vens C, Rosso M-N, Danchin EGJ. Identifying discriminative classification-based motifs in biological sequences. 2011 [cited 2020 Jun 24];27(9):1231–8. Available from: http://dtai.cs.kuleuven.be/ml/systems/merci.

79. Ebrahimi S, Ghasemi-Basir HR, Majzoobi MM, Rasouli-Saravani A, Hajilooi M, Solgi G. HLA-DRB1*04 may predict the severity of disease in a group of Iranian COVID-19 patients. Hum Immunol. 2021 Oct;82(10):719–25.

80. de Sousa E, Ligeiro D, Lerias JR, Zhang C, Agrati C, Osman M, et al. Mortality in COVID-19 disease patients: Correlating the association of major histocompatibility complex (MHC) with severe acute respiratory syndrome 2 (SARS-CoV-2) variants. Int J Infect Dis. 2020 Sep;98:454–9.

81. Langton DJ, Bourke SC, Lie BA, Reiff G, Natu S, Darlay R, et al. The influence of HLA

genotype on the severity of COVID-19 infection. HLA. 2021 Jul;98(1):14–22.

82. Harvey WT, Carabelli AM, Jackson B, Gupta RK, Thomson EC, Harrison EM, et al. SARS-CoV-2 variants, spike mutations and immune escape. Nat Rev Microbiol. 2021 Jul;19(7):409–24.

83. Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, et al. Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. Cell. 2020 Aug;182(4):812-827.e19.

84. Agrawal P, Kumar R, Usmani SS, Dhall A, Patiyal S, Sharma N, et al. GPSRdocker: a Docker-based resource for genomics, proteomics and systems biology. BioRxiv. 2019;827766.

85. Tavasolian F, Rashidi M, Hatam GR, Jeddi M, Hosseini AZ, Mosawi SH, et al. HLA, Immune Response, and Susceptibility to COVID-19. Front Immunol. 2020;11:601886.

86. Blackwell JM, Jamieson SE, Burgner D. HLA and infectious diseases. Clin Microbiol Rev. 2009 Apr;22(2):370–85, Table of Contents.

87. Shi Y, Wang Y, Shao C, Huang J, Gan J, Huang X, et al. COVID-19 infection: the perspectives on immune responses. Vol. 27, Cell death and differentiation. England; 2020. p. 1451–4.

88. Garcia LF. Immune Response, Inflammation, and the Clinical Spectrum of COVID-19. Front Immunol. 2020;11:1441.

89. Miyadera H, Tokunaga K. Associations of human leukocyte antigens with autoimmune diseases: challenges in identifying the mechanism. J Hum Genet. 2015 Nov;60(11):697–702.

90. Alves C, Souza T, Meyer I, Toralles MBP, Brites C. Immunogenetics and infectious diseases: special reference to the mayor histocompatibility complex. Braz J Infect Dis. 2006 Apr;10(2):122–31.

91. Alvarez B, Barra C, Nielsen M, Andreatta M. Computational Tools for the Identification and Interpretation of Sequence Motifs in Immunopeptidomes. Proteomics. 2018 Jun;18(12):e1700252.

92. Chen B, Khodadoust MS, Olsson N, Wagar LE, Fast E, Liu CL, et al. Predicting HLA class II antigen presentation through integrated deep learning. Nat Biotechnol. 2019 Nov;37(11):1332–43.

93. Dimitrov I, Garnev P, Flower DR, Doytchinova I. MHC Class II Binding Prediction-A Little Help from a Friend. J Biomed Biotechnol. 2010;2010:705821.

94. Luo H, Ye H, Ng HW, Shi L, Tong W, Mendrick DL, et al. Machine Learning Methods for Predicting HLA-Peptide Binding Activity. Bioinform Biol Insights. 2015;9(Suppl 3):21–9.