Methods 232 (2024) 18–28

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

HLA-DR4Pred2: An improved method for predicting HLA-DRB1*04:01 binders

Sumeet Patiyal [1], Anjali Dhall [2], Nishant Kumar [3], Gajendra P.S. Raghava [4,*]

Department of Computational Biology, Indraprastha Institute of Information Technology, Okhla Phase 3, New Delhi 110020, India

ARTICLE INFO

Keywords:
HLA-DRB1*04:01
Immunotherapy
Machine Learning
Prediction Method
BLAST search
Antigen binders

ABSTRACT

HLA-DRB1*04:01 is associated with numerous diseases, including sclerosis, arthritis, diabetes, and COVID-19, emphasizing the need to scan for binders in the antigens to develop immunotherapies and vaccines. Current prediction methods are often limited by their reliance on the small datasets. This study presents HLA-DR4Pred2, developed on a large dataset containing 12,676 binders and an equal number of non-binders. It's an improved version of HLA-DR4Pred, which was trained on a small dataset, containing 576 binders and an equal number of non-binders. All models were trained, optimized, and tested on 80 % of the data using five-fold cross-validation and evaluated on the remaining 20 %. A range of machine learning techniques was employed, achieving maximum AUROC of 0.90 and 0.87, using composition and binary profile features, respectively. The performance of the composition-based model increased to 0.93, when combined with BLAST search. Additionally, models developed on the realistic dataset containing 12,676 binders and 86,300 non-binders, achieved a maximum AUROC of 0.99. Our proposed method outperformed existing methods when we compared the performance of our best model to that of existing methods on the independent dataset. Finally, we developed a standalone tool and a webserver for HLADR4Pred2, enabling the prediction, design, and virtual scanning of HLA-DRB1*04:01 binding peptides, and we also released a Python package available on the Python Package Index (https://webs.iiitd.edu.in/raghava/hladr4pred2/; https://github.com/raghavagps/hladr4pred2; https://pypi.org/project/hladr4pred2/).

1. Introduction

The human leukocyte antigen (HLA) complex is a highly polymorphic genomic region located at chromosome 6 in the human genome [1–3]. 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 is the largest repository of HLA-related sequences, reports thousands of human major histocompatibility complex (MHC) associated alleles and genomic sequences [5–7]. HLA class-I molecules display intracellular peptides to $CD8^+$ T-cells, whereas HLA class-II molecules are composed of two polypeptide chains (α and β) and presents extracellular peptides to $CD4^+$ T-cells. HLA-class II alleles are mainly presented on antigen presenting cells, for instance, B-cells, macrophages, dendritic cells (DCs), etc. [8–10]. The binding groove of MHC-II molecules is open at both ends, allowing longer peptides to extend beyond the groove from the flanking regions, as shown in Fig. 1 [11,12]. The majority of MHC class-II alleles present peptides that are derived from pathogenic proteins [13,14]. MHC class-II alleles bind peptides and present them on the cell surface, where they interacts with T-cell receptors (Fig. 1A). This interaction activates $CD4^+$ T-cells, which

Abbreviations: IEDB, Immune Epitope DataBase; HLA, Human Leukocyte Antigen; DCs, Dendritic Cells; DT, Decision Tree; RF, Random Forest; LR, Logistic Regression; XGB, eXtreme Gradient Boosting; KNN, K-Nearest Neighbor; GNB, Gaussian Naïve Bayes; ET, Extremely randomized Tree; SVC, Support Vector Classifier; MCC, Mathews Correlation Coefficient; AUROC, Area Under the Receiver Operating Characteristics curve; MERCI, Motif-Emerging and with Classes-Identification.

* Corresponding author at: Department of Computational Biology, Indraprastha Institute of Information Technology, Delhi, Okhla Industrial Estate, Phase III (Near Govind Puri Metro Station), New Delhi 110020, India.

E-mail addresses: sumeetp@iiitd.ac.in (S. Patiyal), anjalid@iiitd.ac.in (A. Dhall), nishantk@iiitd.ac.in (N. Kumar), raghava@iiitd.ac.in (G.P.S. Raghava).

[1] ORCID ID: https://orcid.org/0000-0003-1358-292X.
[2] ORCID ID: https://orcid.org/0000-0002-0400-2084.
[3] ORCID ID: https://orcid.org/0000-0001-7781-9602.
[4] ORCID ID: https://orcid.org/0000-0002-8902-2876.

https://doi.org/10.1016/j.ymeth.2024.10.007
Received 22 July 2024; Received in revised form 27 September 2024; Accepted 15 October 2024
Available online 19 October 2024

secrete cytokines, such, as interferon-gamma, tumor necrosis factor, and granulocyte–macrophage colony-stimulating factor (GM-CSF) to initiate the immune responses.

In the past, several studies have shown that the HLA-DR4 gene is highly correlated with several diseases [15–18], especially, HLA-DRB1 * 04:01 is associated with the development of multiple sclerosis [19,20], autoimmune disorders [21], type 1 diabetes [22], Lyme disease [23], COVID-19 severity, and rheumatoid arthritis [24]. HLA-DR4 molecules play a significant role in autoimmune disorders initiation and progression (Fig. 1B). 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 [25–29]. Studies also reveal that, patients positive with HLA-DR4 associated alleles have maximum chances of having autoimmune disorders, therefore it could be 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 [30,31]. Virtual scanning of HLA-DR4 binders at genome level is not practically feasible due to the time, cost, and complexity involved in experimental techniques. A number of computational methods have been developed to predict HLA-DR4 binders, which can facilitate the scientific community in performing the scanning of binders at large scale [32–36]. In 2004, our group developed the HLADR4Pred method, which is widely used and cited by the scientific community. One of the major limitations of HLADR4Pred is that it is trained on a small dataset that was available in 2004.We have provided the comprehensive comparison between the HLADR4Pred and HLADR4Pred2 in Supplementary Table 1.

In this study, we present an upgraded version of HLADR4Pred, which was trained using the largest dataset available in the immune epitope database (IEDB). We deployed state-of-the-art techniques to enhance the accuracy of predicting HLA-DRB1*04:01 binders. Initially, we created two datasets: the main dataset containing 12,676 HLA-DRB1*04:01 binders and an equal number of non-binders, and the alternate dataset containing 12,676 binders and 86,300 non-binders. In order to classify binders and non-binders, we developed classification models using a wide range of machine learning techniques. All models were evaluated using internal and external validation techniques. It is a known fact that the similar sequences or patterns have similar functions, this fact has been utilized in this study to improve the accuracy of binder prediction. In this study, we used Basic Local Alignment Search Tool (BLAST) for sequence similarity search and MERCI software for motif/pattern discovery and searching. The culmination of our work is an ensemble method that combines machine learning and similarity-based approaches, enabling the precise prediction of HLA-DRB1*04:01 binders (Fig. 2).

2. Material and methods

2.1. Dataset Creation and Preprocessing

In this study, we have extracted experimentally validated HLA Class-II allele HLA-DRB1*04:01 binding peptides from IEDB [37]. Initially, the total number of binding peptides extracted from IEDB was 19665, with length varying from 8 to 32 amino acids. After eliminating identical peptides and peptides containing non-natural amino acids, we were left with 12,880 unique peptides. Further analysis of the peptide lengths revealed that 98.4 % (i.e., 12,676 peptides) ranged from 9 to 22 amino acids in length (as depicted in Supplementary Fig. 1). Consequently, we selected these 12,676 peptides to constitute our positive dataset. In such prediction methods, one of the primary challenges is acquiring an experimentally validated negative dataset, which consist of non-binders of HLA-DRB1*04:01 allele. We generated peptides of length 9 to 22 residues from protein in Swiss-Prot database. We randomly picked up 12,676 peptides from Swiss-port generated peptides and assigned them as non-binders. Finally, we create a main dataset that contain of 12,676 experimentally validated binders obtained from IEDB and 12,676 non-binders generated from proteins in Swiss-Prot database.

In addition to main dataset, we also generated an alternate dataset or realistic dataset that contain 12,676 binders and 86,300 non-binders obtained from IEDB. These 86,300 non-binders were obtained from IEDB after removing HLA-DRB1*04:01 binders and peptides that do not have lengths between 9 and 22 amino acids. Both the dataset was further divided into training and independent dataset, where 80 % of the data constitute training and the remaining 20 % make independent dataset [56]. To avoid the biasness in the length distribution in training and independent dataset, we have arranged all peptides as per their length and then transferred every fifth peptide into the independent dataset and rest constitutes the training dataset.

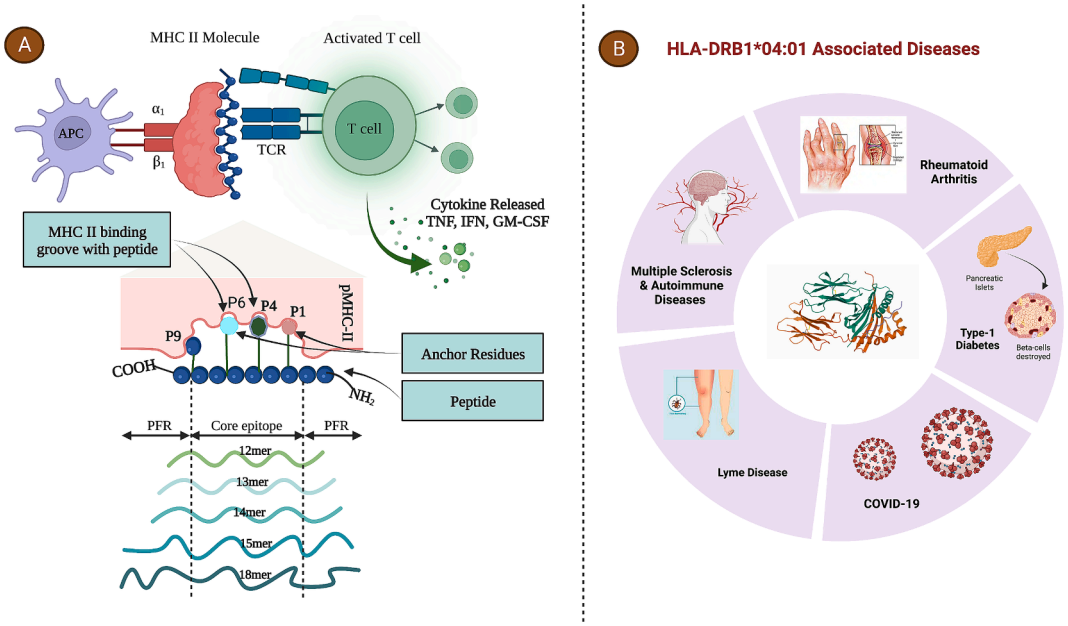


Fig. 1. (A) Pictorial representation of peptide presented by MHC-II molecules to T-cells, anchor residues of peptides bound to the allele-specific pockets of MHC-II; (B) Association of HLA-DRB1*04:01 allele with number of diseases.

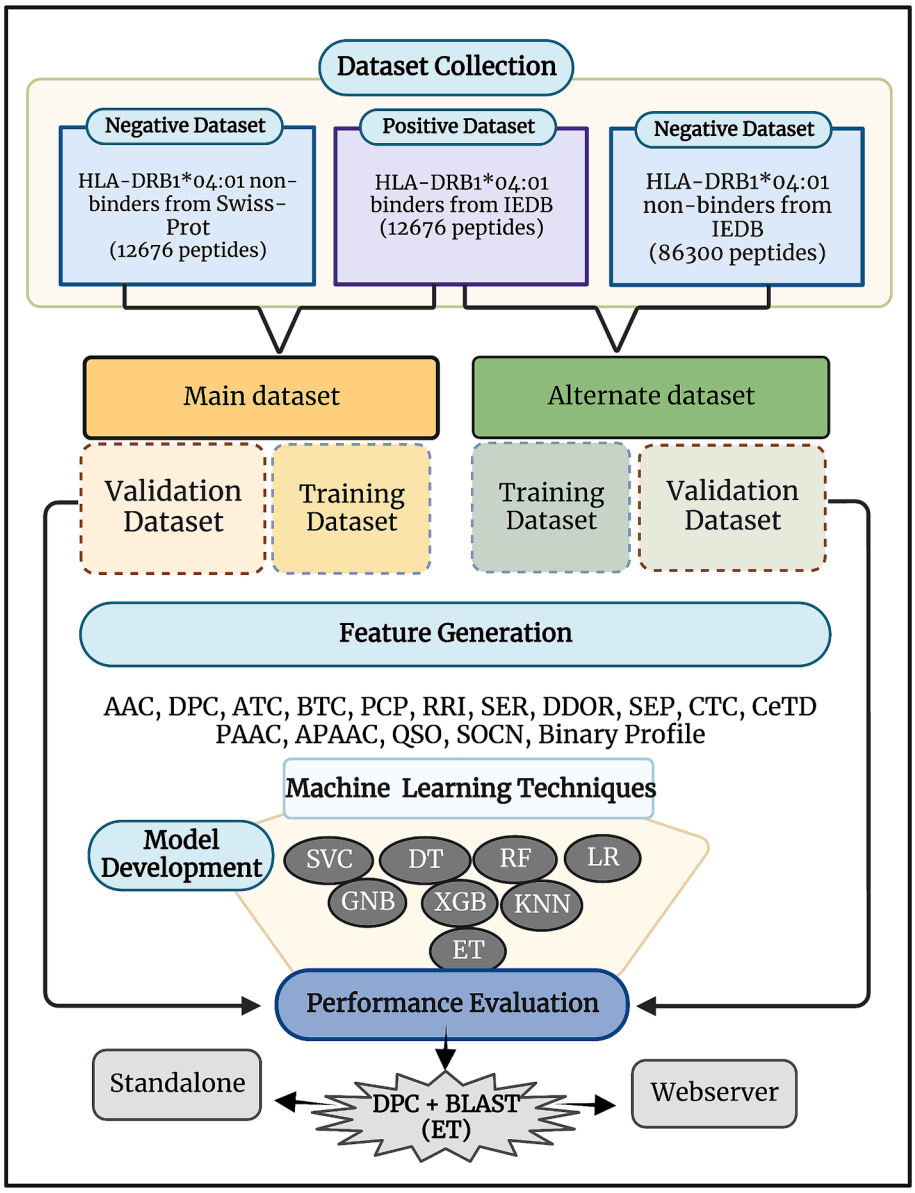


Fig. 2. A flowchart shows overall workflow of algorithm implemented in this study.

2.2. Composition analysis

To check the abundance of each amino acid in each dataset, we calculated the composition of each amino acid using equation (1). We have utilized the amino acid composition module of Pfeature [38] 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.

2.3. Position conservation analysis

In order to explore the position specific preference of residues, we have created the logos using the software named 'Weblogo' [39] web-server. A stack of amino acids that are measured in bits is graphically represented by this software. The overall height of each stack represents the sequence conservation at that position. On the other hand, the height of symbols within a stack reflects the relative frequency of the relevant amino at that particular position [38]. Ensuring the peptide's fix length parameter is a prerequisite for creating a logo. Since, the minimum length of the considered peptide is 9, hence to achieve the fix length criteria, we have taken the first 9 residues from the N-terminal and last 9 residues from the C-terminal. Finally, by concatenating both the regions, we have created a fix length peptide with 18 residues. We have created the "weblogo" for HLA-DRB1*04:01 binders and non-binders in the, main, and alternate dataset.

2.4. Generation of features

In this study to represent the sequence as a numerical vector, we have utilized the composition and binary profile module of Pfeature [38]. By using this tool, we have computed an extensive set of features, including 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 (SEP), 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 computed the 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 profiles (N_9C_9). In Supplementary Table 2, we have reported the length of the vector size generated by composition, and binary profile based features. In Table 1, we have shown the example sequences of different length and highlighted the regions in the sequences which is designated as N_9, C_9 and N_9C_9, respectively.

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 2.

2.5. Model development

In order to train and develop prediction models, we have utilized a diverse set of classifiers using scikit-learn [40] library of python 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).

2.6. Cross-Validation

To build more robust and accurate prediction models, we adopted the five-fold cross-validation technique that avoids overfitting and minimizes the bias in our generated models. Moreover, it also allows us to assess the efficiency of the prediction models. As per the standard norms, we have implemented the five-fold cross validation technique on the training dataset and kept the independent dataset untouched. As per the standard protocols, 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 a testing dataset. Finally, the overall performance is the mean of the performances of five iterations [57].

2.7. Evaluation of parameters

To ensure the fair evaluation of the different models developed using various classifiers, we have used the well-established evaluation parameters. In this method, we implemented both threshold-dependent and threshold-independent parameters. For threshold-dependent parameters, we considered sensitivity, specificity, accuracy, F1-score, kappa, and Mathews correlation coefficient (MCC). Meanwhile, 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:

$$Sensitivity = \frac{TP}{TP + FN} \tag{2}$$

$$Specificity = \frac{TN}{TN + FP} \tag{3}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$F1 - Score = \frac{2TP}{2TP + FP + FN} \tag{5}$$

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

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

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

2.8. 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 parameters for each of our machine learning classifiers and optimised them by maximizing the AUROC.

2.9. Similarity search

In order to predict, if the query peptide is a binder of a HLA-

Table 2
Generation of NC22 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**

Table 1
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**

DRB1*04:01 using similarity search, we have implemented BLAST [41] using the NCBI-blast executable version 2.13.0. We have created the custom database using our training dataset by implementing "make-blastdb" 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 as binder or non-binder.

2.10. 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 [42] with default 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.

2.11. Webserver architecture

We have developed the user-friendly updated version of our old webserver HLADR4Pred and named it as HLA-DR4Pred 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 compatibility was tested and confirmed to be compatible with all the modern devices, including mobile, tablet, laptop, iMac, and desktop. The server is designed with six major modules, including predict, scan, design, blast, motif-scan, and standalone, to address the needs and tasks of various users efficiently.

3. Results

3.1. Composition analysis

In the present study, we have calculated the average composition of each residue in HLA-DRB1*04:01 binders and non-binders in the dataset. The amino acid composition is calculated using Pfeature [38]. The average residue composition for each dataset is provided in Fig. 3, 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 the negative dataset generated randomly using the Swiss-Prot [43] database and general proteome signifies that the negative dataset is not biased towards a particular amino acid or the nature of amino acids.

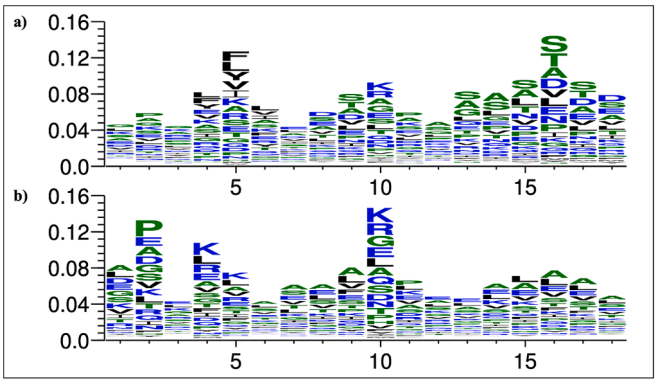


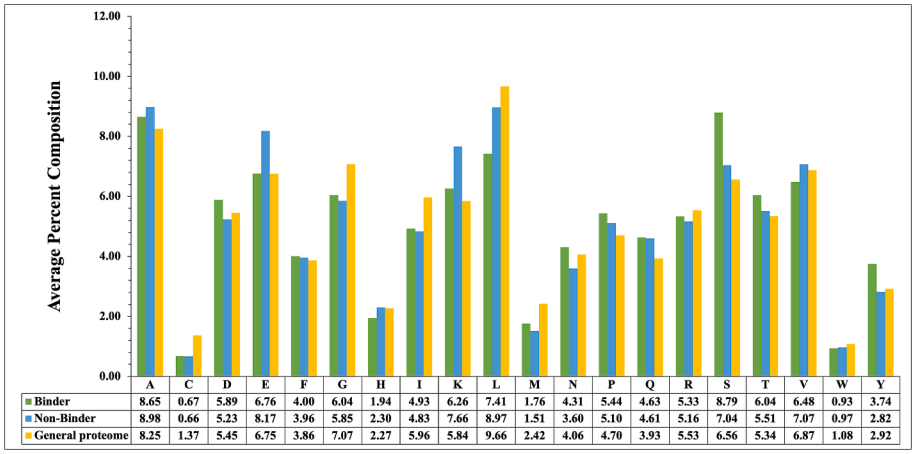
Fig. 4. Positional preference of residues presented by web logo for A) HLA-DRB1*04:01 binders, B) HLA-DRB1*04:01 non-binders.

3.2. Position preference analysis

In this study, the preference of particular residues at a specific position in a peptide was studied by creating the "Weblogo" for HLA-DRB1*04:01 binders and non-binders for main, and alternate dataset, as shown in Fig. 4. In the case of HLA-DRB1*04:01 binders, 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 positively 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 is also preferred by positively charged residues 'K/R'; and positions 14–18 showed abundance for residues 'A/L'.

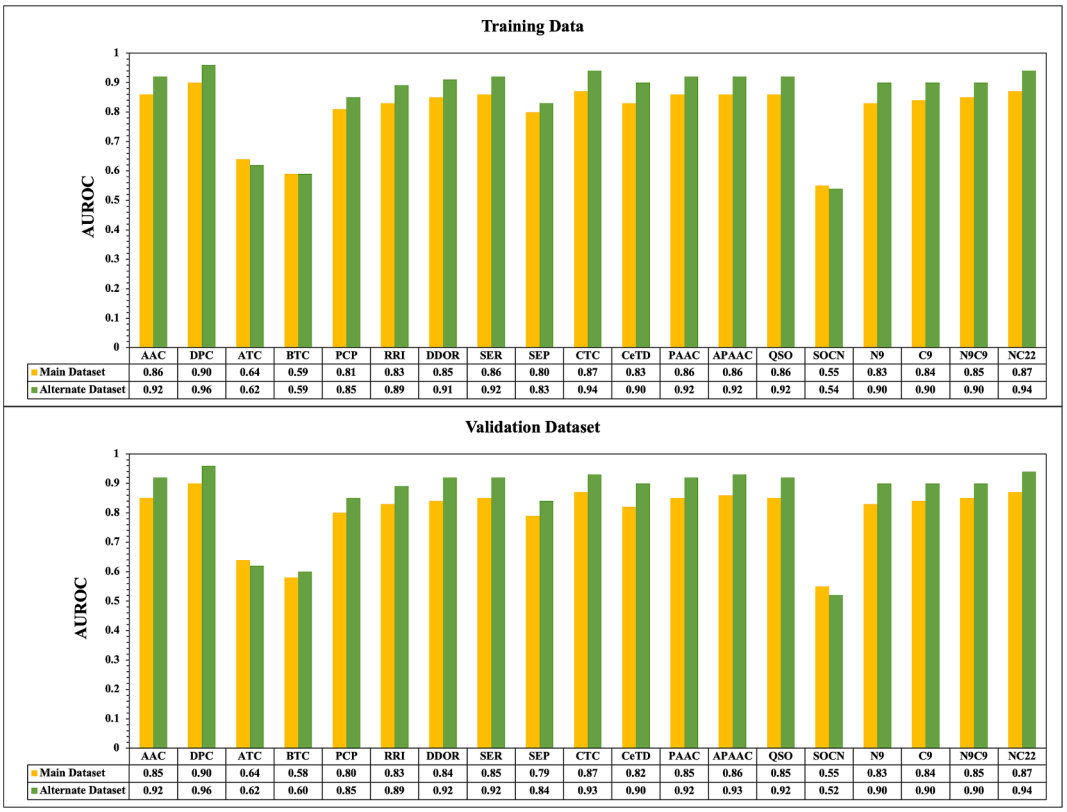
3.3. Performance of models on composition and binary profile based module

We have calculated the fifteen different types of composition features and the binary profiles for different patterns such as N_9, C_9, N_9C_9, and NC_{22} using the Pfeature module [38] to develop the prediction models using eight different classifiers using sklearn [40] 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 externally validated on the independent dataset of the main, and alternate datasets. The comprehensive performance of all the performing models developed on the training and independent dataset



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Binder	8.65	0.67	5.89	6.76	4.00	6.04	1.94	4.93	6.26	7.41	1.76	4.31	5.44	4.63	5.33	8.79	6.04	6.48	0.93	3.74
Non-Binder	8.98	0.66	5.23	8.17	3.96	5.85	2.30	4.83	7.66	8.97	1.51	3.60	5.10	4.61	5.16	7.04	5.51	7.07	0.97	2.82
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	4.70	3.93	5.53	6.56	5.34	6.87	1.08	2.92

Fig. 3. Average percent amino acid composition of HLA-DRB1*04:01 binder, non-binders and general proteome.



	AAC	DPC	ATC	BTC	PCP	RRI	DDOR	SER	SEP	CTC	CeTD	PAAC	APAAC	QSO	SOCN	N9	C9	N9C9	NC22
Main Dataset	0.86	0.90	0.64	0.59	0.81	0.83	0.85	0.86	0.80	0.87	0.83	0.86	0.86	0.86	0.55	0.83	0.84	0.85	0.87
Alternate Dataset	0.92	0.96	0.62	0.59	0.85	0.89	0.91	0.92	0.83	0.94	0.90	0.92	0.92	0.92	0.54	0.90	0.90	0.90	0.94

	AAC	DPC	ATC	BTC	PCP	RRI	DDOR	SER	SEP	CTC	CeTD	PAAC	APAAC	QSO	SOCN	N9	C9	N9C9	NC22
Main Dataset	0.85	0.90	0.64	0.58	0.80	0.83	0.84	0.85	0.79	0.87	0.82	0.85	0.86	0.85	0.55	0.83	0.84	0.85	0.87
Alternate Dataset	0.92	0.96	0.62	0.60	0.85	0.89	0.92	0.92	0.84	0.93	0.90	0.92	0.93	0.92	0.52	0.90	0.90	0.90	0.94

Fig. 5. The performance of ET based classifier developed using fifteen different types of composition based features, and four different types binary profile based features on main, and alternate dataset.

using different types of features is reported in Supplementary Table 3. It was observed that the ET classifier performs best among all the other ML models. As shown in Fig. 5, the ET classifier based model developed on DPC features outperformed all the other models developed on other features, with an AUROC of 0.90 on the independent data of the main dataset; and an AUROC of 0.96 on the independent data of the alternate dataset. CTC based model performed second best with an AUROC of 0.87 on independent data of the main dataset. However, the binary profile pattern NC_{22} also outperformed the other patterns with an AUROC of 0.87 on independent data of the main dataset.

3.4. Performance of models on combined features

Further, we have combined all the features to develop the vector of size 2382 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 independent dataset. As shown in Fig. 6, the ET-based model developed using combined features outperformed all the other classifiers by attaining the maximum AUROC of 0.88 on the independent data of the main dataset. Supplementary Table 4 comprises the threshold-dependent and threshold-independent performance measures of all the classifiers on the main, and alternate datasets.

3.5. Performance of the hybrid model

On observing the performances of various machine learning classifiers on different types of features, it was found that the ET-based model developed on DPC features outperformed all the other features with an AUROC of 0.90 on the independent data of the main dataset. Consequently, in order to enhance the performance, we have integrated the ET-based model of DPC with similarity search using BLAST [41], and call it the hybrid model. We have implemented BLAST by varying the e-values in order to find the optimal e-value, at which we can achieve the maximum AUROC. Table 3 captures the results for each dataset at different e-values for training as well as independent datasets. We varied the e-value from e^{-6} to e^{2}, and the main dataset was able to achieve an AUROC of 0.93 on the independent dataset at an e-value of 1.0, followed by an alternate dataset achieving an AUROC of 0.99 on the independent dataset. Comprehensive results are provided in Supplementary Table 5. The optimal parameters for the DPC-based ET model, identified using the GridSearchCV approach, are presented in Supplementary Table 6. The same model has been implemented at the backend of the server and respective standalone packages.

3.6. Motif analysis

In this study, we have implemented MERCI software [42] with default parameters to obtain the specific regions i.e. motifs from the main dataset, which are highly specific to HLA-DRB1*04:01 binders but absent in non-binder, a similar procedure was repeated for non-binders, where we searched for non-binder specific motifs which are exclusively present in the non-binders and absent in the binder sequences. In Table 4, we have reported the motifs specific to binders and non-binders, along with their coverage in the positive and negative datasets. Residue T, V, F, P, Q, and T are dominant in binders, where residue D, Y, and K cover most of the motifs.

3.7. Comparison with the available methods

It is of utmost importance to evaluate the newly developed method's effectiveness by making comparison with the existing methods, in order to gain the insights into the pros and cons of a newly developed approach. Since, HLADR4Pred2 is an update of HLADR4Pred [32], a comprehensive comparison is required to understand the advantages of the newer version over older versions. Supplementary Table 1

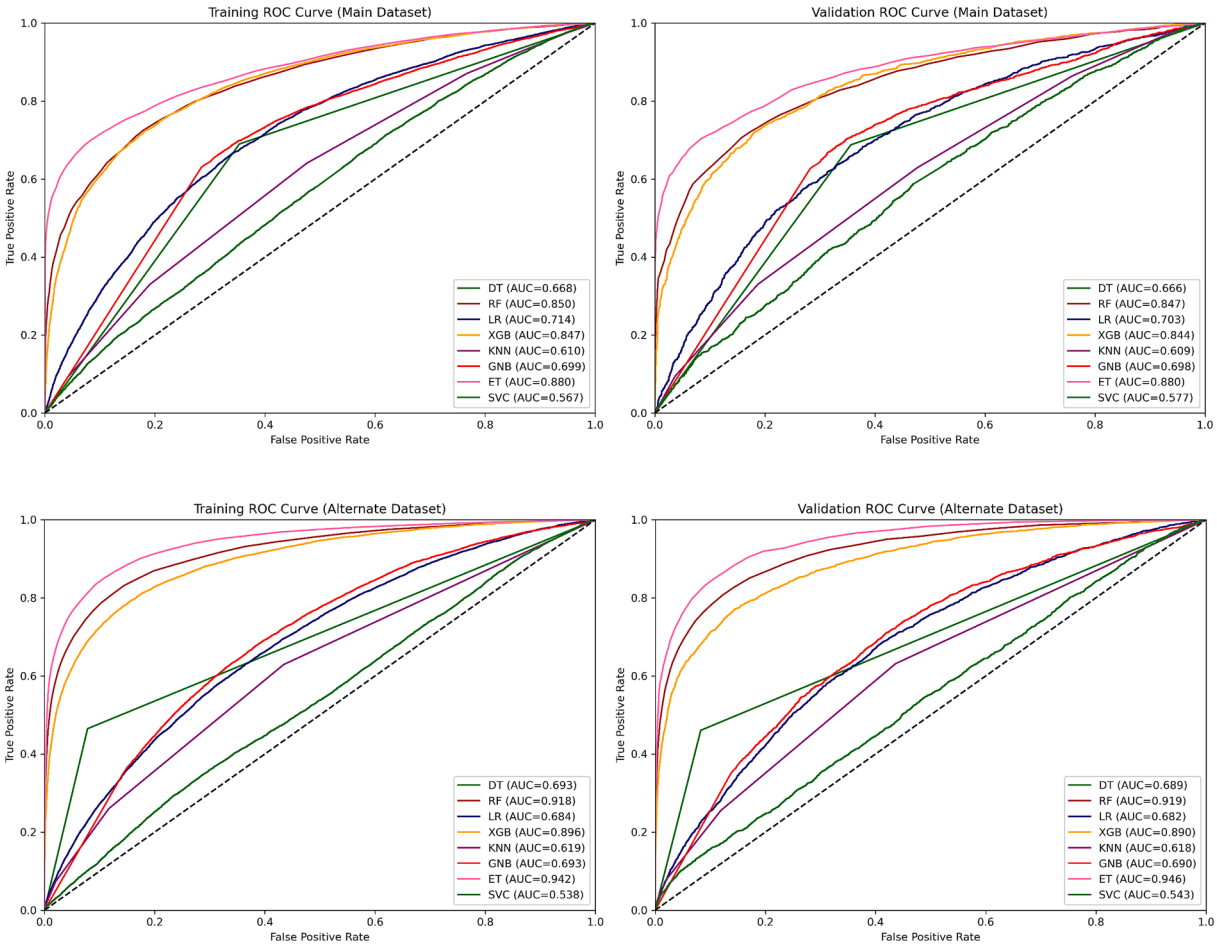


Fig. 6. AUROC score curve showing the performance for all model developed using various classifiers for, main, and alternate dataset on combined features on training and validation dataset.

Table 3
Performance of hybrid model at different e-values for main, and alternate dataset on training and independent dataset.

E-value	Dataset	Main dataset			Alternate dataset		
		Acc	AUROC	MCC	Acc	AUROC	MCC
1.00E-06	**Training**	82.19	0.9	0.64	88.99	0.95	0.64
	Independent	81.79	0.9	0.64	89.58	0.95	0.65
1.00E-05	**Training**	82.25	0.9	0.65	88.35	0.95	0.63
	Independent	82.01	0.91	0.64	89.36	0.95	0.65
1.00E-04	**Training**	82.38	0.9	0.65	88.86	0.95	0.63
	Independent	82.36	0.91	0.65	90.14	0.95	0.66
1.00E-03	**Training**	82.68	0.91	0.65	88.64	0.95	0.63
	Independent	82.67	0.91	0.65	90.13	0.96	0.67
1.00E-02	**Training**	83.22	0.91	0.66	88.57	0.96	0.63
	Independent	83.21	0.92	0.67	90.17	0.96	0.67
1.00E-01	**Training**	84.57	0.92	0.69	88.4	0.96	0.63
	Independent	85.02	0.93	0.7	90.14	0.96	0.67
1.00E + 00	**Training**	**84.77**	**0.92**	**0.70**	**88.4**	**0.98**	**0.63**
	Independent	**85.31**	**0.93**	**0.71**	**90.14**	**0.99**	**0.67**
1.00E + 01	**Training**	84.82	0.92	0.7	88.44	0.96	0.63
	Independent	86.14	0.93	0.72	89.42	0.96	0.65
1.00E + 02	**Training**	84.26	0.92	0.69	89.59	0.96	0.66
	Independent	85	0.93	0.7	89.83	0.96	0.66

*Acc: Accuracy; AUROC: Area under the receiver operating characteristic curve; MCC: Matthews Correlation Coefficient.

accumulates differences in the HLADR4Pred and HLADR4Pred2 at the level of the dataset, implemented features, prediction approach, web-server and standalone. 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 independent dataset and tested the performance of the existing methods on the same. Propred [33] is able to predict the HLA-DRB1*04:01 binding sites and able to achieve 55.26 % accuracy with AUROC 0.74, whereas NetMHCIIpan 4.0 [35,36] achieved accuracy of

Table 4
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	158	158	Y-D-G-K-D	1932	1932
F-T-P-E-T	138	296	R-K-W-E-A	527	2459
T-P-E-T-N	92	388	S-D-H-E	249	2708
D-Q-T-V-I	90	478			
F-V-K-D-Q	225	703			
I-F-T-P-E	180	883			
K-D-Q-T-V-I	90	973			
S-I-F-T-P	270	1198			
V-K-D-Q-T	180	1378			

65.82 % with AUROC 0.72, followed by TEPITOPE [44] with accuracy of 67.75 % with balanced sensitivity and specificity, SMM-align [45] predicts the MHC class-II binding affinity using stabilization matrix-alignment method and achieved accuracy of 67.95 %. Artificial neural network-based method i.e. NNAlign [34], 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, on which the predictions are made, and it attained the accuracy of 69.41 % on the independent dataset. MixMHCpred [46–48] achieves an accuracy of 75.84 % with AUROC 0.83, while TLimmuno2 [49] reached an accuracy of 80.63 % with AUROC 0.85. Finally, the older version of HLADR4Pred2 achieved an 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 methods with an AUROC of 0.93 and an accuracy of 85.31 %. To evaluate the significance of the difference in MCC between HLADR4Pred2.0 and other methods, we applied a Z-test for comparing correlation coefficients. This approach involves transforming the MCC values into Fisher's Z-scores and then calculating the statistical significance of the differences using a two-tailed Z-test. The Z-test results, as shown in Table 5, indicate that the MCC values of HLADR4Pred2.0 are significantly higher than those of all other methods, with p-values well below the 0.05 significance threshold. These results demonstrate that HLADR4Pred2.0 significantly outperforms the other methods in terms of MCC, providing a higher degree of correlation between predicted and actual outcomes. The improvement in MCC highlights the robustness of HLADR4Pred2.0 in accurately classifying binders and non-binders for HLA-DRB1*04:01. These results in Table 5 showed that HLADR4Pred2 is a reliable method which has outperformed the other methods on the independent dataset of the main dataset which was not used while training or testing the model.

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

Recent studies reported that HLA-DRB1*04:01 binding sites are associated with the severity of COVID-19 patients [50–52]. The mutations associated with spike protein in COVID-19 variants can alter the binding of peptides [53,54]. In order to understand the effect of mutations in different variants of COVID-19 with the HLA-binding peptides, we utilized the "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 the Centres for Disease Control and Prevention (CDC portal) (https://www.cdc.gov/hai/data/portal/), the alpha variant possess seven mutation named N501Y, A570D, D613G, P681H, T716I, D981A and D1118H, whereas beta variant incorporated 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, and L981F. Currently, we created the mutated proteins of different variants of COVID-19 and predict the binding peptides and the 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 6.

3.9. Webserver Implementation

We created the user-friendly webserver "HLADR4Pred 2.0" to assist the scientific community which is available at URL 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. The module "PREDICT" allows users to predict the potential of an uncharacterized peptide to be a HLA-DRB1*04:01 binder. The module "SCAN" allows user to provide sequences with length more than 22, which is a constraint in the predict module. In this module, 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 module "DESIGN" 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. In the "BLAST" module, user can make the predictions if a submitted sequence(s) is a binder or non-binder by performing similarity search using BLAST. In the module "MOTIF-SCAN," 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. In the "Standalone" module, we provided both Python- and Perl-based versions of HLADR4Pred2, along with links to the GitHub repository and the Python package index, which is available on PyPI and

Table 5
Comparison of HLADR4Pred2.0 approach with the existing methods on independent dataset of the main dataset.

Methods	Year	Sens	Spec	Acc	AUROC	F1	Kappa	MCC	p-value
Propred	2001	78.378	44.156	55.263	0.735	0.532	0.181	0.219	9.13E-98
Hladr4pred	2004	54.098	79.447	75.036	0.690	0.430	0.279	0.289	2.20E-77
TEPITOPE	2005	68.278	67.336	67.747	NA	0.297	−0.347	0.353	3.34E-60
SMM-Align	2007	68.535	67.495	67.948	NA	0.292	−0.357	0.358	6.40E-59
NNAlign	2017	68.946	68.410	68.643	NA	0.288	−0.367	0.371	1.25E-55
Consensus IEDB	–	69.409	69.404	69.406	NA	0.283	−0.380	0.385	3.75E-52
MixMHCpred	2023	76.000	75.673	75.838	0.829	0.760	0.517	0.517	2.66E-23
TLimmuno2	2023	80.770	80.490	80.630	0.847	0.807	0.613	0.613	4.23E-08
HLADR4Pred2.0	–	88.320	82.300	85.310	0.930	0.860	0.710	0.710	–

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

Table 6
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

can be accessed through the pip package manager.

4. Discussion

The MHC Class-II allele 'HLA-DRB1*04:01′ is a key regulator of immune responses and has been linked to various autoimmune disorders and the severity of COVID-19 [50,51,55]. A study by Kamiza et al., reported that among the 44 alleles of HLA-DRB1 gene; HLA-DRB1*04:01 allele significantly reduced the survival of cervical cancer patients [55]. Interestingly, in the recent vaccine strategies, peptide-based vaccines such as "immunosense" vaccine, emerged as an important tool for activating the immune response. The HLA restricted peptides can be utilized to activate the T-cell responses, offering a promising targeted therapeutic approach for treating both infectious disease and cancer. These HLA-restricted peptides are allele-specific; based on the structure of HLA molecule and peptide-binding groove. Therefore, the peptide's binding affinity may differ for each HLA allele type.

As a results, over the past three decades, numerous methods have been developed for predicting binders of MHC Class-I and Class-II alleles. Despite the tremendous efforts made over the years, the accuracy of these methods remains far from perfect. Hence, there is a pressing need to develop highly accurate methods for predicting MHC binders, particularly for important and clinically relevant alleles. Therefore, there is a crucial need to develop a method, that can predict HLA-DRB1*04:01 binders with high precision. In 2004, we introduced the HLADR4Pred method for predicting HLA-DRB1*04:01 binders, which has since been widely utilized and referenced by the scientific community. In this study, we have undertaken a systematic effort to develop an updated and highly accurate method for predicting HLA-DRB1*04:01 binders. We have extracted experimentally validated 12,676 positive peptides (i.e. HLA-DRB1*04:01 binding peptides and 86,300 negative peptides (non-binding peptides) from IEDB. The positional analysis reveals that in HLA-DRB1*04:01 binding peptides specific amino acids are preferred for instance; hydrophobic residues (L/F/Y/I/V) at 4th, 5th, and 6th; while polar and uncharged amino acids 'S/T/A' preferred at 9th and 10th position. This indicates composition and binary profile based features are extremely important for accurate predictions. Here, we have developed various machine learning models using eight different classifiers such as SVC, DT, RF, XGB, KNN, LR, ET, GNB and SVC. Our results shows ExtraTree based classifier based models outperformed the other classifiers. Specifically, dipeptide composition- based models achieved the highest AUROC of 0.90 on the main dataset, and AUROC of 0.96 on the alternate dataset with an accuracy of more than 82 % on training and independent datasets (Supplementary Table 3). Then after, we aggregate alignment-free (ML-based) and alignment-based (BLAST search) algorithm and attained maximum AUROC 0.93 on the independent dataset of the main dataset. To compare the performance of our model with the existing methods, we have considered different methods, as shown in Table 5, and found out that HLADR4Pred2 has outperformed all the other methods with AUROC of 0.93.

5. Conclusion

HLADR4Pred2 is an improved version of HLADR4Pred, which was trained on a small dataset containing only 576 binders and an equal number of non-binders. Apart from HLADR4Pred, there are a number of other methods with the ability to predict the binders for HLA class-II alleles. Therefore, it is important to benchmark the performance of the other existing methods in comparison to HLADR4Pred2. To achieve this, we have isolated the independent dataset and evaluate the performance of the existing methods on the same. We also reported the motifs specific to HLA-DRB1*04:01 binders and non-binders. The average residue composition shows that serine residue is abundant in HLA-DRB1*04:01 binding peptides in comparison to the non-binding peptides. Additionally, we have developed a web server and standalone package using the best features and classifiers. HLADR4Pred2 incorporates five modules such as PREDICT, SCAN, DESIGN, BLAST, and MOTIF-SCAN. HLADR4Pred2 tool predicts 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

packages is available at https://webs.iiitd.edu.in/raghava/hladr4pred2/standalone.php.

6. Limitation of the study

While HLADR4Pred2 offers substantial improvements in predicting HLA-DRB1*04:01 binders, there are a few limitations that present opportunities for future refinement. One limitation is the model's current focus on the HLA-DRB1*04:01 allele, which restricts its broader applicability to other HLA alleles. Expanding the model to incorporate additional alleles would enhance its utility across a wider range of immunological studies. Furthermore, although the dataset used for training is robust and carefully curated, it may not fully reflect the diversity found in global populations or rare peptide sequences. This could impact the model's generalization to less-studied or novel peptides, particularly in specific clinical or geographic contexts.

Additionally, while HLADR4Pred2 excels in predicting linear peptides, it does not account for conformational epitopes that play a significant role in many immune responses. Incorporating structural information into future versions could further improve the model's predictive power. However, it is important to note that HLADR4Pred2 provides both a web server and standalone options, offering flexibility to a broad range of users. While the model efficiently handles large datasets, future updates could optimize it for extreme cases involving very large or complex peptide datasets. Despite these limitations, HLADR4Pred2 delivers strong predictive accuracy, and further validation on diverse datasets would help solidify its position as a versatile and reliable tool in the field of HLA-peptide binding prediction. These additional validations will ensure the model's robustness and applicability across a broader range of biological contexts.

Author contributions

SP, AD, and GPSR collected and processed the datasets. SP, AD, and GPSR implemented the algorithms and developed the prediction models. AD, SP, and GPSR analysed the results. SP created the back-end and front-end user interface of the web server. SP, NK, AD, and GPSR performed the writing, reviewing and draft preparation of the manuscript. GPSR conceived and coordinated the project. All authors have read and approved the final manuscript.

Funding

The current work has been supported by the Department of Biotechnology (DBT) grant BT/PR40158/BTIS/137/24/2021.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Authors are thankful to the University Grants Commission (UGC), Department of Science and Technology, and Department of Biotechnology (DBT) for fellowships and financial support, and the Department of Computational Biology, IIITD, New Delhi for infrastructure and facilities. We would like to acknowledge that the Figures were created using BioRender.com.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ymeth.2024.10.007.

Data availability

All the datasets used in this study are available at the "HLA-DR4Pred 2.0" web server, https://webs.iiitd.edu.in/raghava/hladr4pred2/dataset.php.

References

[1] N.B. Crux, S. Elahi, 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. 8 (2017) 832, https://doi.org/10.3389/fimmu.2017.00832.
[2] T. Shiina, K. Hosomichi, H. Inoko, J.K. Kulski, The HLA genomic loci map: expression, interaction, diversity and disease, J. Hum. Genet. 54 (2009) 15–39, https://doi.org/10.1038/jhg.2008.5.
[3] S.Y. Choo, The HLA system: genetics, immunology, clinical testing, and clinical implications, Yonsei Med. J. 48 (2007) 11–23, https://doi.org/10.3349/ymj.2007.48.1.11.
[4] M. Wang, M.H. Claesson, Classification of human leukocyte antigen (HLA) supertypes, Methods Mol. Biol. 1184 (2014) 309–317, https://doi.org/10.1007/978-1-4939-1115-8_17.
[5] J. Robinson, D.J. Barker, X. Georgiou, M.A. Cooper, P. Flicek, S.G.E. Marsh, IPD-IMGT/HLA database, Nucleic Acids Res. 48 (D1) (2020) D948–D955, https://doi.org/10.1093/nar/gkz950.
[6] Y.M. Mosaad, Clinical role of human leukocyte antigen in health and disease, Scand. J. Immunol. 82 (2015) 283–306, https://doi.org/10.1111/sji.12329.
[7] D. Zheng, T. Liwinski, E. Elinav, Interaction between microbiota and immunity in health and disease, Cell Res. 30 (2020) 492–506, https://doi.org/10.1038/s41422-020-0332-7.
[8] P. Leone, E.C. Shin, F. Perosa, A. Vacca, F. Dammacco, V. Racanelli, MHC class I antigen processing and presenting machinery: organization, function, and defects in tumor cells, J. Natl Cancer Inst. 105 (2013) 1172–1187, https://doi.org/10.1093/jnci/djt184.
[9] L.N. Adler, W. Jiang, K. Bhamidipati, M. Millican, C. Macaubas, S.C. Hung, et al., The other function: class II-restricted antigen presentation by B cells, Front. Immunol. 8 (2017) 319, https://doi.org/10.3389/fimmu.2017.00319.
[10] J.L. Sanchez-Trincado, M. Gomez-Perosanz, P.A. Reche, Fundamentals and methods for T- and B-cell Epitope prediction, J. Immunol. Res. 2017 (2017) 2680160, https://doi.org/10.1155/2017/2680160.
[11] C.J. Holland, D.K. Cole, A. Godkin, Re-directing CD4(+) T cell responses with the flanking residues of MHC class II-bound peptides: the core is not enough, Front. Immunol. 4 (2013) 172, https://doi.org/10.3389/fimmu.2013.00172.
[12] M. Wieczorek, E.T. Abualrous, J. Sticht, M. Álvaro-Benito, S. Stolzenberg, F. Noé, et al., Major histocompatibility complex (MHC) class I and MHC class II proteins: conformational plasticity in antigen presentation, Front. Immunol. 8 (2017) 292, https://doi.org/10.3389/fimmu.2017.00292.
[13] M. Nielsen, O. Lund, S. Buus, C. Lundegaard, MHC class II epitope predictive algorithms, Immunology 130 (2010) 319–328, https://doi.org/10.1111/j.1365-2567.2010.03268.x.
[14] K.L. Rock, E. Reits, J. Neefjes, Present yourself! By MHC class I and MHC class II molecules, Trends Immunol. 37 (2016) 724–737, https://doi.org/10.1016/j.it.2016.08.010.
[15] G.M. Dunston, R.M. Halder, Vitiligo is associated with HLA-DR4 in black patients. A preliminary report, Arch. Dermatol. 126 (1990) 56–60, https://doi.org/10.1001/archderm.1990.01670250062009.
[16] J.D. Taurog, HLA-DR4 and the spondyloarthropathies, Ann. Rheum. Dis. 61 (2002) 193–194, https://doi.org/10.1136/ard.61.3.193.
[17] T. Shi, W. Lv, L. Zhang, J. Chen, H. Chen, Association of HLA-DR4/HLA-DRB1*04 with Vogt-Koyanagi-Harada disease: a systematic review and meta-analysis, Sci. Rep. 4 (2014) 6887, https://doi.org/10.1038/srep06887.
[18] P. Stastny, E.J. Ball, M.A. Khan, N.J. Olsen, T. Pincus, X. Gao, HLA-DR4 and other genetic markers in rheumatoid arthritis, Br. J. Rheumatol. 27 (Suppl 2) (1988) 132–138, https://doi.org/10.1093/rheumatology/xxvii.suppl_2.132.
[19] D. Brassat, G. Salemi, L.F. Barcellos, G. McNeill, P. Proia, S.L. Hauser, et al., The HLA locus and multiple sclerosis in Sicily, Neurology 64 (2005) 361–363, https://doi.org/10.1212/01.wnl.0000149765.71212.0a.
[20] S. Hoffmann, S. Cepok, V. Grummel, K. Lehmann-Horn, J. Hackermüller, P. F. Stadler, 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. 83 (2008) 219–227, https://doi.org/10.1016/j.ajhg.2008.07.006.
[21] S. Muñiz-Castrillo, A. Vogrig, J. Honnorat, Associations between HLA and autoimmune neurological diseases with autoantibodies, Auto Immun. Highlights. 11 (1) (2020) 2, https://doi.org/10.1186/s13317-019-0124-6.
[22] C.E. Larsen, C.A. Alper, The genetics of HLA-associated disease, Curr. Opin. Immunol. 16 (2004) 660–667, https://doi.org/10.1016/j.coi.2004.07.014.
[23] L. Kovalchuka, J. Eglite, I. Lucenko, M. Zalite, L. Viksna, A. Krumiņa, Associations of HLA DR and DQ molecules with Lyme borreliosis in Latvian patients, BMC. Res. Notes 5 (2012) 438, https://doi.org/10.1186/1756-0500-5-438.
[24] J.L. Newton, S.M.J. Harney, B.P. Wordsworth, M.A. Brown, A review of the MHC genetics of rheumatoid arthritis, Genes Immun. 5 (2004) 151–157, https://doi.org/10.1038/sj.gene.6364045.
[25] B.I. Yamout, R. Alroughani, Multiple sclerosis, Semin. Neurol. 38 (2018) 212–225, https://doi.org/10.1055/s-0038-1649502.

[26] D.M. Maahs, N.A. West, J.M. Lawrence, E.J. Mayer-Davis, Epidemiology of type 1 diabetes, Endocrinol. Metab. Clin. North Am. 39 (2010) 481–497, https://doi.org/10.1016/j.ecl.2010.05.011.
[27] K.M. Gillespie, Type 1 diabetes: pathogenesis and prevention, CMAJ 175 (2006) 165–170, https://doi.org/10.1503/cmaj.060244.
[28] B. McIver, J.C. Morris, The pathogenesis of Graves' disease, Endocrinol. Metab. Clin. North Am. 27 (1998) 73–89, https://doi.org/10.1016/s0889-8529(05)70299-1.
[29] H. Khan, A. Sureda, T. Belwal, S. Çetinkaya, İ. Süntar, S. Tejada, et al., Polyphenols in the treatment of autoimmune diseases, Autoimmun. Rev. 18 (2019) 647–657, https://doi.org/10.1016/j.autrev.2019.05.001.
[30] C. Lundegaard, O. Lund, S. Buus, M. Nielsen, Major histocompatibility complex class I binding predictions as a tool in epitope discovery, Immunology 130 (2010) 309–318, https://doi.org/10.1111/j.1365-2567.2010.03300.x.
[31] M. Nielsen, C. Lundegaard, T. Blicher, B. Peters, A. Sette, S. Justesen, et al., Quantitative predictions of peptide binding to any HLA-DR molecule of known sequence: NetMHCIIpan, PLoS Comput. Biol. 4 (2008) e1000107, https://doi.org/10.1371/journal.pcbi.1000107.
[32] M. Bhasin, G.P.S. Raghava, SVM based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence, Bioinformatics 20 (2004) 421–423, https://doi.org/10.1093/bioinformatics/btg424.
[33] H. Singh, G.P. Raghava, ProPred: prediction of HLA-DR binding sites, Bioinformatics 17 (2001) 1236–1237, https://doi.org/10.1093/bioinformatics/17.12.1236.
[34] M. Nielsen, M. Andreatta, NNAlign: a platform to construct and evaluate artificial neural network models of receptor-ligand interactions, Nucleic Acids Res. 45 (2017) W344–W349, https://doi.org/10.1093/nar/gkx276.
[35] B. Reynisson, B. Alvarez, S. Paul, B. Peters, M. Nielsen, 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. 48 (2020) W449–W454, https://doi.org/10.1093/nar/gkaa379.
[36] E. Karosiene, M. Rasmussen, T. Blicher, O. Lund, S. Buus, M. Nielsen, 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 65 (2013) 711–724, https://doi.org/10.1007/s00251-013-0720-y.
[37] R. Vita, S. Mahajan, J.A. Overton, S.K. Dhanda, S. Martini, J.R. Cantrell, et al., The immune epitope database (IEDB): 2018 update, Nucleic Acids Res. 47 (2019) D339–D343, https://doi.org/10.1093/nar/gky1006.
[38] A. Pande, S. Patiyal, A. Lathwal, C. Arora, D. Kaur, A. Dhall, et al., Pfeature: a tool for computing wide range of protein features and building prediction models, J. Comput. Biol. 30 (2023) 204–222, https://doi.org/10.1089/cmb.2022.0241.
[39] G.E. Crooks, G. Hon, J.M. Chandonia, S.E. Brenner, WebLogo: a sequence logo generator, Genome Res. 14 (2004) 1188–1190, https://doi.org/10.1101/gr.849004.
[40] G. Varoquaux, L. Buitinck, G. Louppe, O. Grisel, F. Pedregosa, A. Mueller, Scikit-learn, GetMob Mob. Comput. Commun. 19 (2015) 29–33, https://doi.org/10.1145/2786984.2786995.
[41] S. McGinnis, T.L. Madden, BLAST: at the core of a powerful and diverse set of sequence analysis tools, Nucleic Acids Res. 32 (Web Server issue) (2004) W20–W25, https://doi.org/10.1093/nar/gkh435.
[42] C. Vens, M.N. Rosso, E.G.J. Danchin, Identifying discriminative classification-based motifs in biological sequences, Bioinformatics 27 (2011) 1231–1238, https://doi.org/10.1093/bioinformatics/btr110.
[43] A. Bairoch, The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000, Nucleic Acids Res. 28 (2000) 45–48, https://doi.org/10.1093/nar/28.1.45.
[44] O. Karpenko, J. Shi, Y. Dai, Prediction of MHC class II binders using the ant colony search strategy, Artif. Intell. Med. 35 (2005) 147–156, https://doi.org/10.1016/j.artmed.2005.02.002.
[45] M. Nielsen, C. Lundegaard, O. Lund, Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method, BMC Bioinf. 9 (2007) 238, https://doi.org/10.1186/1471-2105-8-238.
[46] M. Bassani-Sternberg, C. Chong, P. Guillaume, M. Solleder, H. Pak, P.O. Gannon, et al., Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allostery regulating HLA specificity, PLoS Comput. Biol. 13 (2017) e1005725, https://doi.org/10.1371/journal.pcbi.1005725.
[47] D. Gfeller, P. Guillaume, J. Michaux, H.S. Pak, R.T. Daniel, J. Racle, et al., The length distribution and multiple specificity of naturally presented HLA-I ligands, J. Immunol. 201 (2018) 3705–3716, https://doi.org/10.4049/jimmunol.1800914.
[48] J. Racle, P. Guillaume, J. Schmidt, J. Michaux, A. Larabi, K. Lau, et al., Machine learning predictions of MHC-II specificities reveal alternative binding mode of class II epitopes, Immunity 56 (2023) 1359–1375.e13, https://doi.org/10.1016/j.immuni.2023.03.009.
[49] G. Wang, T. Wu, W. Ning, K. Diao, X. Sun, J. Wang, et al., TLimmuno2: predicting MHC class II antigen immunogenicity through transfer learning, Brief. Bioinform. 24 (3) (2023) bbad116, https://doi.org/10.1093/bib/bbad116.
[50] S. Ebrahimi, H.R. Ghasemi-Basir, M.M. Majzoobi, A. Rasouli-Saravani, M. Hajilooi, G. Solgi, HLA-DRB1*04 may predict the severity of disease in a group of Iranian COVID-19 patients, Hum. Immunol. 82 (2021) 719–725, https://doi.org/10.1016/j.humimm.2021.07.004.
[51] E. de Sousa, D. Ligeiro, J.R. Lérias, C. Zhang, C. Agrati, M. Osman, 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. 98 (2020) 454–459, https://doi.org/10.1016/j.ijid.2020.07.016.
[52] D.J. Langton, S.C. Bourke, B.A. Lie, G. Reiff, S. Natu, R. Darlay, et al., The influence of HLA genotype on the severity of COVID-19 infection, HLA 98 (2021) 14–22, https://doi.org/10.1111/tan.14284.
[53] W.T. Harvey, A.M. Carabelli, B. Jackson, R.K. Gupta, E.C. Thomson, E.M. Harrison, et al., SARS-CoV-2 variants, spike mutations and immune escape, Nat. Rev. Microbiol. 19 (2021) 409–424, https://doi.org/10.1038/s41579-021-00573-0.
[54] B. Korber, W.M. Fischer, S. Gnanakaran, H. Yoon, J. Theiler, W. Abfalterer, et al., Tracking changes in SARS-CoV-2 Spike: evidence that D614G increases infectivity of the COVID-19 virus, Cell 182 (2020) 812–827.e19, https://doi.org/10.1016/j.cell.2020.06.043.
[55] A.B. Kamiza, S. Kamiza, C.G. Mathew, HLA-DRB1 alleles and cervical cancer: a meta-analysis of 36 case-control studies, Cancer Epidemiol. 67 (2020) 101748, https://doi.org/10.1016/j.canep.2020.101748.
[56] P. Razzaghi, K. Abbasi, J.B. Ghasemi, Multivariate pattern recognition by machine learning methods, in: Machine Learning and Pattern Recognition Methods in Chemistry from Multivariate and Data Driven Modeling, Elsevier, 2023, pp. 47–72. https://doi.org/10.1016/b978-0-323-90408-7.00002-2.
[57] A. Gharizadeh, K. Abbasi, A. Ghareyazi, M.R.K. Mofrad, H.R. Rabiee, HGTDR: Advancing drug repurposing with heterogeneous graph transformers, Bioinformatics 40 (2024) btae349. https://doi.org/10.1093/bioinformatics/btae349.

Sumeet Patiyal is currently working as a Ph.D. in Computational biology from the Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India

AnjaliDhall is currently working as a Ph.D. in Computational Biology from the Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India.

NishantKumar is currently working as a Ph.D. in Computational biology from the Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India

GajendraP.S.Raghava is currently working as a Professor and Head of the Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India.