Computers in Biology and Medicine 170 (2024) 108083

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

A method for predicting linear and conformational B-cell epitopes in an antigen from its primary sequence

Nishant Kumar [1], Sadhana Tripathi [1], Neelam Sharma [1], Sumeet Patiyal [1], Naorem Leimarembi Devi, Gajendra P.S. Raghava [*], [2]

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

ARTICLE INFO

Keywords:
Conformational B-Cell epitopes
Linear
Alignment-based models
Machine learning techniques
Random forest

ABSTRACT

B-cell is an essential component of the immune system that plays a vital role in providing the immune response against any pathogenic infection by producing antibodies. Existing methods either predict linear or conformational B-cell epitopes in an antigen. In this study, a single method was developed for predicting both types (linear/conformational) of B-cell epitopes. The dataset used in this study contains 3875 B-cell epitopes and 3996 non-B-cell epitopes, where B-cell epitopes consist of both linear and conformational B-cell epitopes. Our primary analysis indicates that certain residues (like Asp, Glu, Lys, and Asn) are more prominent in B-cell epitopes. We developed machine-learning based methods using different types of sequence composition and achieved the highest AUROC of 0.80 using dipeptide composition. In addition, models were developed on selected features, but no further improvement was observed. Our similarity-based method implemented using BLAST shows a high probability of correct prediction with poor sensitivity. Finally, we developed a hybrid model that combines alignment-free (dipeptide based random forest model) and alignment-based (BLAST-based similarity) models. Our hybrid model attained a maximum AUROC of 0.83 with an MCC of 0.49 on the independent dataset. Our hybrid model performs better than existing methods on an independent dataset used in this study. All models were trained and tested on 80 % of the data using a cross-validation technique, and the final model was evaluated on 20 % of the data, called an independent or validation dataset. A webserver and standalone package named “CLBTope” has been developed for predicting, designing, and scanning B-cell epitopes in an antigen sequence available at (https://webs.iiitd.edu.in/raghava/clbtope/).

1. Introduction

The immune system is an interactive and interconnected network of cells, tissues, and organs that work together to defend the host against foreign invaders [1]. It recognizes and responds to a wide range of disease-causing pathogens, such as parasites, fungi, bacteria, and viruses. The immune system can be subdivided into innate and adaptive immunity [2]. Innate immune response provides the first line of protection against invasive pathogens and activates the adaptive immune system. It is a non-specific, immediate reaction that does not produce any immunological memory [3]. On the other hand, adaptive immunity provides antigen-specific defense and generates lifelong immunological memory [4]. It is further composed of B and T-lymphocytes, which identify antigens with distinct specificity. B cells stimulate humoral immunity, whereas T lymphocytes stimulate cell-mediated immunity [5].

Antibodies secreted by B-cells play an important role in the immune response. It specifically binds to antigens at the antigenic determinant site, also known as epitopes. An epitope is a part of proteins that are present on the surface of the antigen. It interacts with the antibody through the B-cell receptor and evokes either cellular or humoral immunological response [6]. Based on the structural characteristics of epitopes and their interaction with antibodies, they can be classified into linear (or continuous) or conformational (discontinuous) epitopes [7]. Linear epitopes are a stretch of continuous residues of amino acids required for binding, whereas conformational epitopes contain amino acids that are far apart from each other but come in close proximity due

[*] Corresponding author. Okhla Industrial Estate, Phase III (Near Govind Puri Metro Station), New Delhi, 110020, Office: A-302 (R&D Block), India..
E-mail addresses: nishantk@iiitd.ac.in (N. Kumar), sadhana20214@iiitd.ac.in (S. Tripathi), neelams@iiitd.ac.in (N. Sharma), sumeetp@iiitd.ac.in (S. Patiyal), leimarembi@gmail.com (N.L. Devi), raghava@iiitd.ac.in (G.P.S. Raghava).
[1] Equal Contribution.
[2] http://webs.iiitd.edu.in/raghava

https://doi.org/10.1016/j.compbiomed.2024.108083
Received 21 June 2023; Received in revised form 26 December 2023; Accepted 27 January 2024
Available online 28 January 2024

Abbreviations

AAC	Amino acid composition
AUROC:	Area Under the Receiver Operating Characteristic
BCEPS	B Cell Epitope Prediction Software
BLAST	Basic Local Alignment Search Tool
DPC	Dipeptide composition
DT	Decision Tree
ET	Extra Tree classifier
GNB	Gaussian Naive Bayes
IDED	IEDB Discontinuous Epitope Dataset
IEDB	Immune Epitope Database
ILED	IEDB Linear Epitope dataset
KNN	k-Nearest Neighbor
LR	Logistic Regression
MCC	Matthew's correlation coefficient
ML:	Machine learning
mRMR	minimum redundancy - maximum relevance
RF	Random Forest
SVC	Support Vector Classifier
XGB	XGBoost

to the polypeptide folding [8]. B-cell epitopes provide a great deal of potential for immunology-related applications, such as the development of epitope-based vaccines, therapeutic antibodies, and disease diagnosis [9,10]. Besides, computational methods go hand in hand with experimental techniques to provide assistance in the identification/prediction of B-cell epitopes which otherwise would be an expensive, time-consuming, and labour-intensive procedure [11].

To predict B-cell epitopes, several computational methods have been developed in the past [12]. These methods can be divided into conformational, linear, and linearized conformational B-cell epitopes. A region/segment on the surface of a protein structure recognized by B-cell receptors is called a conformational or discontinuous B-cell epitope. Numerous methods have been developed to predict conformational B-cell epitopes from protein structures like CEP, DiscoTope, SEPPA, Epitope3D, and SEMA [13–17]. These conformational methods need the tertiary structure of a protein to identify conformational B-cell epitopes in a protein. Thus, the prediction of the tertiary structure of a protein is one of the major bottlenecks in the implementation of conformational methods. A single continuous stretch of amino acids within a protein sequence that can be reacted with anti-protein antibodies is predicted by most modern methods, known as linear/continuous B-cell epitope prediction methods. Initially, linear methods were developed using physicochemical properties [18,19], but most of them have very poor performance [20]. In 2006, machine learning methods were developed to advance the performance of epitope prediction, including ABCpred, BepiPred, and BCPREDS [21–23]. These methods were developed using small datasets mainly extracted from BCIPEP, AntiJen 2.0, and HIVDB. The next-generation methodologies involved the extraction of datasets from IEDB. These methods extensively utilized machine learning and deep learning approaches to predict continuous B-cell epitopes [24–26]. Additionally, specific methodologies were developed to predict conformational epitopes, specifically antibody-interacting residues in an antigen from its primary sequence [27,28]. The subsequent wave of methodologies involved the extraction of datasets from IEDB. These methods extensively utilized machine learning and deep learning approaches to predict continuous B-cell epitopes [24–26]. Additionally, specific methodologies were developed to forecast conformational epitopes, specifically targeting antibody-interacting residues within an antigen, derived solely from its primary sequence [27,28].

Recently, linearized conformational epitopes were identified from antibody-antigen structures for developing epitope prediction [29]. A strategy to predict both linear and conformational B-cell epitopes has been developed in this study. The dataset used in this study contains linear and linearized conformational B-cell epitopes. Firstly, we developed machine learning based models using the composition of epitope/non-epitope sequences. Secondly, we develop an alignment-based approach using BLAST; a peptide is assigned an epitope if it has high similarity with known B-cell epitopes. It was observed that both techniques have their own limitation. Finally, we created a hybrid model using both alignment-based and alignment-free methods. For the benefit of the scientific community, we developed a standalone software (https://webs.iiitd.edu.in/raghava/clbtope/stand.html) and webserver (https://webs.iiitd.edu.in/raghava/clbtope/) that allow users to predict B-cell epitopes.

2. Material and methods

2.1. Data acquisition

We derived most of the datasets for this study from the following datasets $BCETD_{555}$, $ILED_{2195,}$ and $IDED_{1246}$, obtained from Ras-Carmona et al. [29] a study in which they reduced the sequence redundancy in all datasets using the CD-HIT so that amino acid sequence identity was <80 %. $BCETD_{555}$ is a non-redundant dataset containing linearized conformational 555 B-cell epitopes and 555 non B-cell epitopes retrieved from antibody-antigen structure complexes. $ILED_{2195}$ contains 2195 linear B-cell epitopes and 2195 non-epitopes extracted from the Immune Epitope Database (IEDB). Another conformational dataset $IDED_{1246}$ obtained from IEDB contains 1246 discontinuous B-cell epitopes and 1246 non-epitopes. We merged all three datasets and created a new dataset that contains 7992 epitopes (3996 B-cell epitopes and 3996 non-B-cell epitopes). We have also considered the length of peptide sequences while distributing the dataset. To enhance the quality of the dataset, we have removed the peptides comprising unnatural amino acids 'BJOUXZ' and duplicate sequences. Hence, the final dataset comprises 7871 epitopes (3875 B-cell and 3996 non-B-cell epitopes). Our B-cell epitopes have both types of B-cell epitopes, conformational and linear/continuous.

2.2. Composition analysis

Amino acid composition (AAC) represents the percentage frequency of 20 amino acids within peptide/protein sequences. It is a feature vector of 20 elements in which each component of the sequence represents a fraction of a specific type of amino acid residue in the sequence. The following equation is used to calculate the composition:

$$AAC_i = \frac{R_i}{L} \times 100 \quad (1)$$

Where AAC_i is the percent composition of amino acid i, R_i is the number of residues of type i, and L is the total number of residues in the peptide [30–33].

2.3. Sequence logo

Sequence logos are generated to display sequence conservation patterns that provide a more detailed and accurate explanation of sequence similarity than consensus sequences. We used a web-based application, "weblogo" to generate sequence logos. It generates a graphical representation of a stack of amino acids measured in bits. Each stack's overall height reveals sequence conversation at that place. Furthermore, the relative frequency of the relevant amino acid or nucleic acid at that place in the sequences is indicated by the height of the symbols within the stack [34].

2.4. Feature generation

To develop an effective epitope predictor, it is necessary to transform the input peptide sequences into a set of numerical vectors/features representing the properties of the peptide sequence [35]. The extracted features must preserve the peptide's sequence information to the greatest extent possible and reflect intrinsic correlation with the peptide classification [32]. We used the standalone software of 'Pfeature' [36] to extract features from the peptide sequence. By using the composition module, we have calculated the ten different types of features such as amino acid composition (AAC), dipeptide composition (DPC), atomic composition (ATC), physico-chemical properties based composition (PCP), residue repeat information (RRI), Shannon entropy for all residues (SER), composition enhanced transition and distribution (CeTD), pseudo amino acid composition (PAAC), amphiphilic pseudo amino acid composition (APAAC), and quasi-sequence order (QSO). As indicated in Table 1, we created a vector of 780 features in this investigation using a composition-based feature module.

2.5. Motif based features

A motif is a repeated pattern of sequences in a set of peptides or proteins. Identifying high-class motifs for positive dataset characterization involves comparing positive dataset motifs with those of negative datasets [31]. In this study, we apply the classification scheme proposed by Koolman and Rohm (1996), which contains seven non-overlapping properties aliphatic (A, G, I, L, V), sulfur (C, M), aromatic (F, Y, W), neutral (S, T, N, Q), acidic (D, E), basic (R, H, K), and, Cyclic (P) [37].

2.6. Feature selection

Feature selection is a crucial stage in classification. In most datasets, only some features contribute to determining the endpoint; some are redundant or noisy. "Pfeature" generates a significant vector of features. It is a major challenge to identify the relationship between the features of the data. The ultimate objective of feature selection is to choose a proper subset of features from the initial feature set that can reduce the likelihood of over-fitting and boost effectiveness by simplifying the model [33,35,38–40]. There are several feature selection techniques available in the literature. In this study, we have implemented the mRMR (minimum redundancy-maximum relevance) feature selection technique. A primary concern of this study is the ability to generate better models as fast as possible.

2.6.1. Machine learning

Machine learning (ML) algorithms are powerful and potent techniques that are capable of transforming data into reliable decisions. We implement several machine learning algorithms [40] to develop a powerful predictor that can efficiently and accurately predict the B-cell epitopes. Our study adopts supervised classification algorithms, where each ML algorithm has a unique set of hyperparameters that should be tuned and modified to identify the optimal combination and, as a result, the best prediction and solution to the problem [33]. One primary task is choosing an appropriate classification algorithm [35]. We have used the following ML algorithms: Decision Tree (DT), Random Forest (RF), Gaussian Naive Bayes (GNB), Logistic Regression (LR), XGBoost (XGB), k-nearest neighbors (KNNs), and Extra-Trees (ET) classifier [41]. We must carefully train the ML algorithm to classify unknown datasets accurately.

Table 1
List of descriptors with a brief description and length of the vector.

S. No	Type of Descriptor	Length of the vector
1.	AAC (Amino acid composition)	20
2.	DPC (Dipeptide composition)	400
3.	APAAC (Amphiphilic pseudo amino acid composition)	29
4.	ATC (Atomic composition)	5
5.	CETD (Composition-enhanced transition distribution)	187
6.	SER (Shannon entropy of residues)	20
7.	PAAC (Pseudo amino acid composition)	23
8.	PCP (Physico-chemical properties composition)	30
9.	QSO (Quasi-sequence order)	46
10.	RRI (Residue repeat Information)	20
	Combined Features	780

2.7. Model's evaluation

We have implemented a five-fold cross-validation algorithm to evaluate our model. As per the standard protocols, this technique splits data randomly into k-folds. We implemented a cross-validation technique with k = 5, then we trained our model on four folds, and the rest of one fold was used to test the model, and the same cycle was repeated five times. In the results, we have shown the average scores of all repetitions [40,41].

2.8. Similarity search

For protein/peptide annotation, a well-known similarity search-based technique called "BLAST" was used. In this technique, all peptide sequences whose functions are known are aligned with the query peptide sequence. The query peptide is annotated based on its alignment score with known peptides. Our study implemented the BLAST-based technique blastp (BLAST+ 2.7.1), a peptide-peptide BLAST, to predict B-cell epitopes and non B-cell epitopes [42–45]. BLAST formatted database was constructed using the training dataset against which the query sequences (sequences in the test set) were hit at various e-values that range from 1e-6 to 1e+3. Based on the top hits, the query sequence was classified as positive if the top hit was positive and vice versa if the top hit was negative.

2.9. Evaluation parameters

Several parameters are considered to measure the ML algorithm's quality or predictive performance [31]. The parameters considered to evaluate model performance are threshold-dependent and threshold-independent. Threshold-dependent parameters include Sensitivity, Specificity, Accuracy, and Matthew's correlation coefficient (MCC), whereas Area Under the Receiver Operating Characteristic (AUROC) is considered a threshold-independent parameter. The major advantage of using the AUROC is that we can select a threshold that best suits our model and evaluate the overall discriminatory power of a model across all possible thresholds. These performance evaluation criteria have been widely applied to evaluate the model's effectiveness and are well-defined in the literature [30,40,41,46]. The measurements are defined as follows:

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

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

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

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{5}$$

Where, TN, TP, FN, and FP stand for true negative, true positive, false negative, and false positive, respectively.

2.10. Web server implementation

We have built a web server called "CLBTope" to predict B-cell epitopes (https://webs.iiitd.edu.in/raghava/clbtope). The user-friendly and device-compatible front end of CLBTope was created using HTML5, CSS, and PHP scripts. Users can submit a sequence in FASTA format through the online server's Predict, Design, Protein Scan, Motif Scan, and Blast Scan modules.

3. Results

3.1. Compositional analysis

AAC provides the occurrence frequency of a given peptide, which allows us to distinguish between the B-cell epitopes from non B-cell epitopes easily. Here, we compute the average composition of B-cell epitopes, non B-cell epitopes, and general proteome, as shown in Fig. 1. The general proteome is the amino acid composition of the Swiss-Prot database. The amino-acid residues lysine (K), aspartic acid (D), glutamic acid (E), asparagine (N), arginine (R), proline (P), and glutamine (Q) are most abundant in the positive dataset. In contrast, residues alanine (A), cysteine (C), phenylalanine (F), isoleucine (I), valine (V), and leucine (L) are highly conserved in the negative dataset.

3.2. Positional analysis

In this study, we have built/constructed a sequence logo to understand and inspect the inclination of a particular residue at a specific position in B-cell epitopes. As depicted in Fig. 2, the hydrophilic residue glutamic acid (E) is a highly abundant and conserved residue at almost all the positions because epitopes are the specific region that resides on the surface of the protein [47] where interactions with antibodies or immune receptors occur. The hydrophilic amino acids are mostly located on the surface of the proteins, which increases the likelihood of antibody binding, whereas serine (S) is more preserved at the 3rd, 6th, 7th, and 21st positions; however, lysine (K) dominates the 18th position.

3.3. B-cell epitope prediction models

B-cell epitope identification is of practical interest for developing antibodies with desired specificity and designing vaccines [29]. We have developed a B-cell epitope prediction model as a classification problem for ML that can accurately distinguish B-cell epitopes from non B-cell epitopes. In this study, we construct a dataset from "BCEPS" that includes 3875 B-cell epitopes, which we call the positive dataset, and 3996 non B-cell epitopes, called a negative dataset. The system architecture of the method is provided in Fig. 3.

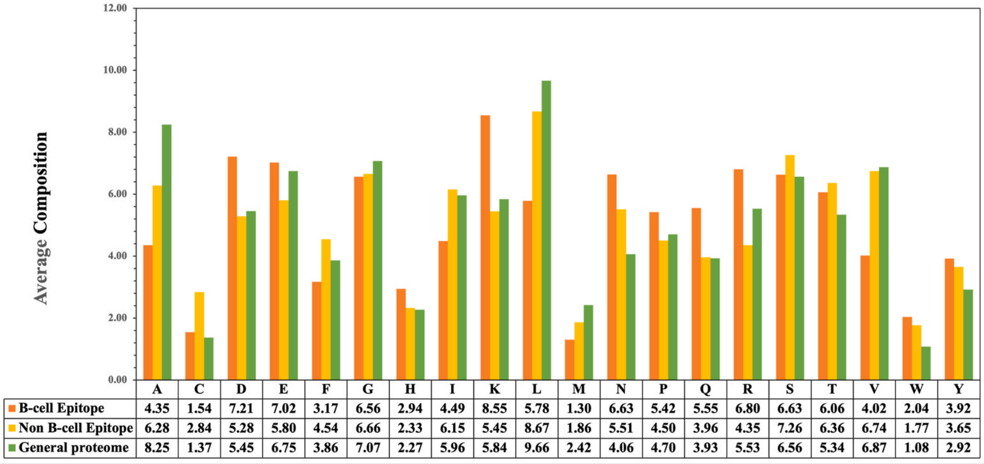
We specially applied various ML algorithms such as DT, RF, LR, XGB, KNN, GNB, and ET to build a model. One of the crucial phases is the selection of the most appropriate ML algorithm for classification. Firstly, we develop prediction models on ten composition-based features extracted using the "Pfeature" tool. It was observed that RF (Random forest) performs best among other ML algorithms and achieved almost similar performance on all the descriptors, achieving AUROC above 0.72. In the 5-fold cross-validation, our model achieved the highest AUROC of 0.802 on DPC-400 features on the validation dataset, significantly better than other descriptors shown in Fig. 4. In Supplementary Table S1, comprehensive results of several classifiers are provided.

In our analysis, most descriptors have a performance that is comparatively similar to DPC. Therefore, we concatenate all the descriptors [33,35] that achieve AUROC above 0.760, producing a hybrid feature vector of size 760, including the proportion of DPC-400. The hybrid feature vector includes DPC, PAAC, SER, APAAC, AAC, SPC, PCP, QSO, and CETD. Here, we observed that RF performs best as it achieves maximum performance with an AUROC 0.800 and has a balanced sensitivity and specificity, as highlighted in Fig. 5. For the performance metrics, including sensitivity, specificity, accuracy, AUROC, and MCC, please refer to Supplementary Table S2.

3.4. Performance of selected features

Many of the features include unrelated features that consume a lot of computing time to train the model. The model is simplified when developed on extracted truly relevant features [32]. In this study, we adopt the mRMR algorithm [38] to remove unrelated features or interdependence between features. We analyzed the performance on the top 400, 200, and 100 selected sets of hybrid features and observed that the maximum AUROC achieved is 0.79 among the above selected sets. Fig. 6 shows the 5-fold cross-validation prediction results on hybrid features with feature selection (mRMR) and without feature selection. Supplementary Table S3 includes all of the results.

The performance on DPC composition (a vector of 400 features) is still superior to the hybrid feature in terms of AUROC, shown in Fig. 7. Furthermore, we focus on the DPC descriptor to improve the performance and selection of the best features using the mRMR feature selection algorithm. We extract a unique set of features from the DPC descriptor and evaluate the performance on top-200, 150, and 100 selected features shown in Supplementary Table S4. It was shown that a DPC-400 features is superior to the performance of these selected sets of features. Fig. 8 compares different classifiers based on DPC feature sets for training and validation datasets (in terms of AUROC).



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
🟧 B-cell Epitope	4.35	1.54	7.21	7.02	3.17	6.56	2.94	4.49	8.55	5.78	1.30	6.63	5.42	5.55	6.80	6.63	6.06	4.02	2.04	3.92
🟨 Non B-cell Epitope	6.28	2.84	5.28	5.80	4.54	6.66	2.33	6.15	5.45	8.67	1.86	5.51	4.50	3.96	4.35	7.26	6.36	6.74	1.77	3.65
🟩 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. 1. Percent average amino acid composition of B-cell epitopes, non B-cell epitopes, and general proteome.

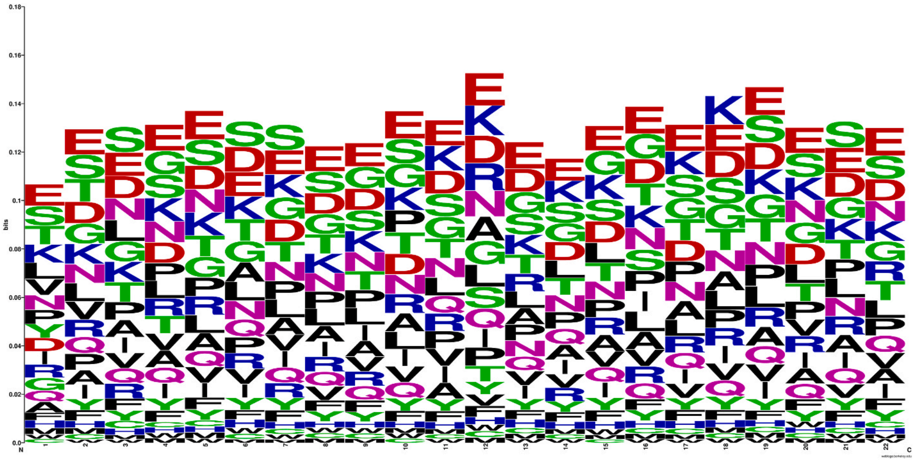


Fig. 2. Sequence logo of B-cell epitopes, glutamic acid (E) residue is dominant at most of the positions.

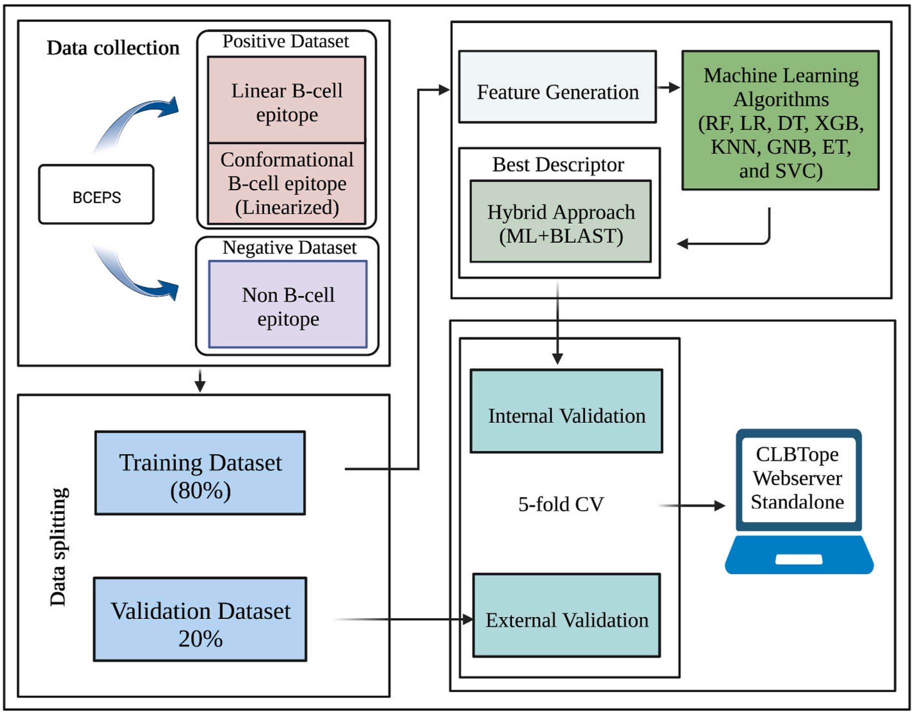


Fig. 3. Complete workflow of the study.

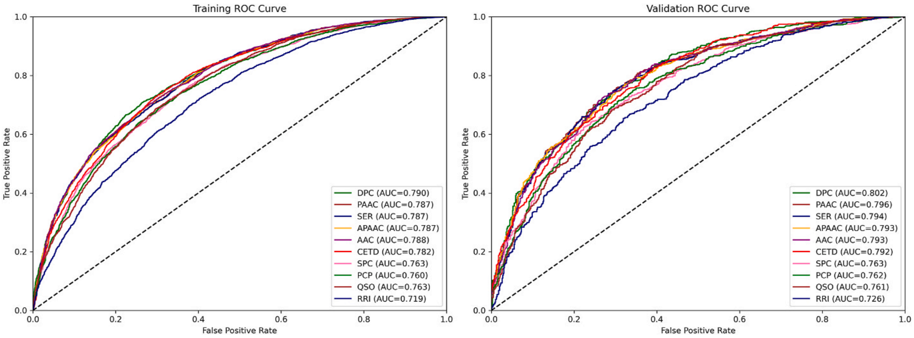


Fig. 4. AUC-ROC score curve showing the performance of ML-based models developed on ten composition-based features/descriptors using RF algorithm.

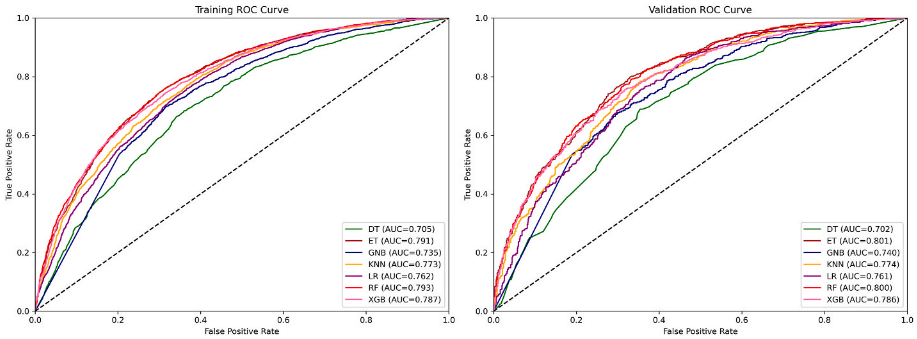
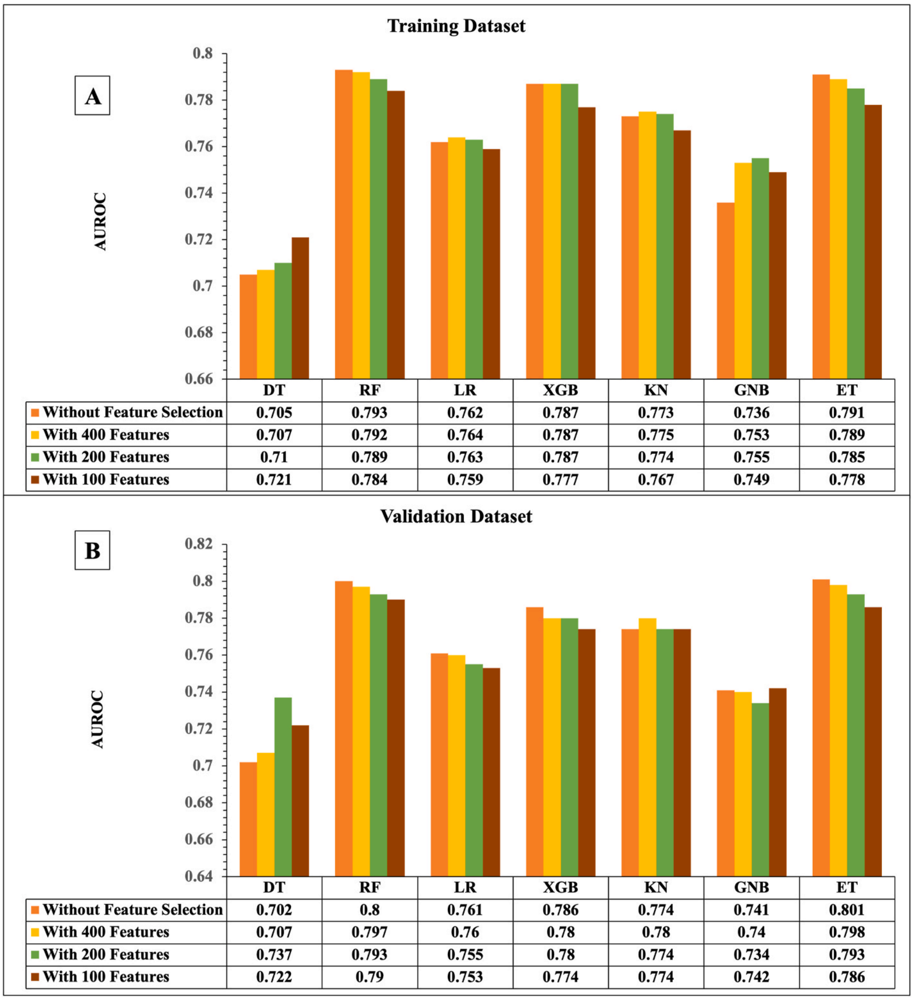


Fig. 5. AUC-ROC score curve showing the performance of machine-learning-based models developed on Hybrid features combination.



	DT	RF	LR	XGB	KN	GNB	ET
🟧 Without Feature Selection	0.705	0.793	0.762	0.787	0.773	0.736	0.791
🟨 With 400 Features	0.707	0.792	0.764	0.787	0.775	0.753	0.789
🟩 With 200 Features	0.71	0.789	0.763	0.787	0.774	0.755	0.785
🟫 With 100 Features	0.721	0.784	0.759	0.777	0.767	0.749	0.778

	DT	RF	LR	XGB	KN	GNB	ET
🟧 Without Feature Selection	0.702	0.8	0.761	0.786	0.774	0.741	0.801
🟨 With 400 Features	0.707	0.797	0.76	0.78	0.78	0.74	0.798
🟩 With 200 Features	0.737	0.793	0.755	0.78	0.774	0.734	0.793
🟫 With 100 Features	0.722	0.79	0.753	0.774	0.774	0.742	0.786

Fig. 6. Performance of different classifiers with and without feature selection (A) Training dataset, (B) Validation dataset.

3.5. *Similarity search approach*

As BLAST was previously used for annotating and assigning functions to proteins based on similarity searches, we have developed a similarity search-based module to improve the model performance further [42, 48]. We implemented a similar concept (blastp) for annotating the given peptide as a B-cell epitope or non B-cell epitope. The query sequences (test set sequences) were searched against a local database that we created using the training dataset at e-values ranging from $1e^{-6}$ to $1e^{+3}$. Each query sequence has been categorized as B-cell epitope or non B-cell epitope based on top-hit. For instance, if the query sequence has the top-hit against a B-cell epitope, then the query peptide is assigned as a B-cell epitope. Otherwise, it is assigned as a non B-cell epitope. As shown in Table 2, the probability of correct prediction (chits) ranges from 1.94 % to 70.32 % on the B-cell epitope dataset and 1.63 %–64.83 % on the non B-cell epitope dataset.

4. Hybrid methods

We combine two approaches to develop models called hybrid methods. We integrate ML scores with BLAST and MERCI scores.

Hybrid1 approach: In this model, we incorporate BLAST similarity

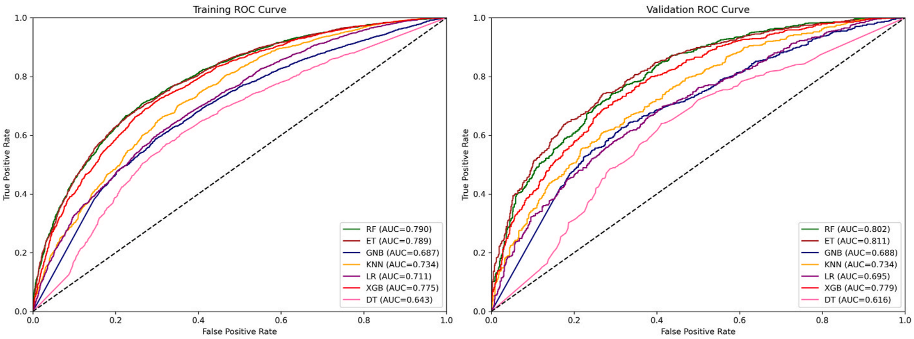
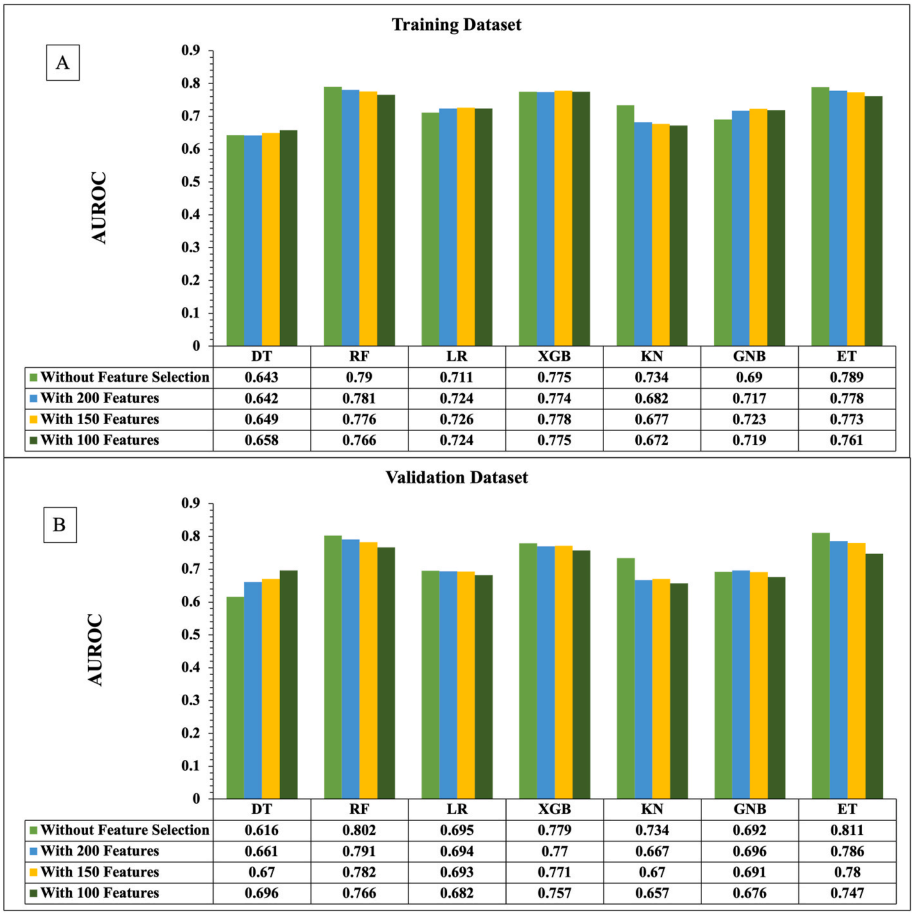


Fig. 7. AUC-ROC score curve showing the performance of machine-learning-based models developed on DPC-400 features.



	DT	RF	LR	XGB	KN	GNB	ET
Without Feature Selection	0.643	0.79	0.711	0.775	0.734	0.69	0.789
With 200 Features	0.642	0.781	0.724	0.774	0.682	0.717	0.778
With 150 Features	0.649	0.776	0.726	0.778	0.677	0.723	0.773
With 100 Features	0.658	0.766	0.724	0.775	0.672	0.719	0.761

	DT	RF	LR	XGB	KN	GNB	ET
Without Feature Selection	0.616	0.802	0.695	0.779	0.734	0.692	0.811
With 200 Features	0.661	0.791	0.694	0.77	0.667	0.696	0.786
With 150 Features	0.67	0.782	0.693	0.771	0.67	0.691	0.78
With 100 Features	0.696	0.766	0.682	0.757	0.657	0.676	0.747

Fig. 8. Performance of different classifiers on DPC-400, 200, 150, and 100 features (A) Training dataset, (B) Validation dataset.

search score and machine learning prediction. The RF-based classifier performs best when utilizing the DPC-400 feature vector; the complete results are shown in Supplementary Table S1. Therefore, our first step is to compute the ML prediction score using our best-performing model. Secondly, we use BLAST at different e-values to classify a given peptide. Then, we assigned "+0.5" for a correct positive prediction (B-cell epitope),'-0.5' for a correct negative prediction (non B-cell epitope), and '0' if the hit is not found [42,48]. To predict B-cell epitope/non B-cell epitope, we integrate the BLAST and ML scores for each query sequence to get the prediction performance. Finally, we developed a hybrid model and calculated the performance at different e-values (see Fig. 9). As shown in Fig. 9, at e-value '1', we obtained a maximum AUROC of 0.829 with an accuracy of 74.460 on validation datasets. The complete presentation of performance measures, including Sensitivity, Specificity, Accuracy, AUROC, and MCC, is shown in Supplementary Table S5.

Hybrid2 approach: We combine the DPC-400 ML and motif scores in this approach. To find new sequences engaged in a biological process of interest, we used a technique called "MERCI" that locates motifs made up of particular amino acids and physico-chemical properties that can be used as discriminators. This tool includes an option to find gapped motifs and introduces the two parameters FP and FN. The FP represents the minimal frequency threshold for the positive sequences, and the FN represents the maximal frequency threshold for the negative sequences [37]. To identify the specific motifs in the positive dataset, we have constructed a positive dataset known as B-cell epitopes and a negative dataset of non B-cell epitopes. By using this dataset, this tool identifies top k motifs that occur most frequently in the positive dataset [37]. We started individually with simple 1-gap, 2-gap, and Koolman and Rohm

Table 2
The performance of BLAST-based search on the validation dataset.

E-value	Positive Dataset			Negative Dataset		
	Chits	Whits	No-hits	Chits	Whits	No-hits
1.00E-06	15 (1.94 %)	1 (0.13 %)	759 (97.94 %)	13 (1.63 %)	1 (0.13 %)	785 (98.25 %)
1.00E-05	18 (2.32 %)	3 (0.39 %)	754 (97.29 %)	19 (2.38 %)	1 (0.13 %)	779 (97.5 %)
1.00E-04	34 (4.39 %)	9 (1.16 %)	732 (94.45 %)	27 (3.38 %)	9 (1.13 %)	763 (95.49 %)
1.00E-03	57 (7.35 %)	12 (1.55 %)	706 (91.10 %)	41 (5.13 %)	11 (1.38 %)	747 (93.49 %)
1.00E-02	85 (10.97 %)	17 (2.19 %)	673 (86.84 %)	59 (7.38 %)	19 (2.38 %)	721 (90.24 %)
1.00E-01	165 (21.29 %)	25 (3.23 %)	585 (75.48 %)	89 (11.14 %)	21 (2.63 %)	689 (86.23 %)
1.00E+00	282 (36.39 %)	30 (3.87 %)	463 (59.74 %)	118 (14.77 %)	25 (3.13 %)	656 (82.10 %)
1.00E+01	386 (49.81 %)	64 (8.26 %)	325 (41.94 %)	205 (25.66 %)	75 (9.39 %)	519 (64.96 %)
1.00E+02	518 (66.84 %)	205 (26.45 %)	52 (6.71 %)	467 (58.45 %)	265 (33.17 %)	67 (8.39 %)
2.00E+02	538 (69.42 %)	227 (29.29 %)	10 (1.29 %)	505 (63.20 %)	280 (35.04 %)	14 (1.75 %)
1.00E+03	545 (70.32 %)	229 (29.55 %)	1 (0.13 %)	518 (64.83 %)	281 (35.17 %)	0 (0.00 %)

[a]Chits: correct hits; Whits: wrong hits.

classification without gaps. We also consider Koolman and Rohm's classification with 1-gap and 2-gap. We integrate the DPC-400 ML score with the Motif score to calculate the performance of the hybrid model. Finally, we achieved maximum AUROC with default parameters of 0.67 and AUROC with parameter FN of 0.71.

4.1. Performance comparison with existing methods

We also performed the comparison with some existing B-cell prediction methods such as iBCEEL [49], ABCPred [21], LBTope [24], CBTope [27], BCEPred [20], BepiPred3.0 [28], iLBE [50], EpiDope [51], BepiBlast [52], SEMA [17], and BCEPS [29] on the validation dataset. Our assessment results show that our tool outperforms other methods in terms of performance measures, including Sensitivity, Specificity, Accuracy, AUROC, and MCC, as shown in Supplementary Table S6. Fig. 10 shows the AUC-ROC score curve showing the comparative analysis of existing methods of B-cell prediction.

5. Case study: analysis of B-cell epitope in SARS-CoV-2 receptor binding domain

The receptor binding domain (RBD) regions are immunogenic, the RBD contains the interacting surface for ACE2 binding. Previous studies of SARS-CoV-2 have indicated that most potent mAbs bind close to the ACE2 interacting surface on the RBD to block the interaction with ACE2 expressed on target cells or disrupt the pre-fusion conformation [53]. Furthermore, In our investigation, we utilized the "Protein Scan" module of the "CLBTope" method, employing a default cut-off value of 0.53 with a peptide length of 11 residues, to predict B-cell epitopes within the "SARS-CoV-2 Receptor binding domain," which comprises 195 amino acids. Our method successfully identified 103 B-cell epitopes and 82 non B-cell epitopes within this domain. These predicted B-cell epitopes represent potential targets for the immune system to recognize and generate an immune response against the SARS-CoV-2 virus. Table 3 displays the top 10 highly potential B-cell epitopes found within the polyprotein. For a comprehensive view of the results, including a more extensive list of identified peptides, the complete findings have been documented in Supplementary Table S7.

6. Webserver implementation

To provide better accessibility and navigation to the scientific community, a user-friendly web server has been developed named

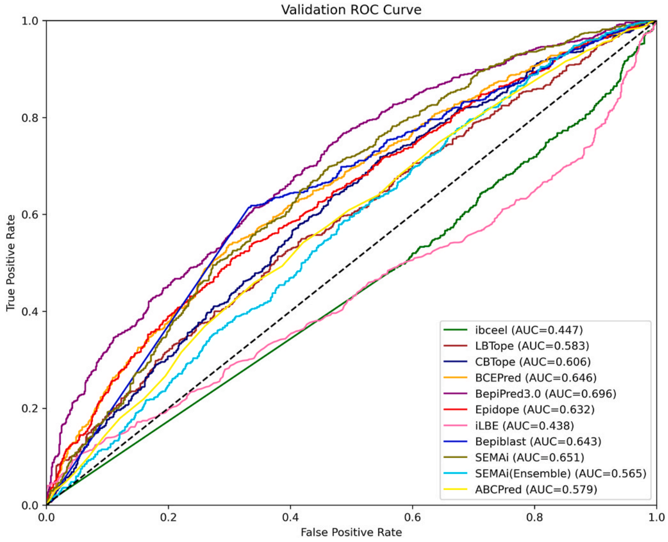


Fig. 10. AUC-ROC score curve showing the Comparative analysis of existing methods of B-cell prediction.

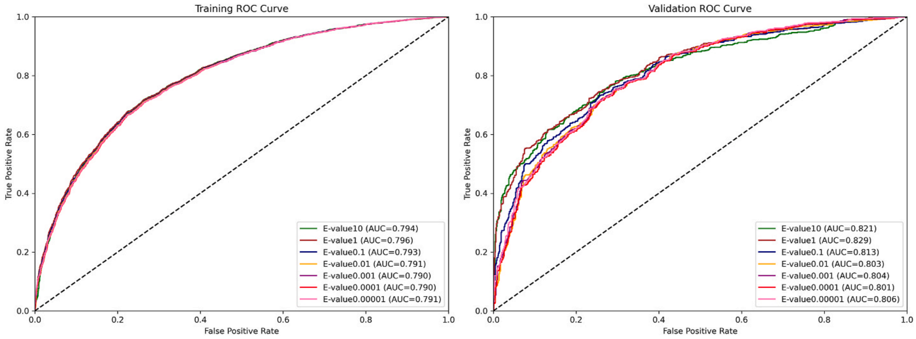


Fig. 9. AUC-ROC score curve showing the model performance developed using the Hybrid method (BLAST and DPC-400) on the training and validation dataset.

Table 3
Potential B-cell epitopes predicted by our tool in SARS-CoV-2 Receptor binding domain, selected based on Hybrid score.

Start	End	Sequence	Hybrid_Score
37	47	YNSASFSTFKC	1.09
39	49	SASFSTFKCYG	1.02
38	48	NSASFSTFKCY	1.01
40	50	ASFSTFKCYGV	0.97
48	58	YGVSPTKLNDL	0.94
41	51	SFSTFKCYGVS	0.93
44	54	TFKCYGVSPTK	0.93
47	57	CYGVSPTKLND	0.92
42	52	FSTFKCYGVSP	0.91
49	59	GVSPTKLNDLC	0.91

"CLBTope" (https://webs.iiitd.edu.in/raghava/clbtope/) for the prediction of B-cell epitopes, either the epitope is linear or conformational (linearized). We implement our top-performing model in the webserver with the following modules: "Predict", "Design", "Scan", "Motif Scan", and "Blast Scan". Users can classify submitted sequences as B-cell epitopes or non B-cell epitopes using the "Predict" module. It allows us to either paste or upload a file containing single or multiple peptide sequences in FASTA format with length between 11 and 25. Using the 'Protein Scan' module, users can scan or identify regions in an amino-acid sequence corresponding to B-cells or non-B-cells. Users can generate all B-cell epitope analogs created in the submitted sequence through the' Design' module. The 'Blast Scan' module allows them to search the query sequence against the database of known B-cell epitopes. A query sequence is predicted as a B-cell epitope or non B-cell epitope depending upon the match or hit in the database. If the query sequence to be searched is present in the database, then it is accounted for a match or hit in the database and is predicted/classified as a B-cell epitope; otherwise, the sequence is classified as non B-cell epitope. A FASTA file format is also available for users to download the positive and negative datasets we used in this study.

7. Discussion and conclusion

Epitope-based peptide vaccines have an advantage over conventional vaccines as they facilitate the precise delivery of targeted vaccines. Predicting B-cell epitope not only aids immunologists in designing epitope-based peptide vaccines but also advances the development of accurate diagnostic kits and effective treatment [54,55]. In this study, we introduced a new method called "CLBTope" for predicting B-cell epitope and non B-cell epitope. Our model is trained on a dataset comprising 3875 B-cell epitopes (positive dataset) and 3996 non B-cell epitopes (negative dataset) obtained from BCEPS. Utilizing "Pfeature", we extract features from peptide sequences and apply various ML algorithms: DT, RF, LR, XGB, KNN, GNB, and ET, developing prediction models on ten composition-based features. Among all models, the RF-based model demonstrates superior performance, achieving a maximum AUROC of 0.802 on the DPC descriptor, which consists of a vector of 400 features. We further compute the hybrid feature vector of size 760 and develop our RF model, achieving an AUROC of 0.800. The hybrid vector was further reduced using the mRMR feature selection algorithm and analyzed the performance on the top 400, 200, and 100, achieving the maximum AUROC of 0.79. To enhance the model's performance further, we developed an RF-based hybrid model by integrating BLAST and DPC-400 features. This new method achieves a maximum AUROC of 0.829 on the validation dataset. We have also compared our method with the previously available methods such as iBCEEL [49], ABCPred [21], LBTope [24], CBTope [27], BCEPred [20], BepiPred3.0 [28], iLBE [50], EpiDope [51], BepiBlast [52], SEMA [17], and BCEPS [29]. We anticipate that this method will significantly contribute to the scientific community working in this domain. We have provided a user-friendly webserver CLBTope (https://webs.iiitd.edu.in/raghava/clbtope/) for predicting, scanning, and designing B-cell epitopes.

Funding Source

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

Data availability Statement

All the datasets used in this study are available at the "CLBTope" web server, https://webs.iiitd.edu.in/raghava/clbtope/algo.php#dataset.

BioRxiv doi

https://doi.org/10.1101/2023.01.18.524531.

CRediT authorship contribution statement

Nishant Kumar: Conceptualization, Methodology, Software, Validation, Writing – original draft, Writing – review & editing. **Sadhana Tripathi:** Methodology, Writing – original draft, Writing – review & editing. **Neelam Sharma:** Methodology, Writing – review & editing. **Sumeet Patiyal:** Data curation, Methodology, Software, Validation, Writing – original draft, Writing – review & editing. **Naorem Leimarembi Devi:** Methodology, Writing – review & editing. **Gajendra P.S. Raghava:** Conceptualization, Data curation, Methodology, Project administration, Software, Validation, Visualization, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare no competing financial and non-financial interests.

Acknowledgements

Authors are thankful to the University Grants Commission (UGC), Department of Bio-Technology (DBT), DBT-RA program in Biotechnology and Life Sciences, and Department of Science and Technology (DST-INSPIRE) for fellowships and financial support, and the Department of Computational Biology, IIITD New Delhi for infrastructure and facilities. We would like to acknowledge that 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.compbiomed.2024.108083.

References

[1] J. Parkin, B. Cohen, An overview of the immune system, Lancet 357 (2001) 1777–1789.
[2] D.D. Chaplin, Overview of the immune response, J. Allergy Clin. Immunol. 125 (2010) S3–S23.
[3] B.P. Kaur, E. Secord, Innate immunity, Pediatr. Clin. North Am. 66 (2019) 905–911.
[4] A.E. Thompson, JAMA patient page. The immune system, JAMA 313 (2015) 1686.
[5] J.S. Marshall, R. Warrington, W. Watson, H.L. Kim, An introduction to immunology and immunopathology, Allergy Asthma Clin. Immunol. 14 (2018) 49.
[6] C.B. Palatnik-de-Sousa, I.S. Soares, D.S. Rosa, Editorial: epitope discovery and synthetic vaccine design, Front. Immunol. 9 (2018) 826.
[7] D.J. Barlow, M.S. Edwards, J.M. Thornton, Continuous and discontinuous protein antigenic determinants, Nature 322 (1986) 747–748.
[8] T.A. Najar, S. Khare, R. Pandey, S.K. Gupta, R. Varadarajan, Mapping protein binding sites and conformational epitopes using cysteine labeling and yeast surface display, Structure 25 (2017) 395–406.

[9] Y.T. Lo, T.C. Shih, T.W. Pai, L.P. Ho, J.L. Wu, H.Y. Chou, Conformational epitope matching and prediction based on protein surface spiral features, BMC Genom. 22 (2021) 116.
[10] K.A. Galanis, K.C. Nastou, N.C. Papandreou, G.N. Petichakis, D.G. Pigis, V. A. Iconomidou, Linear B-cell epitope prediction for in silico vaccine design: a performance review of methods available via command-line interface, Int. J. Mol. Sci. 22 (2021).
[11] E.E.G. Kozlova, L. Cerf, F.S. Schneider, B.T. Viart, N.G. C, B.T. Steiner, S. de Almeida Lima, F. Molina, C.G. Duarte, L. Felicori, C. Chavez-Olortegui, R. A. Machado-de-Avila, Computational B-cell epitope identification and production of neutralizing murine antibodies against Atroxlysin-I, Sci. Rep. 8 (2018) 14904.
[12] M.C. Jespersen, S. Mahajan, B. Peters, M. Nielsen, P. Marcatili, Antibody specific B-cell epitope predictions: leveraging information from antibody-antigen protein complexes, Front. Immunol. 10 (2019) 298.
[13] U. Kulkarni-Kale, S. Bhosle, A.S. Kolaskar, CEP: a conformational epitope prediction server, Nucleic Acids Res. 33 (2005) W168–W171.
[14] J.V. Kringelum, C. Lundegaard, O. Lund, M. Nielsen, Reliable B cell epitope predictions: impacts of method development and improved benchmarking, PLoS Comput. Biol. 8 (2012) e1002829.
[15] C. Zhou, Z. Chen, L. Zhang, D. Yan, T. Mao, K. Tang, T. Qiu, Z. Cao, SEPPA 3.0-enhanced spatial epitope prediction enabling glycoprotein antigens, Nucleic Acids Res. 47 (2019) W388–W394.
[16] B.M. da Silva, Y. Myung, D.B. Ascher, D.E.V. Pires, epitope3D: a machine learning method for conformational B-cell epitope prediction, Briefings Bioinf. 23 (2022).
[17] T.I. Shashkova, D. Umerenkov, M. Salnikov, P.V. Strashnov, A.V. Konstantinova, I. Lebed, D.N. Shcherbinin, M.N. Asatryan, O.L. Kardymon, N.V. Ivanisenko, SEMA: antigen B-cell conformational epitope prediction using deep transfer learning, Front. Immunol. 13 (2022) 960985.
[18] J.L. Pellequer, E. Westhof, PREDITOP: a program for antigenicity prediction, J. Mol. Graph. 11 (1993) 204–210, 191-202.
[19] M. Odorico, J.L. Pellequer, BEPITOPE: predicting the location of continuous epitopes and patterns in proteins, J. Mol. Recogn. 16 (2003) 20–22.
[20] S. Saha, BcePred: prediction of continuous B-cell epitopes in antigenic sequences using physico-chemical properties, Springer-Verlag Berlin. Heidelberg. 3239 (2004).
[21] S. Saha, G.P. Raghava, Prediction of continuous B-cell epitopes in an antigen using recurrent neural network, Proteins 65 (2006) 40–48.
[22] J.E. Larsen, O. Lund, M. Nielsen, Improved method for predicting linear B-cell epitopes, Immunome Res. 2 (2006) 2.
[23] Y. El-Manzalawy, D. Dobbs, V. Honavar, Predicting linear B-cell epitopes using string kernels, J. Mol. Recogn. 21 (2008) 243–255.
[24] H. Singh, H.R. Ansari, G.P. Raghava, Improved method for linear B-cell epitope prediction using antigen's primary sequence, PLoS One 8 (2013) e62216.
[25] Y. Lian, M. Ge, X.M. Pan, EPMLR: sequence-based linear B-cell epitope prediction method using multiple linear regression, BMC Bioinf. 15 (2014) 414.
[26] T. Liu, K. Shi, W. Li, Deep learning methods improve linear B-cell epitope prediction, BioData Min. 13 (2020) 1.
[27] H.R. Ansari, G.P. Raghava, Identification of conformational B-cell Epitopes in an antigen from its primary sequence, Immunome Res. 6 (2010) 6.
[28] J.N. Clifford, M.H. Hoie, S. Deleuran, B. Peters, M. Nielsen, P. Marcatili, BepiPred-3.0: improved B-cell epitope prediction using protein language models, Protein Sci. 31 (2022) e4497.
[29] A. Ras-Carmona, H.F. Pelaez-Prestel, E.M. Lafuente, P.A. Reche, BCEPS: a web server to predict linear B cell epitopes with enhanced immunogenicity and cross-reactivity, Cells 10 (2021).
[30] S. Gupta, P. Kapoor, K. Chaudhary, A. Gautam, R. Kumar, C. Open Source Drug Discovery, G.P. Raghava, In silico approach for predicting toxicity of peptides and proteins, PLoS One 8 (2013) e73957.
[31] X. Liang, F. Li, J. Chen, J. Li, H. Wu, S. Li, J. Song, Q. Liu, Large-scale comparative review and assessment of computational methods for anti-cancer peptide identification, Briefings Bioinf. (2021) 22.
[32] L. Qian, Y. Wen, G. Han, Identification of cancerlectins using support vector machines with fusion of G-gap dipeptide, Front. Genet. 11 (2020) 275.
[33] J.L. Blanco, A.B. Porto-Pazos, A. Pazos, C. Fernandez-Lozano, Prediction of high anti-angiogenic activity peptides in silico using a generalized linear model and feature selection, Sci. Rep. 8 (2018) 15688.
[34] G.E. Crooks, G. Hon, J.M. Chandonia, S.E. Brenner, WebLogo: a sequence logo generator, Genome Res. 14 (2004) 1188–1190.
[35] R. Yang, C. Zhang, R. Gao, L. Zhang, A novel feature extraction method with feature selection to identify golgi-resident protein types from imbalanced data, Int. J. Mol. Sci. 17 (2016) 218.
[36] A. Pande, S. Patiyal, A. Lathwal, C. Arora, D. Kaur, A. Dhall, G. Mishra, H. Kaur, N. Sharma, S. Jain, S.S. Usmani, P. Agrawal, R. Kumar, V. Kumar, G.P.S. Raghava, Pfeature: a tool for computing wide range of protein features and building prediction models, J. Comput. Biol. (2022).
[37] C. Vens, M.N. Rosso, E.G. Danchin, Identifying discriminative classification-based motifs in biological sequences, Bioinformatics 27 (2011) 1231–1238.
[38] M. Radovic, M. Ghalwash, N. Filipovic, Z. Obradovic, Minimum redundancy maximum relevance feature selection approach for temporal gene expression data, BMC Bioinf. 18 (2017) 9.
[39] N. Sharma, L.D. Naorem, S. Jain, G.P.S. Raghava, ToxinPred2: an improved method for predicting toxicity of proteins, Briefings Bioinf. 23 (2022).
[40] S. Jain, A. Dhall, S. Patiyal, G.P.S. Raghava, IL13Pred: a method for predicting immunoregulatory cytokine IL-13 inducing peptides, Comput. Biol. Med. 143 (2022) 105297.
[41] A. Dhall, S. Patiyal, N. Sharma, S.S. Usmani, G.P.S. Raghava, Computer-aided prediction and design of IL-6 inducing peptides: IL-6 plays a crucial role in COVID-19, Briefings Bioinf. 22 (2021) 936–945.
[42] D. Kaur, C. Arora, G.P.S. Raghava, A hybrid model for predicting pattern recognition receptors using evolutionary information, Front. Immunol. 11 (2020) 71.
[43] G.M. Boratyn, A.A. Schaffer, R. Agarwala, S.F. Altschul, D.J. Lipman, T.L. Madden, Domain enhanced lookup time accelerated BLAST, Biol. Direct 7 (2012) 12.
[44] M. Kumar, M.M. Gromiha, G.P. Raghava, SVM based prediction of RNA-binding proteins using binding residues and evolutionary information, J. Mol. Recogn. 24 (2011) 303–313.
[45] H. Singh, G.P. Raghava, BLAST-based structural annotation of protein residues using Protein Data Bank, Biol. Direct 11 (2016) 4.
[46] N. Sharma, S. Patiyal, A. Dhall, N.L. Devi, G.P.S. Raghava, ChAlPred: a web server for prediction of allergenicity of chemical compounds, Comput. Biol. Med. 136 (2021) 104746.
[47] J.V. Kringelum, M. Nielsen, S.B. Padkjaer, O. Lund, Structural analysis of B-cell epitopes in antibody:protein complexes, Mol. Immunol. 53 (2013) 24–34.
[48] N. Sharma, S. Patiyal, A. Dhall, A. Pande, C. Arora, G.P.S. Raghava, AlgPred 2.0: an improved method for predicting allergenic proteins and mapping of IgE epitopes, Briefings Bioinf. 22 (2021).
[49] B. Manavalan, R.G. Govindaraj, T.H. Shin, M.O. Kim, G. Lee, iBCE-EL: a new ensemble learning framework for improved linear B-cell epitope prediction, Front. Immunol. 9 (2018) 1695.
[50] M.M. Hasan, M.S. Khatun, H. Kurata, iLBE for computational identification of linear B-cell epitopes by integrating sequence and evolutionary features, Dev. Reprod. Biol. 18 (2020) 593–600.
[51] M. Collatz, F. Mock, E. Barth, M. Holzer, K. Sachse, M. Marz, EpiDope: a deep neural network for linear B-cell epitope prediction, Bioinformatics 37 (2021) 448–455.
[52] A. Ras-Carmona, A.A. Lehmann, P.V. Lehmann, P.A. Reche, Prediction of B cell epitopes in proteins using a novel sequence similarity-based method, Sci. Rep. 12 (2022) 13739.
[53] W. Dejnirattisai, D. Zhou, H.M. Ginn, H.M.E. Duyvesteyn, P. Supasa, J.B. Case, Y. Zhao, T.S. Walter, A.J. Mentzer, C. Liu, B. Wang, G.C. Paesen, J. Slon-Campos, C. Lopez-Camacho, N.M. Kafai, A.L. Bailey, R.E. Chen, B. Ying, C. Thompson, J. Bolton, A. Fyfe, S. Gupta, T.K. Tan, J. Gilbert-Jaramillo, W. James, M. Knight, M. W. Carroll, D. Skelly, C. Dold, Y. Peng, R. Levin, T. Dong, A.J. Pollard, J.C. Knight, P. Klenerman, N. Temperton, D.R. Hall, M.A. Williams, N.G. Paterson, F.K. R. Bertram, C.A. Siebert, D.K. Clare, A. Howe, J. Radecke, Y. Song, A.R. Townsend, K.A. Huang, E.E. Fry, J. Mongkolsapaya, M.S. Diamond, J. Ren, D.I. Stuart, G. R. Screaton, The antigenic anatomy of SARS-CoV-2 receptor binding domain, Cell 184 (2021) 2183–2200, e2122.
[54] H.W. Wang, T.W. Pai, Machine learning-based methods for prediction of linear B-cell epitopes, Methods Mol. Biol. 1184 (2014) 217–236.
[55] A.W. Purcell, J. McCluskey, J. Rossjohn, More than one reason to rethink the use of peptides in vaccine design, Nat. Rev. Drug Discov. 6 (2007) 404–414.

1. Nishant Kumar is currently working as Ph.D. in Computational biology from Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India.

2. Sadhana Tripathi is currently working as M.Tech in Computational biology from Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India.

3. Neelam Sharma is currently working as Ph.D. in Computational biology from Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India.

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

5. Dr. Naorem Leimarembi Devi is currently working as a DBT-Research Associate in Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India.

6. Gajendra P. S. Raghava is currently working as Professor and Head of Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India.