Computers in Biology and Medicine 158 (2023) 106864

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

A web server for predicting and scanning of IL-5 inducing peptides using alignment-free and alignment-based method

Leimarembi Devi Naorem, Neelam Sharma, Gajendra P.S. Raghava *

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

ARTICLE INFO

Keywords:
IL-5 inducing peptides
Alignment-based method
Alignment-free method
Machine learning techniques
BLAST
Motifs

ABSTRACT

Interleukin-5 (IL-5) can act as an enticing therapeutic target due to its pivotal role in several eosinophil-mediated diseases. The aim of this study is to develop a model for predicting IL-5 inducing antigenic regions in a protein with high precision. All models in this study have been trained, tested and validated on experimentally validated 1907 IL-5 inducing and 7759 non-IL-5 inducing peptides obtained from IEDB. Our primary analysis indicates that IL-5 inducing peptides are dominated by certain residues like Ile, Asn, and Tyr. It was also observed that binders of a wide range of HLA alleles can induce IL-5. Initially, alignment-based methods have been developed using similarity and motif search. These alignment-based methods provide high precision but poor coverage. In order to overcome this limitation, we explore alignment-free methods which are mainly machine learning-based models. Firstly, models have been developed using binary profiles and eXtreme Gradient Boosting-based model achieved a maximum AUC of 0.59. Secondly, composition-based models have been developed and our dipeptide-based random forest model achieved a maximum AUC of 0.74. Thirdly, random forest model developed using selected 250 dipeptides and achieved AUC 0.75 and MCC 0.29 on validation dataset; best among alignment-free models. In order to improve the performance, we developed an ensemble or hybrid method that combined alignment-based and alignment-free methods. Our hybrid method achieved AUC 0.94 with MCC 0.60 on a validation/independent dataset. The best hybrid model developed in this study has been incorporated into the user-friendly web server and a standalone package named 'IL5pred' (https://webs.iiitd.edu.in/raghava/il5pred/).

1. Introduction

Innate immunity is considered the first line of defense against invading pathogens [1]. In contrast, adaptive immunity act as the second line of defense and is referred to as antigen-specific immune response. The immunological response against a pathogen is mediated via major histocompatibility complex (MHC) by presenting the antigen on the surface of antigen-presenting cells including dendritic, B-cells and macrophages [1]. MHC class II is responsible for processing and presenting exogenous antigens. This antigenic peptide activates $CD4^+$ T-helper cells (Th) hence inducing the release of different cytokines [2]. Interleukin-5 (IL-5) is a crucial cytokine secreted by Th2, type-2 innate lymphoid, mast, basophils and eosinophils cells [3,4]. These cells produced IL-5 upon stimulation by several environmental pollutants, inhaled allergens and microbes [5]. IL-5 is a glycosylated homodimeric protein with 45–60 kDa molecular weight, composed of two helical bundle motifs [6]. The gene that encodes IL-5 is found in the same cluster as IL-3, IL-4, IL-13 and granulocyte-macrophage colony-stimulating factor (GM-CSF) [7]. IL-5 exerts its pleiotropic actions via IL-5 receptor (IL-5R), which is comprised of an α and a βc chain. The α subunit recognises the IL-5 molecule, whereas the βc subunit recognises either IL-3 or GM-CSF. Thereby promoting eosinophils maturation, activation, survival and discharge from the bone marrow into the bloodstream and finally to the airways [5-8]. When activated by IL-5, eosinophils degranulate and produce antimicrobial cytotoxins that are harmful to nearby cells and tissues [9]. Although IL-5 is critical in eosinophil development, it has also been linked to the onset and severity of a number of diseases, including asthma, autoimmune disorders, allergy, atopic dermatitis, eosinophilic esophagitis, and cancer [5,10–13]. IL-5 production during asthma exacerbation can cause pulmonary eosinophilia, enhancing airway smooth muscle contraction and increased mucus production [14]. It has been demonstrated in earlier studies that IL-5 is induced by a wide variety of MHC class II alleles and has a role in MHC class II regulation [15–17]. The overall biological effects of IL-5 on eosinophils are depicted in Fig. 1.

* Corresponding author.
E-mail addresses: leimarembi@gmail.com (L.D. Naorem), neelams@iiitd.ac.in (N. Sharma), raghava@iiitd.ac.in (G.P.S. Raghava).

https://doi.org/10.1016/j.compbiomed.2023.106864
Received 1 November 2022; Received in revised form 6 March 2023; Accepted 30 March 2023
Available online 4 April 2023

List of abbreviations

MHC	Major Histocompatibility Complex
IL-5	Interleukin-5
IEDB	Immune Epitope Database
BLAST	Basic Local Alignment Search Tool
HLA	Human Leukocyte Antigen
MERCI	Motif-EmeRging and with Classes-Identification
AAC	Amino Acid Composition
DPC	Dipeptide composition
ML	Machine Learning
RF	Random Forest
KNN	K-nearest neighbour
DT	Decision Tree
GNB	Gaussian Naïve Bayes
XGB	Extreme Gradient Boosting
LR	Logistic Regression
Acc	Accuracy
Sens	Sensitivity
Spec	Specificity
MCC	Matthews correlation coefficient
AUC	Area under the receiver operating characteristic curve

In the past, numerous computational methods have been developed for designing peptide/epitope-based subunit vaccines and immunotherapy. In the initial phase, methods have been developed for predicting Human Leukocyte Antigen (HLA) class I and II binders. Major HLA class I binder prediction include ProPred1 [18], nHLAPred [19], PSSMHCpan [20], NetMHCpan [21] and NetMHC-3.0 [22]. Major HLA class II binder prediction methods include ProPred [23], HLA-DR4Pred [24], NetMHCII [25] and NetMHCIIpan [26]. Besides, another method HLA_{nc}Pred [27] has been developed to predict non-classical HLA binders. Recently, the trend has changed as a number of researchers are developing methods for predicting peptides that can induce specific types of cytokines. CytoPred is a method that can predict and classify cytokines with high accuracy [28]. Currently, the following methods, IL4pred [29], IL-6Pred [30], IL-10Pred [31], ILeukin10Pred [32], IL17eScan [33], and IL-13Pred [34] have been developed to predict IL-4, IL-6, IL-10, IL-17, and IL-13 inducing peptides respectively. Moreover, PIP-EL [35] and ProInflam [36] are the tools used to predict the peptides inducing a group of cytokines. These above-mentioned prediction tools utilize compositional and binary profile features of the peptide sequence. For instance, IL4pred and IL17eScan use amino acid, dipeptide composition and amino acid pairs, IL10pred uses dipeptide composition as input features, and many more.

Taking into consideration the importance of IL-5 in several diseases there is a necessity to develop a computational method that can predict IL-5 inducing peptides. To the best of our knowledge, no method has been developed to predict IL-5 inducing peptides. Thus in this study, we proposed a novel method IL5pred to identify these peptides. We retrieved the experimentally validated dataset from the well-established database Immune Epitope Database (IEDB) [37]. We have implemented Basic Local Alignment Search Tool (BLAST) to predict the peptides based on their similarity to known IL-5 inducers. A motif-based approach was utilised to identify motifs in IL-5 inducers. Next, a wide range of

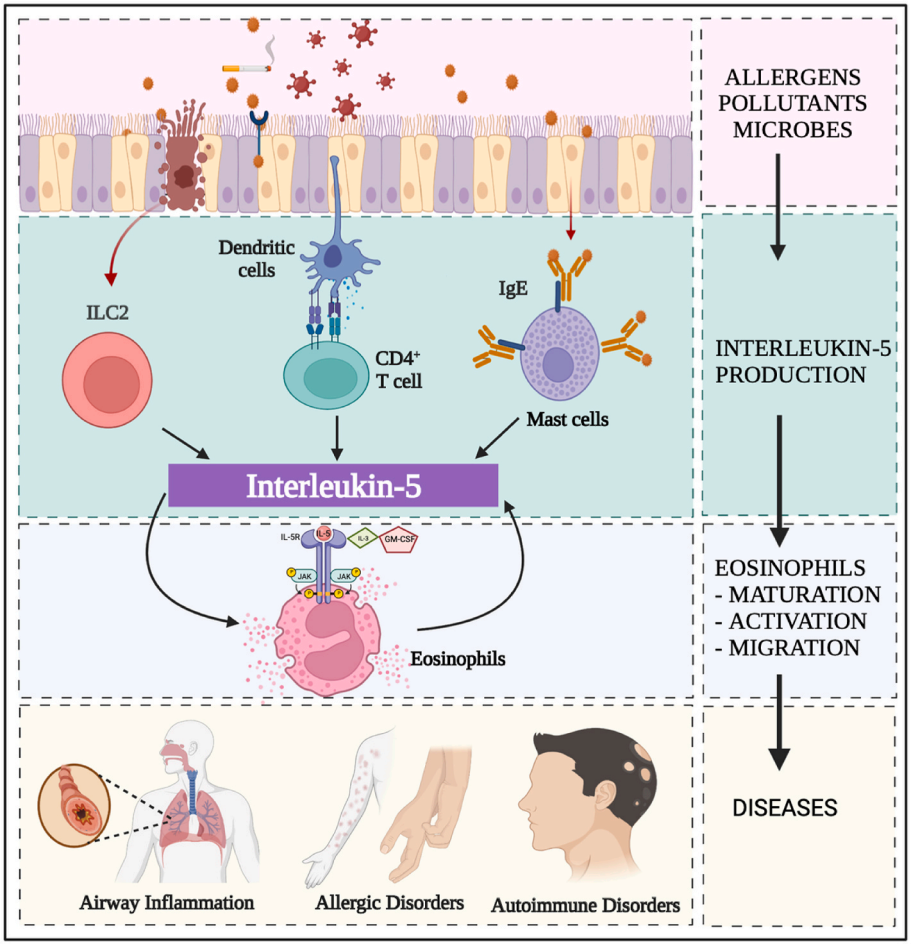


Fig. 1. Biological effects of IL-5 in the development of eosinophils and its linkage to the onset and severity of several diseases.

sequence-based features including composition and binary profiles for each peptide was computed using Pfeature tool [38]. We developed several machine-learning models utilising these features to predict IL-5 inducers with better accuracy and the best model was incorporated into the web server that can efficiently classify IL-5 inducers and non-inducers.

2. Materials and methods

2.1. Dataset preparation

The most challenging aspect of designing a bioinformatics tool is gathering enough experimentally validated data. The current study retreived IL-5 inducing and non-IL-5 inducing peptides from IEDB [37], a freely accessible database that consists of a huge amount of experimentally verified immune epitopes. We created two datasets: main dataset and alternate dataset.

2.1.1. Main dataset

We extracted 2802 IL-5 inducing peptides and 8674 non-IL-5 inducing peptides, which are experimentally validated MHC class II binders from IEDB tested on *Homo sapiens*. Further, all peptides containing non-standard characters (i.e., 'B,' 'J,' 'O', 'U,' 'X' and 'Z') and duplicates were removed. We have chosen unique linear peptides with lengths ranging from 9 to 20 amino acids. The peptides found to be common or exactly matched in both datasets were also eliminated. Finally, 1907 IL-5 inducing and 7759 non-IL-5 inducing peptides were obtained and labelled as positive and negative datasets, respectively.

2.1.2. Alternate dataset

We have also created an alternate dataset comprising IL-5 inducing as a positive dataset and random peptides as a negative dataset (generated from the Swiss-Prot database) [39]. We randomly extracted 1907 peptides and assigned them as negative peptides to create a balanced dataset. Finally, the alternate dataset comprises 1907 IL-5 inducing peptides and 1907 random peptides.

2.2. Sequence logos

To investigate the preference of individual amino acids at a particular position, we generated the sequence logo using the R package "ggseqlogo" [40]. It gives the graphical representation with residue positions on x-axis and the bit score indicating the conservation of residues at a specific position on y-axis. This package processes a fixed-length input vector, thus, we considered the minimum length of 9 amino acids from N-terminal for each peptide sequence.

2.3. Motif identification using MERCI

The detection of motifs in peptides is crucial for annotating the function of the sequence. The current study employed a publicly available software, Motif-EmeRging and with Classes-Identification (MERCI) [41] to search motifs in IL-5 inducing peptides. This software used a Perl script to find the motifs exclusively present in positive and negative peptide sequences. This MERCI program examined both the positive and negative peptide sequences to extract the specific patterns. Thus, we adopted a two-step strategy to identify the motifs exclusively present in positive and negative peptide sequences. Initially, we retrieved the motifs present in IL-5 inducing peptides by inputting IL-5 inducing peptides as positive and IL-5 non-inducing peptides as negative. We next flipped the datasets providing MERCI with IL-5 non-inducing peptides as positive datasets and inducing peptides as negative, thus obtaining the motifs for IL-5 non-inducing peptides.

2.4. Feature generation

In order to develop any prediction model, it is necessary to extract a set of pertinent features for each protein or peptide sequence. Different feature extraction methods have been employed in various studies. Hong et al., have proposed novel protein encoding strategies by integrating them with convolution neural networks [42] and deep learning methods to annotate the proteins as well as to restrain false predictions [43]. Another study by Xia et al. has developed a method named PFmulDL, in which multiple deep-learning algorithms have been amalgamated for protein function interpretation. In this study, authors have combined recurrent neural network (RNN) along with the multi-kernel CNN method to annotate the function and then transfer learning has been implemented for improving the efficiency of the model [44]. SVM-Prot-2016 is another machine learning-based method incorporating K nearest neighbour (kNN) and probabilistic neural networks (PNN) for predicting the functions of the proteins from their sequence regardless of their similarity [45].

In the study, we have used Pfeature [38], a standalone tool, to generate distinct features from peptide sequence information. Using this tool, we computed several compositional-based features as well as binary profile-based features for each peptide sequence. The major composition features such as amino acid composition (AAC), dipeptide composition (DPC), tripeptide composition (TPC), atom composition (ATC), Physico-chemical properties repeat composition (PRI) along with their vector length are tabulated in Table S1. Following are some of the descriptions of the composition-based features used in the current study to develop several alignment-free prediction models using machine learning (ML) techniques.

2.5. Amino acid composition-based features

AAC has been used efficaciously in a number of sequence-based classification techniques [46]. It is the primary feature that describes the fraction of each amino acid residue present in a peptide sequence with 20 length vector. Equation (1) is used to calculate AAC for each amino acid residue.

$$\boldsymbol{AAC(i)} = \frac{\boldsymbol{Ri}}{\boldsymbol{N}} \times \mathbf{100} \tag{1}$$

AAC(i) represents the per cent amino acid composition of residue type (i); Ri represents the number of residues of type i, and N represents the length of the sequence.

2.6. Dipeptide composition-based features

DPC is another frequently utilised input parameter for peptide composition-based classification. It captures comprehensive information about the pairwise composition of the amino acids in the peptide sequence with a fixed vector of length 400 (20*20). The following equation is used to calculate DPC.

$$\mathrm{DPC(i)} = \frac{\text{Total number of dipeptides (i)}}{\mathrm{N}-1} \times 100 \tag{2}$$

DPC(i) represents the per cent of dipeptide composition of residue type i, and N represents the length of the sequence.

2.7. Selection and ranking of features

A total of 9553 features were computed for both main and alternate datasets. Only relevant features are used for developing the classification model. Thus, selecting relevant features from a larger set of features is the most important step but challenging. Although there are a number of techniques for feature selection, we employed a support vector classifier (SVC) with L1 regularization employing the Scikit-learn package [47]. This method chooses the non-zero coefficients and implements the

L1 penalty to select the small set of relevant features. During the optimization process, the L1 regularization handles the sparse matrix by selecting some model features. Another important regularization parameter used in this technique is "C" parameter which is set to the default value (i.e. 0.01), and as lower the C value, fewer features are selected [48]. Feature-selector tool was further used to rank these features based on their importance. This tool rank features that are frequently used to divide the dataset among all trees, using the DT-based algorithm Light Gradient Boosting Machine [49].

2.8. Binary profile-based features

Previous research studies have shown that the binary profile is one of the essential features for classifying peptides [50]. Generally, it is challenging to create a fixed-length pattern given the varied length of the peptides. In our study, the length of the peptides varies from 9 to 20, thus we generated a fixed length binary profile by extracting the fixed length segments from either N or C terminus of the peptide [50]. Since the minimum length of the peptides is 9, thus we created binary profiles for N_9, C_9 and also for combined terminal residues N_9C_9 after calculating the fixed length patterns.

2.9. Alignment-based search using BLAST

In this study, BLAST has been used for alignment-based search. It is the most frequently used tool for annotating protein/peptide and nucleotide sequences [51]. We implemented blastp-short for short peptide sequences (8–30 amino acids) to identify IL-5 inducing peptides based on the similarity of peptides with IL-5 inducers and non-inducers. To identify IL-5 inducers and non-inducers, the top hit of BLAST was considered at different E-value cut-offs.

2.10. Alignment-free approach for classification

To generate more accurate prediction methods, several studies have employed alignment-free approach using different ML techniques [52,53]. For this, Random Forest (RF) [54], K-nearest neighbour (KNN) [55], Decision Tree (DT) [56], Gaussian Naïve Bayes (GNB) [57], XGBoost (XGB) [58], and Logistic Regression (LR) [59] were employed using Scikit's sklearn package from Python [47]. Different hyper-parameters were tuned in these classification algorithms, and the results achieved on the best parameters were reported.

2.11. Five-fold cross-validation

The datasets were split into 80:20 ratio, in which 80% of the data was utilised for training and 20% for validation purposes. Five-fold cross-validation (CV) was used on 80% training dataset to train, test and evaluate the classification models. The training dataset is split into five equal folds, four of which are utilised for training and the remaining for testing. This method is iterated five times, with each fold being tested separately. This method has been extensively applied in several studies by researchers in the past [50,60]. The complete workflow of IL5pred is shown in Fig. 2.

2.12. Performance evaluation metrics

Model evaluation is a crucial stage in determining the model's efficiency. Thus, the performance of different ML models was evaluated using standard evaluation metrics, including threshold-dependent and independent parameters. The threshold-dependent parameters consist of Accuracy (Acc), Sensitivity (Sens), Specificity (Spec), and Matthews correlation coefficient (MCC), and the Area under the receiver operating characteristic curve (AUC) is a threshold-independent metric. Sens (equation (3)) is also known as recall and can be defined as the true

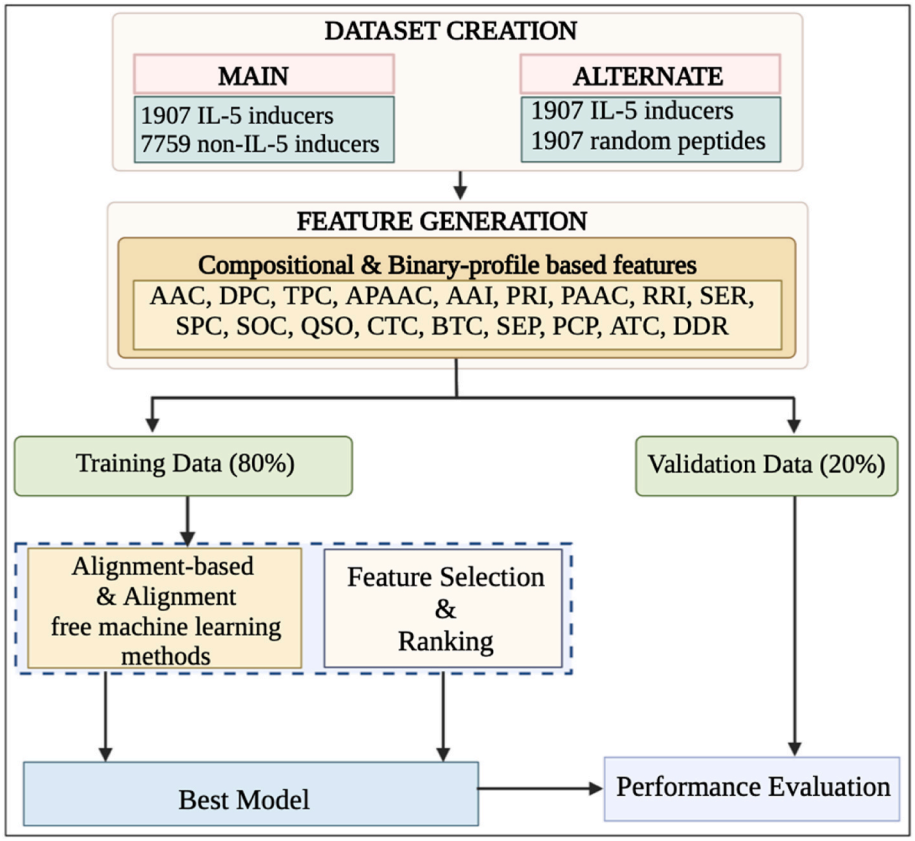


Fig. 2. Complete architecture of IL5pred, including dataset collection, feature generation, machine learning techniques and performance evaluation.

positive rate, while Spec (equation (4)) is the true negative rate. Acc (equation (5)) denotes the percentage of the correctly predicted IL-5 inducers and non-inducers, and MCC (equation (6)) is the relation between the predicted and actual values. These performance metrics are commonly used and well-annotated in the previous studies [61,62] and can be calculated as:

$$\text{Sens} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100 \tag{3}$$

$$\text{Spec} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100 \tag{4}$$

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}} \times 100 \tag{5}$$

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

TP, TN, FP, and FN annotate true positive, true negative, false positive, and false negative respectively.

2.13. Hybrid or ensemble approach

The study explores the potential of a hybrid model that integrates BLAST, motif and alignment-free approach using ML. Initially, BLAST was used to predict the peptide sequence at E-value of 10^{-1}. The scores of '+0.5' and '-0.5' were assigned for the correct positive and negative predictions respectively, and the score of '0' was provided for no hits. Second, the MERCI program was used to classify the same peptide sequence, and the scores of '+0.5' and '-0.5' were assigned if the motifs were present in positive and negative datasets and 0 for absent. The peptide sequences not identified using BLAST (no hits) and MERCI were further predicted using ML models. Ultimately, the overall score was calculated by combining BLAST, MERCI as well as ML prediction scores. Then, the peptides were assigned as IL-5 inducers and non-IL-5 inducers based on the overall score. This hybrid approach has been applied in different research studies in recent years [60,63].

3. Results

3.1. Compositional analysis

The AAC for IL-5 inducing and non-IL-5 inducing peptides was computed. It has been found that the average composition of Phe, Gly, Ile, Lys, Asn, Arg, and Tyr are higher in IL-5 inducing peptides. In contrast, residues like Ala, Thr, Glu, Pro and Val are not preferred in IL-5 inducing peptides. Additionally, we also computed the AAC of random peptides generated from Swiss-Prot. The average AAC for IL-5 inducing, non-IL-5 inducing and random peptides is depicted in Fig. 3.

3.2. HLA alleles distribution analysis

The distribution of epitopes from IEDB database was also examined for different HLA alleles and is depicted in Fig. 4. It has been observed that 9 alleles were found common in both IL-5 inducing and non-IL-5 inducing assays. 24 peptides showed IL-5 inducing whereas only one peptide showed non-IL-5 inducing response against HLA-DRB1*04:01 allele. Similarly, the assays against the HLA-DR allele included 22 IL-5 inducing peptides and 9 non-IL-5 inducing peptides.

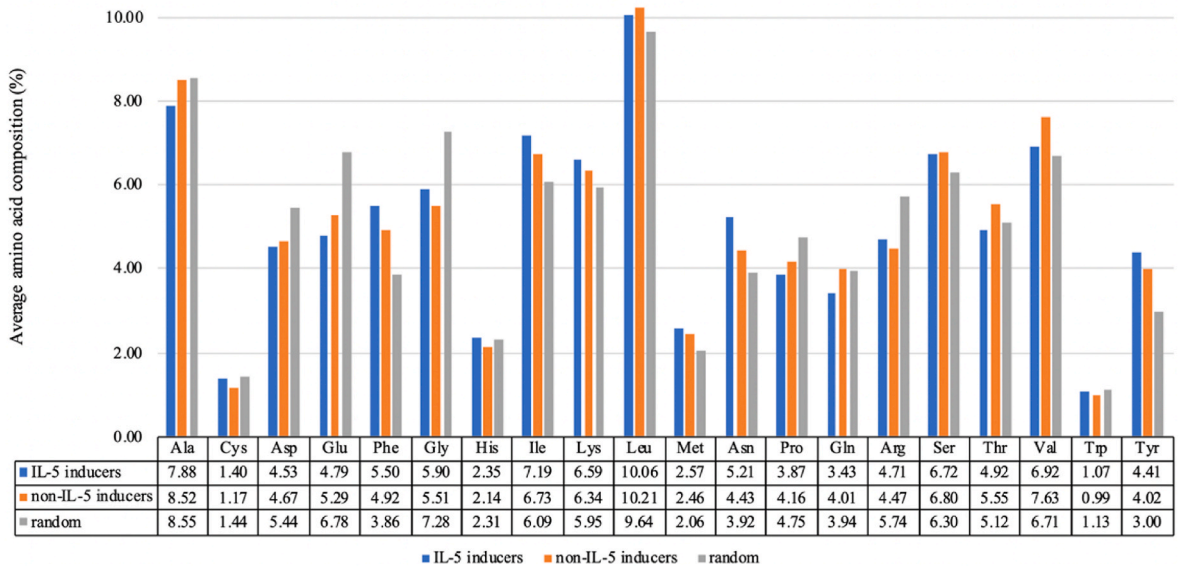
3.3. Positional preference of residues

We generated the sequence logo for IL-5 inducing peptides to identify the positional preference of individual amino acid residues at specific positions. This analysis exhibited that hydrophobic residues such as Leu, Ile, Val, Phe and Ala are highly predominant in IL-5 inducing peptides. The sequence logo of 9 N-terminal residues of IL-5 inducers is depicted in Fig. 5.

3.4. Alignment-based approach

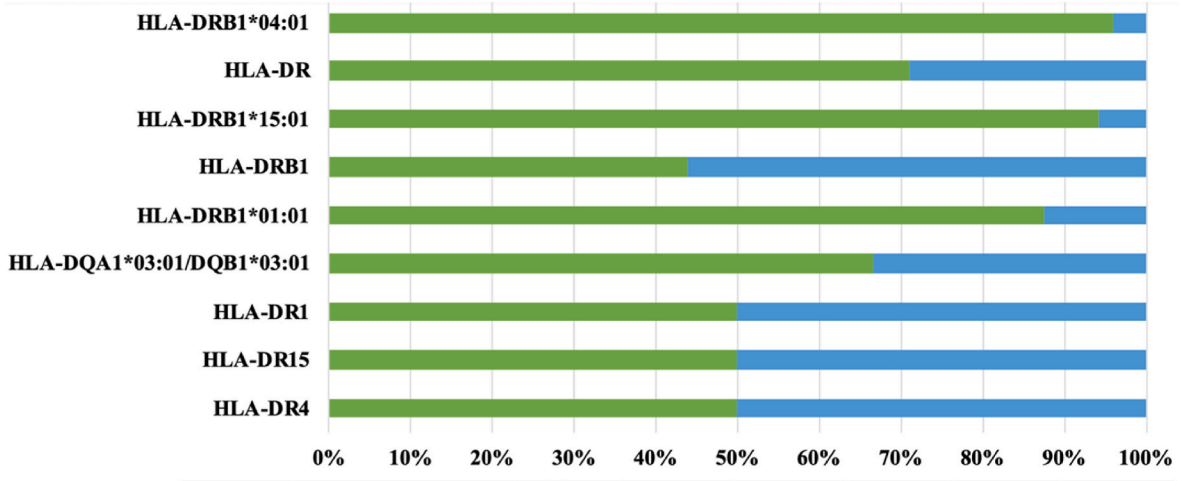
3.4.1. BLAST similarity search

The study utilised BLAST for developing alignment-based models to distinguish between IL-5 inducing and non-IL-5 inducing peptides. A five-fold CV was carried out for evaluating the performance of BLAST. The training dataset was divided into five equal folds, out of which the peptides in four folds were used to build BLAST database, and the sequences in the fifth fold were searched against that database. This process has been iterated five times to avoid any biasness. To evaluate the



	Ala	Cys	Asp	Glu	Phe	Gly	His	Ile	Lys	Leu	Met	Asn	Pro	Gln	Arg	Ser	Thr	Val	Trp	Tyr
IL-5 inducers	7.88	1.40	4.53	4.79	5.50	5.90	2.35	7.19	6.59	10.06	2.57	5.21	3.87	3.43	4.71	6.72	4.92	6.92	1.07	4.41
non-IL-5 inducers	8.52	1.17	4.67	5.29	4.92	5.51	2.14	6.73	6.34	10.21	2.46	4.43	4.16	4.01	4.47	6.80	5.55	7.63	0.99	4.02
random	8.55	1.44	5.44	6.78	3.86	7.28	2.31	6.09	5.95	9.64	2.06	3.92	4.75	3.94	5.74	6.30	5.12	6.71	1.13	3.00

Fig. 3. Amino acid composition of IL-5, non-IL-5 and random peptides.



	HLA-DR4	HLA-DR15	HLA-DR1	HLA-DQA1*03:01/DQB1*03:01	HLA-DRB1*01:01	HLA-DRB1	HLA-DRB1*15:01	HLA-DR	HLA-DRB1*04:01
IL-5 inducers	1	2	3	2	7	11	16	22	24
non-IL-5 inducers	1	2	3	1	1	14	1	9	1

Fig. 4. Distribution of HLA alleles among assays reporting IL-5 inducers and non-IL-5 inducers. Green color denotes IL-5 inducers and blue color indicates non-IL-5 inducers.

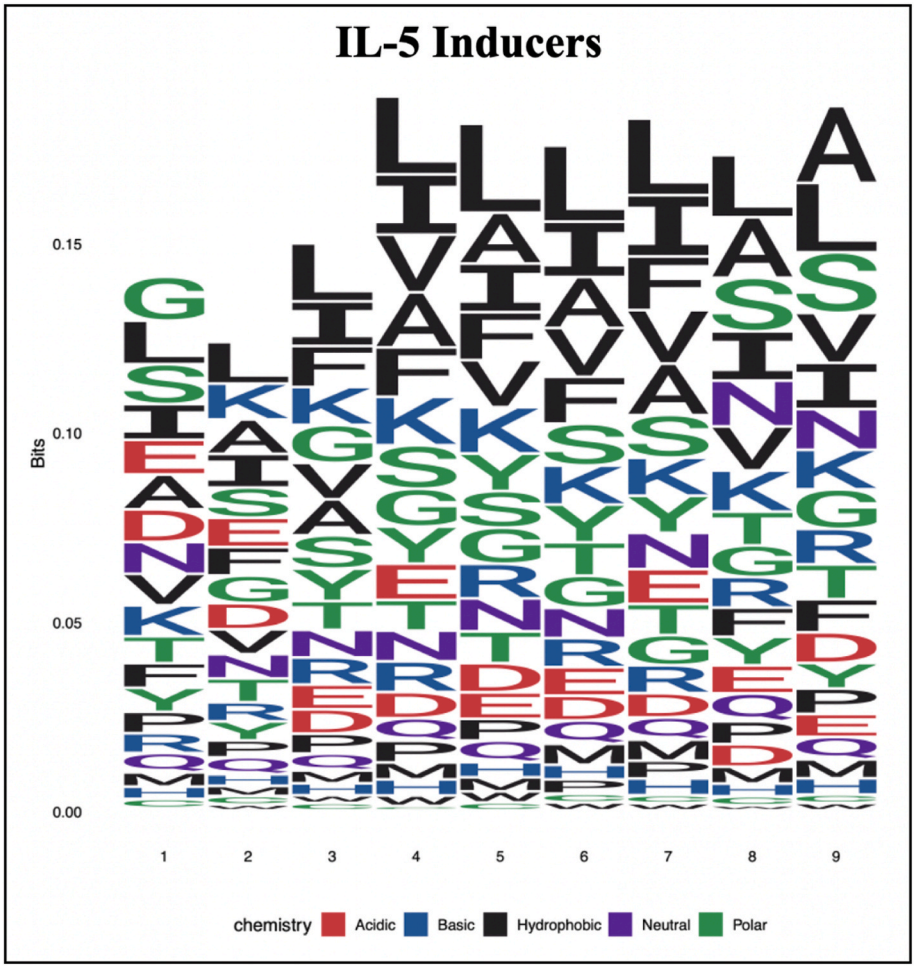


Fig. 5. Sequence logo depicting positional conservation of nine amino acid residues at N- terminus for IL-5 inducers. Amino acid residues such as Leu, Ile, Val, Phe and Ala are highly predominant in IL-5 inducing peptides.

performance of BLAST on the validation dataset, a BLAST database was generated using all the sequences of the training dataset and then each sequence in the validation dataset was searched against it. The peptide is assigned as IL-5 and non-IL-5 inducer based on the top hit generated by BLAST. For the main dataset, the number of correct hits (sensitivity) rises from 0.13% to 5.56% for the training and 0.1%–6.72% for the validation datasets. Besides this, it has attained a specificity of 36.39% for training and 40.74% for validation datasets with E-value ranging from 10^{-6} to 10^{-1}. As shown in Table 1, the wrong hits (error) increase proportionally with the increase in correct hits for both datasets. Hence, it can be inferred from the result that alignment-based method using BLAST alone is not competent in classifying IL-5 and non-IL-5 inducers since it produces a large number of wrong hits as well as no hits.

3.4.2. Motif-based approach

In order to extract the motifs solely found in IL-5 inducing and non-IL-5 inducing peptides of the main dataset, we have used the MERCI program. The motifs such as ENSL, LYVGS, HFFN, and NANR are exclusively present in IL-5 inducing peptides. Alternatively, VGL, YYA, RSP, and PAG motifs are solely found in non-IL-5 inducing peptides. The detailed results are provided in Table 2.

3.5. Alignment-free approach

3.5.1. Binary-profile based features

We also computed binary-profile-based features to develop alignment-free classification model using several ML techniques. We have built ML models for the N_9, C_9 and also for combined terminal residues N_9C_9. We found that RF-based model developed using binary profile of C_9 achieved an AUC of 0.57 on training and 0.58 on validation dataset. RF-model developed using binary profile of N_9 achieved an AUC of 0.55 on training and 0.57 on validation dataset. In the case of binary profile of N_9C_9, XGB-based model achieved an AUC of 0.59 on training and validation datasets. The performance of ML models developed using binary-profile-based features on main dataset is given in Table 3, while the result for the alternate dataset is provided in Table S2. We found that the binary profile-based model does not perform quite well when classifying the peptide sequences.

3.5.2. All compositional features

A vector of 9553 features was computed against each sequence for both main and alternate datasets using Pfeature standalone tool [38]. We computed composition-based features including AAC, DPC, TPC and others to distinguish IL-5 and non-IL-5 inducing peptides. These 9553 features were used to develop alignment-free classification model using several ML techniques. The performance of ML models developed using all compositional features on the main dataset is listed in Table 4. We found that LR-based model attained an AUC of 0.68 on training and validation datasets. In addition, the performance of ML models developed using all compositional features on the alternate dataset is tabulated in Table S3.

Alternatively, SVC-L1 method was then used to reduce these features to 318 (main dataset) and 164 (alternate dataset). These selected features for the main dataset were utilised to build several prediction models and their performance is tabulated in Table 5. The performance of machine learning models developed using selected features on the alternate dataset is shown in Table S4.

Further, we ranked these reduced features using feature-selector tool based on their importance. The detailed results for the top-ranked features are tabulated in Table S5. Based on these top-ranked features (10, 50, 100, 150, ….), several classification models were developed. The performance of the best models developed using the top-ranked features on the main dataset is listed in Table 6. In addition, the performance of other classification models developed using these ranked features on the main and alternate datasets was also computed and tabulated in Tables S6 and S7.

3.5.3. Different types of compositional features

These compositional features were used to develop alignment-free classification model using several ML techniques. The performance of the best models developed using different composition-based features on the main dataset is listed in Table 7. Among all, we found that DPC-based RF model achieved an AUC of 0.74 on training and validation datasets. It could be inferred from the results that alignment-free ML-based model using DPC has performed significantly better than other compositional-based features models. Furthermore, the performance of other classification models developed using compositional-based features on the main and alternate datasets was computed and tabulated in Supplementary Tables S8 and S9.

3.5.4. Cascade machine-learning model

In addition, we employed a bilayer cascade ML-based approach to distinguish IL-5 and non-IL-5 with improved accuracy. In this approach, the prediction was carried out using two layers of machine learning algorithms. In the first layer, 108 ML models were generated by amalgamating the prediction scores of 18 different types of features computed using six different ML techniques. The second layer has been trained on the scores generated by the first layer. Hence, the second

Table 2
Motifs exclusively present in IL-5 inducing and non-IL-5 inducing peptide sequences.

S. No.	Motifs (IL-5 inducers)	No. of sequences	Motifs (Non-IL-5 inducers)	No. of sequences
1	E N S L	10	V G L	42
2	L E N S	10	Y Y A	35
3	L E N S L	10	R S P	34
4	H F F N	7	P A G	33
5	L Y V G S	7	V K P	31
6	L Y V G S K	7	F R V	29
7	L Y V G S K T	7	V F A	29
8	L Y V G S K T K	7	M L R	28
9	N A N R	7	I G K	27
10	V L E N S	7	L K Y	27
11	V L E N S L	7		

Table 1
The performance of alignment-based method, developed using BLAST-based similarity on the main dataset.

E-value	Training				Validation			
	IL-5 inducers		non-IL-5 inducers		IL-5 inducers		non-IL-5 inducers	
	Chits* (Sens)	Whits (error)	Chits (Spec)	Whits (error)	Chits (Sens)	Whits (error)	Chits (Spec)	Whits (error)
10^{-1}	**430 (5.56%)**	**340 (4.4%)**	**2814 (36.39%)**	**354 (4.58%)**	**130 (6.72%)**	**99 (5.02%)**	**788 (40.74%)**	**97 (5.12%)**
10^{-2}	249 (3.22%)	207 (2.68%)	1468 (18.99%)	219 (2.83%)	80 (4.14%)	66 (3.31%)	368 (19.03%)	64 (3.41%)
10^{-3}	149 (1.93%)	127 (1.64%)	729 (9.43%)	129 (1.67%)	49 (2.53%)	37 (1.71%)	193 (9.98%)	33 (1.91%)
10^{-4}	98 (1.27%)	73 (0.94%)	375 (4.85%)	71 (0.92%)	36 (1.86%)	20 (0.98%)	81 (4.19%)	19 (1.03%)
10^{-5}	28 (0.36%)	25 (0.32%)	70 (0.91%)	27 (0.35%)	12 (0.62%)	10 (0.47%)	13 (0.67%)	9 (0.52%)
10^{-6}	10 (0.13%)	2 (0.03%)	8 (0.1%)	4 (0.05%)	2 (0.1%)	3 (0.1%)	0 (0%)	2 (0.16%)

* Chits: Correct hits; Whits: Wrong hits; Sens: Sensitivity; Spec: Specificity.

Table 3
The performance of machine learning-based models developed using binary profile of terminal residues of peptides on main dataset.

C_9										
ML	Training					Validation				
	Sens	Spec	Acc	AUC	MCC	Sens	Spec	Acc	AUC	MCC
DT	26.65	76.98	67.11	0.52	0.03	26.6	78.48	67.99	0.53	0.05
RF	52.9	56.19	55.55	0.57	0.07	58.82	53.27	54.4	0.58	0.1
LR	51.65	53.91	53.47	0.54	0.04	54.99	52.95	53.36	0.55	0.06
XGB	53.03	54.28	54.04	0.55	0.06	55.5	52.17	52.84	0.55	0.06
KNN	67.15	36.65	42.63	0.53	0.03	67.01	36.49	42.66	0.53	0.03
GNB	53.03	52.09	52.28	0.53	0.04	58.31	50.49	52.07	0.57	0.07
N_9										
DT	50.59	51.43	51.27	0.52	0.02	58.31	44.98	47.67	0.53	0.03
RF	50.53	55.84	54.8	0.55	0.05	54.73	53.79	53.98	0.57	0.07
LR	47.89	57.08	55.28	0.54	0.04	44.5	55.8	53.52	0.52	0
XGB	52.77	53.27	53.17	0.54	0.05	51.15	51.72	51.6	0.52	0.02
KNN	49.87	56.66	55.33	0.55	0.05	53.2	53.66	53.57	0.57	0.06
GNB	50.86	51.05	51.01	0.52	0.02	54.22	46.66	48.19	0.52	0.01
N_9C_9										
DT	24.27	78.88	68.17	0.52	0.03	27.88	78.61	68.36	0.53	0.06
RF	57.65	54.1	54.8	0.59	0.09	62.4	50.23	52.69	0.58	0.1
LR	53.5	52.61	52.78	0.54	0.05	59.08	52.11	53.52	0.56	0.09
XGB	54.95	56.93	56.54	0.59	0.1	55.75	55.02	55.17	0.59	0.09
KNN	66.56	40.44	45.56	0.56	0.06	62.4	40.31	44.78	0.54	0.02
GNB	51.65	52.56	52.38	0.53	0.03	59.08	48.74	50.83	0.56	0.06

*Sens: Sensitivity; Spec: Specificity; Acc: Accuracy; MCC: Matthews correlation coefficient; AUC: Area under receiver operating characteristic curve; RF: Random Forest; LR: Logistic Regression; XGB: eXtreme Gradient Boosting; KNN: K-nearest neighbour; GNB: Gaussian Naïve Bayes; DT: Decision Tree.

Table 4
The performance of machine learning models developed using all compositional features on the main dataset.

ML	Parameters	Training					Validation				
		Sens	Spec	Acc	AUC	MCC	Sens	Spec	Acc	AUC	MCC
DT	c = entropy, md = 10, mf = auto	52.97	53.02	53.01	0.55	0.05	80.31	29.29	39.61	0.57	0.09
RF	c = entropy, md = 100, mf = log2, ne = 1000	61.74	62.61	62.44	0.68	0.20	61.38	59.17	59.62	0.67	0.17
LR	C = 0.01, p = l2, s = liblinear	**63.65**	**64.24**	**64.12**	**0.68**	**0.23**	**64.19**	**62.22**	**62.62**	**0.68**	**0.21**
XGB	g = 0.5, it = gain, ne = 1000	62.14	59.25	59.82	0.65	0.17	62.66	58.78	59.57	0.65	0.17
KNN	a = auto, m = Manhattan, nn = 91, w = distance	53.63	54.09	54.00	0.56	0.06	61.13	50.94	53.00	0.58	0.10
GNB	var_smo = 1e-09	41.29	75.87	69.09	0.59	0.15	48.08	76.28	70.58	0.62	0.22

*Sens: Sensitivity; Spec: Specificity; Acc: Accuracy; MCC: Matthews correlation coefficient; AUC: Area under receiver operating characteristic curve; RF: Random Forest; LR: Logistic Regression; XGB: eXtreme Gradient Boosting; KNN: K-nearest neighbour; GNB: Gaussian Naïve Bayes; DT: Decision Tree, c: criterion; md: max_depth; mf: max_features; s: solver; g: gamma; it = importance_type; ne: n_estimators; m: metric; w: distance; nn: n_neighbors; var_smo: variance smoothing; p: penalty; a: algorithm.

Table 5
The performance of machine learning models developed using selected features on the main dataset.

318 Features											
ML	Training						Validation				
	Threshold	Sens	Spec	Acc	AUC	MCC	Sens	Spec	Acc	AUC	MCC
DT	0.48	57.98	53.96	54.75	0.58	0.10	49.87	60.60	58.43	0.56	0.09
RF	0.19	69.86	65.03	65.97	0.74	0.28	68.80	59.88	61.69	0.70	0.23
LR	0.48	67.41	66.92	67.02	0.74	0.28	61.64	62.87	62.62	0.67	0.20
XGB	0.17	63.65	63.05	63.17	0.70	0.22	62.40	61.05	61.32	0.67	0.19
KNN	0.18	63.52	59.20	60.05	0.65	0.18	61.89	56.32	57.45	0.62	0.15
GNB	1	93.27	29.46	41.97	0.61	0.21	85.68	27.48	39.25	0.57	0.12

*Sens: Sensitivity; Spec: Specificity; Acc: Accuracy; MCC: Matthews correlation coefficient; AUC: Area under receiver operating characteristic curve; RF: Random Forest; LR: Logistic Regression; XGB: eXtreme Gradient Boosting; KNN: K-nearest neighbour; GNB: Gaussian Naïve Bayes; DT: Decision Tree.

Table 6
Performance of the best random forest-based models developed on top-ranked features of the main dataset.

Features	Parameters	Training					Validation				
		Sens	Spec	Acc	AUC	MCC	Sens	Spec	Acc	AUC	MCC
Top10	c = entropy, md = 100, mf = auto, ne = 1000	54.02	56.90	56.34	0.59	0.09	57.55	55.54	55.95	0.60	0.11
Top50	c = entropy, md = 50, mf = log2, ne = 1000	64.18	59.44	60.37	0.67	0.19	64.19	56.12	57.76	0.65	0.16
Top100	c = entropy, md = 50, mf = log2, ne = 500	62.53	65.15	64.64	0.69	0.22	62.66	63.38	63.24	0.67	0.21
Top150	c = entropy, md = 50, mf = log2, ne = 1000	64.51	64.11	64.19	0.71	0.23	66.24	60.73	61.84	0.70	0.22
Top200	c = entropy, md = 50, mf = log2, ne = 1000	64.91	66.06	65.83	0.73	0.25	65.47	62.67	63.24	0.70	0.23

*Sens: Sensitivity; Spec: Specificity; Acc: Accuracy; MCC: Matthews correlation coefficient; AUC: Area under receiver operating characteristic curve; c: criterion; md: max_depth; mf: max_features; ne: n_estimators.

Table 7
Performance of best machine learning-based models developed on compositional-based features of the main dataset.

Features	ML	Training					Validation				
		Sens	Spec	Acc	AUC	MCC	Sens	Spec	Acc	AUC	MCC
AAC	RF	61.61	61.73	61.71	0.67	0.19	63.17	59.69	60.39	0.66	0.19
AAI	RF	59.57	58.72	58.89	0.63	0.15	60.61	56.38	57.24	0.62	0.14
APAAC	RF	63.26	61.68	61.99	0.68	0.20	60.10	59.37	59.51	0.65	0.16
ATC	GNB	54.29	51.98	52.43	0.54	0.05	57.55	47.89	49.85	0.54	0.04
BTC	XGB	49.54	60.28	58.17	0.57	0.08	49.11	59.88	57.70	0.56	0.07
CTC	RF	59.76	59.57	59.61	0.64	0.16	62.40	56.77	57.91	0.65	0.15
DDR	RF	56.86	59.27	58.80	0.62	0.13	59.59	56.90	57.45	0.62	0.13
DPC	**RF**	**68.01**	**66.20**	**66.56**	**0.74**	**0.28**	**68.80**	**66.11**	**66.65**	**0.74**	**0.29**
PAAC	RF	61.21	60.65	60.76	0.67	0.18	63.17	60.53	61.07	0.66	0.19
PCP	RF	58.58	62.56	61.78	0.66	0.17	58.31	61.24	60.65	0.66	0.16
PRI	RF	55.34	58.67	58.02	0.60	0.11	52.43	55.74	55.07	0.56	0.07
QSO	RF	62.01	57.26	58.19	0.64	0.15	63.17	54.18	56.00	0.62	0.14
RRI	RF	56.99	60.10	59.49	0.62	0.14	54.73	57.29	56.77	0.59	0.10
SEP	DT	51.19	58.41	57.00	0.57	0.08	53.45	52.60	52.74	0.56	0.05
SER	RF	61.94	60.76	60.99	0.67	0.18	60.87	60.34	60.45	0.66	0.17
SOC	LR	53.83	49.36	50.23	0.52	0.03	52.69	48.74	49.54	0.51	0.01
SPC	RF	61.28	58.70	59.21	0.65	0.16	60.10	60.01	60.03	0.66	0.16
TPC	RF	68.67	59.86	61.59	0.71	0.23	71.10	58.59	61.12	0.71	0.24

*Sens: Sensitivity; Spec: Specificity; Acc: Accuracy; MCC: Matthews correlation coefficient; AUC: Area under receiver operating characteristic curve; RF: Random Forest; LR: Logistic Regression; XGB: eXtreme Gradient Boosting; GNB: Gaussian Naïve Bayes; DT: Decision Tree.

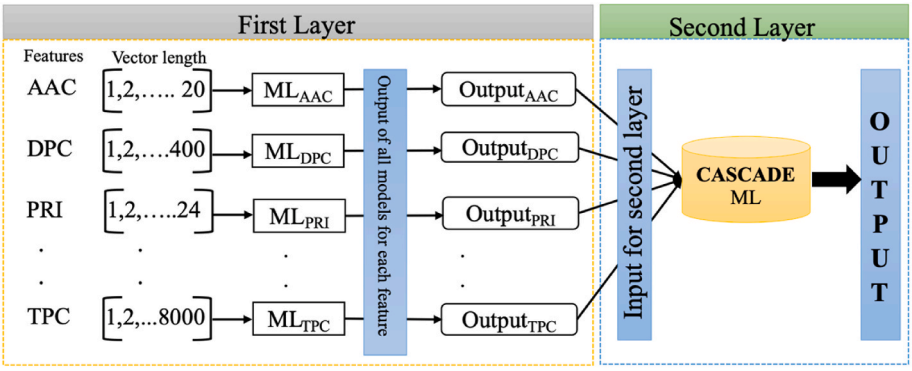


Fig. 6. Diagrammatic representation of bilayer cascade ML-based approach. The first layer consists of different ML models based on individual features. The second layer has been trained on the scores of the output generated by the first layer.

layer learns from the results of the first layer classifiers and produces a final cascade ML model [64] (Fig. 6). We found that among all the ML models, XGB-based classifier attained an AUC of 0.69 on training and 0.67 on validation dataset for the main dataset (Table S10). The performance of the alternate dataset using cascade approach is given in Table S10.

3.5.5. *Selected dipeptide compositional features*

Since we achieved the best performance on DPC features using RF-based model, thus we utilised these features to enhance the classification performance. Feature-selector tool was used to rank these features based on their importance. Using these top-ranked features, we developed several machine learning-based models and their performances are tabulated in Table 8. We observed that RF-based model on DPC-250 features attains the best performance with an AUC of 0.74 on training

Table 8
The performance of best random forest-based models developed using the composition of selected dipeptides.

Features	Parameters	Training				Validation					
		Sens	Spec	Acc	AUC	MCC	Sens	Spec	Acc	AUC	MCC
Top50	c = entropy, md = 50, mf = log2, ne = 500	58.64	60.15	59.86	0.63	0.15	59.34	60.86	60.55	0.64	0.16
Top100	c = entropy, md = 100, mf = log2, ne = 1000	63.33	63.11	63.15	0.69	0.21	67.52	63.19	64.06	0.72	0.25
Top150	c = gini, md = 100, mf = log2, ne = 500	64.05	66.89	66.34	0.72	0.25	69.57	65.91	66.65	0.74	0.29
Top200	c = gini, md = 100, mf = log2, ne = 1000	64.58	68.42	67.67	0.73	0.27	65.22	70.06	69.08	0.74	0.29
Top250	c = gini, md = 100, mf = log2, ne = 800	**65.70**	**68.21**	**67.72**	**0.74**	**0.28**	**66.24**	**69.15**	**68.56**	**0.75**	**0.29**
Top300	c = entropy, md = 100, mf = log2, ne = 800	68.34	65.01	65.66	0.74	0.27	71.36	65.65	66.81	0.75	0.30
Top350	md = 100, mf = log2, ne = 1000	67.81	66.72	66.93	0.74	0.28	68.29	67.14	67.37	0.75	0.29
All400	mf = log2, md = 100, ne = 1000	68.01	66.20	66.56	0.74	0.28	68.80	66.11	66.65	0.74	0.29

*Sens: Sensitivity; Spec: Specificity; Acc: Accuracy; MCC: Matthews correlation coefficient; AUC: Area under receiver operating characteristic curve; c: criterion; md: max_depth; mf: max_features, ne: n_estimators.

and 0.75 on validation dataset respectively, with balanced sensitivity and specificity for main dataset. The same procedure was carried out on alternate dataset and the detailed results are provided in Supplementary Table S11.

We developed several ML-based methods to predict IL-5 using compositional and binary profile-based features computed from the peptide sequence. We could infer from the results that alignment-free ML-based model using DPC-250 features has performed significantly better than other feature-based models. Hence, in the current study, the hybrid model was developed using DPC-250 features of the peptides.

3.5.6. *Hybrid approach*

To combat the shortcomings of individual techniques, we have explored the potential of our hybrid method that combines BLAST, motif with ML using DPC-250 features. This approach was built to classify the IL-5 inducers with high precision. For this, the composition-based model is integrated with MERCI and BLAST-based approaches. Followed by the MERCI approach, peptides were distinguished using BLAST with an E-value of 10^{-1}. The sequence is assigned as IL-5 and non-IL-5 inducers based on the top hit of BLAST, while the peptides that are unpredicted by BLAST were further predicted using DPC-based ML model. This hybrid method considerably improved the accuracy on validation dataset, hence overcoming the limitation of each method as shown in Table 9. RF-based model performed the best on main dataset and achieved AUC of 0.92 and 0.94 on the training and validation dataset. The results for the alternate dataset are provided in Supplementary Table S12.

4. Implementation of IL5pred server

A user-friendly web server IL5pred was developed for predicting IL-5 inducing peptides with better efficiency. The main components of the web server are Predict, Design, Protein scan, Motif scan, and Blast scan. The modules are explained below.

(i) **Predict Module**: Enables the user to submit single/multiple peptide sequences in FASTA format and will predict the peptides as IL-5 inducers or non-IL-5 inducers. The module will provide the user prediction scores and results (IL-5 inducers/non-inducers) based on the chosen threshold value.

(ii) **Design Module**: Enables the user to design novel IL-5 inducers with better activity. The user can provide the input sequence in single-line format. It will generate all possible mutants peptides with a single mutation which are further predicted using the model.

(iii) **Protein scan**: Allows the user to identify IL-5 inducing regions in the given protein sequence by generating overlapping patterns based on the chosen length.

(iv) **Motif scan**: Allows the user to search for the motifs present in IL-5 inducing peptide sequences. This module used MERCI program to extract motifs from the given sequence.

(v) **Blast scan**: Uses an alignment-based search method, BLAST. The module allows to search the query sequence against the database of known IL-5 inducing peptides and assign it as IL-5 inducers based on the match from the database.

The web server can be accessed at (https://webs.iiitd.edu.in/raghava/il5pred/). The python-based standalone package of IL5pred was also developed which can be downloaded from (https://webs.iiitd.edu.in/raghava/il5pred/stand.php). The web server is user-friendly and compatible with all contemporary gadgets, such as smartphones, tablets, and desktops.

5. Discussion

IL-5 is an eosinophil regulatory cytokine involved in the maturation, differentiation, activation and migration of eosinophils [5,65]. It plays a role in a wide variety of diseases such as asthma, dermatitis, eosinophilic esophagitis, and hyper-eosinophilic syndrome [12]. Several autoimmune disorders, such as Hashimoto's thyroiditis and Graves' disease, are also associated with elevated IL-5 levels and eosinophilic infiltration [66]. A recent study by Lucas et al. reported that the level of IL-5 is found to be elevated in patients with severe COVID-19 conditions [67]. Past studies have revealed the role of IL-5 in treating several eosinophil-mediated disorders [68] and possess anti-tumor activity. For

Table 9
Performance of hybrid approach that combines BLAST, motif and selected dipeptide composition-based models on the main dataset.

ML	Threshold	Training					Validation				
		Sens	Spec	Acc	AUC	MCC	Sens	Spec	Acc	AUC	MCC
RF	0.2	79.68	82.14	81.66	0.92	0.54	82.86	84.71	84.33	0.94	0.60
DT	0.38	76.52	76.66	76.63	0.83	0.45	82.35	78.55	79.32	0.86	0.52
GNB	0.5	79.29	74.97	75.81	0.82	0.45	81.59	78.22	78.90	0.84	0.51
KNN	0.24	78.63	80.04	79.76	0.91	0.50	76.47	84.06	82.52	0.92	0.54
XGB	0.16	79.55	79.63	79.62	0.91	0.50	83.12	81.66	81.95	0.93	0.56
LR	0.47	78.3	79.50	79.27	0.89	0.49	81.07	80.43	80.56	0.92	0.53

*Sens: Sensitivity; Spec: Specificity; Acc: Accuracy; MCC: Matthews correlation coefficient; AUC: Area under receiver operating characteristic curve; RF: Random Forest; LR: Logistic Regression; XGB: eXtreme Gradient Boosting; KNN: K-nearest neighbour; GNB: Gaussian Naïve Bayes; DT: Decision Tree.

instance, Ikutani et al. have shown the role of IL-5 inducing cells to control eosinophil infiltration in lungs hence preventing tumor metastasis [69]. IL-5 can act as an appealing therapeutic target due to its pivotal role in the majority of eosinophil-mediated diseases [70–72]. Currently, three FDA-approved anti-IL-5 therapeutic agents, such as benralizumab, mepolizumab and reslizumab, are available in the market [9].

Taking into consideration, the role of IL-5 in several diseases there is a necessity to develop a computational method that can predict IL-5 inducing peptides. In the current study, a prediction method, named IL5pred that can be used to anticipate the peptides as IL-5 inducers and non-IL-5 inducers. For this, we have created two datasets namely main and alternate. The main dataset comprises experimentally validated 1907 IL-5 inducing and 7759 non-IL-5 inducing peptides obtained from IEDB. While the alternate dataset contains 1907 IL-5 inducing peptides from main dataset and 1907 random peptides generated from Swiss-Prot. The compositional analysis shows that IL-5 inducing peptides are abundant in Phe, Gly, Ile, Lys, Asn, Arg, and Tyr. Besides composition, the order of the residue is an essential feature and plays a vital role in defining its activity. For this, we also analyzed the residue preference and observed that hydrophobic residues such as Leu, Ile, Val, Phe and Ala are highly preferred at N-terminus of IL-5 inducing peptides. Along with this, we have extracted the motifs exclusively present in IL-5 inducing peptides.

Moreover, a number of sequence-based features including compositional and binary profiles were calculated for the peptides using Pfeature standalone tool [38]. We computed a total of 9553 features and developed numerous prediction models using several classical ML techniques including LR, RF, DT, GNB, KNN and XGB. LR is a supervised learning algorithm used to predict the probability of a target variable [59]. DT is a non-parametric supervised machine learning algorithm. This classifier predicts the target variable using simple decision rules derived from input features [56]. RF is an ensemble learning method used for classification, that combines multiple DTs during training and predicts a single tree as a response variable. Also, it controls the overfitting of the prediction models [54]. XGB classifier implements the gradient boosting algorithm, where an iterative approach is utilised to predict the final output [58]. GNB is a type of Naïve Bayes algorithm based on the Bayes theorem. This classifier is used when the features have continuous values and assume that the features follow Gaussian normal distributions [57]. KNN is a non-parametric supervised learning classifier that classifies a new data point to the target class closer to the nearest neighbour data points [55]. Further, in our study, we attempt to select a minimum set of features with minimum loss in performance to avoid over-optimization of ML models. Taking this into account, we selected different combinations of feature sets using feature selection and ranking methods to accurately predict IL-5 inducers. We also employed a bilayer cascade ML-based approach to better distinguish IL-5 and non-IL-5. For this, we calculated all possible performance metrics to achieve maximum performance in terms of AUC. Further, we choose threshold values that minimise the discrepancy between sensitivity and specificity on the training dataset and compute threshold-dependent parameters such as accuracy and MCC. We found that RF-based model on DPC-250 features achieved an AUC of 0.74 on training and 0.75 on validation datasets with balanced sensitivity and specificity. This RF-based model that used DPC-250 features outperformed other feature-based models.

Apart from this, the commonly used alignment-based method, BLAST was also employed to annotate the peptide sequence. If the query sequence shows a high degree of similarity with a known peptide, then the same function is assigned to the query peptide. It is found from Table 1 that BLAST can predict IL-5 inducers, but it also produces an enormous number of no-hits. To cope with this limitation, alignment-based, and alignment-free approaches using DPC-250 were combined to develop an improved method. This hybrid method, which combines two or more approaches, improved both AUC and accuracy while circumventing the limitations of individual methods. We observed that the hybrid model outperformed other models in the case of the main dataset. Nonetheless, the hybrid model based on DPC-250, motif and BLAST can accurately distinguish IL-5 and non-IL-5 inducing peptides in terms of accuracy.

6. Conclusion

To expedite the research community working in the field of immunotherapy, we have incorporated the best-performing model into the web server. IL5pred is a user-friendly, freely accessible web server which is compatible with laptops and desktops. We expect that researchers will use our prediction method to develop more accurate peptide-based therapeutics for a variety of diseases.

7. Limitation of the study

This is the first systematic attempt to develop *in silico* models for identifying IL-5 inducing peptides. A large enough dataset is required to develop a more robust and accurate classification method. One limitation of this method is the limited number of experimentally validated IL-5 inducing peptides. We have made an effort to create more accurate and reliable method while keeping this constraint in mind.

Author contributions

NLD collected, compiled, and processed the data sets. NLD and NS developed computer programs. NLD and NS implemented the algorithms and prediction models. NLD and NS created the web server. NLD and NS analyzed the results. NLD, NS and GPSR wrote the manuscript. GPSR conceived and coordinated the project and provided overall supervision of the project. All authors have read and approved the final manuscript.

Data availability statement

All the datasets generated for this study are available at the "IL5pred" webserver, https://webs.iiitd.edu.in/raghava/il5pred/stand.php. The source code is hosted on GitHub and can be found at https://github.com/raghavagps/il5pred.

Declaration of competing interest

The authors declare that they have no conflict of interest.

Acknowledgement

We are thankful to the Department of Biotechnology (DBT) for providing an infrastructure grant to the institute. NLD is thankful to DBT-RA program in Biotechnology and Life Sciences for providing Research Associate fellowship. NS is thankful to the Department of Science and Technology (DST-INSPIRE) for providing Senior Research Fellowship. NLD, and NS are thankful to Department of Computational Biology, IIIT-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.2023.106864.

References

[1] J.S. Marshall, R. Warrington, W. Watson, H.L. Kim, An introduction to immunology and immunopathology, Allergy Asthma Clin. Immunol. 14 (2018) 49, https://doi.org/10.1186/s13223-018-0278-1.

[2] J. Zhu, W.E. Paul, CD4 T cells: fates, functions, and faults, Blood 112 (2008) 1557–1569, https://doi.org/10.1182/blood-2008-05-078154.

[3] T. Kouro, K. Takatsu, IL-5- and eosinophil-mediated inflammation: from discovery to therapy, Int. Immunol. 21 (2009) 1303–1309, https://doi.org/10.1093/intimm/dxp102.

[4] I. Raphael, S. Nalawade, T.N. Eagar, T.G. Forsthuber, T cell subsets and their signature cytokines in autoimmune and inflammatory diseases, Cytokine 74 (2015) 5–17, https://doi.org/10.1016/j.cyto.2014.09.011.

[5] C. Pelaia, G. Paoletti, F. Puggioni, F. Racca, G. Pelaia, G.W. Canonica, E. Heffler, Interleukin-5 in the pathophysiology of severe asthma, Front. Physiol. 10 (2019) 1514, https://doi.org/10.3389/fphys.2019.01514.

[6] M.L. Moore, R.S. Peebles, Interleukins | IL-5, in: Encycl. Respir. Med., Elsevier, 2006, pp. 359–363, https://doi.org/10.1016/B0-12-370879-6/00476-2.

[7] M.M. Le Beau, R.S. Lemons, R. Espinosa, R.A. Larson, N. Arai, J.D. Rowley, Interleukin-4 and interleukin-5 map to human chromosome 5 in a region encoding growth factors and receptors and are deleted in myeloid leukemias with a del(5q), Blood 73 (1989) 647–650. http://www.ncbi.nlm.nih.gov/pubmed/2783863.

[8] T. Adachi, R. Alam, The mechanism of IL-5 signal transduction, Am. J. Physiol. 275 (1998) C623–C633, https://doi.org/10.1152/ajpcell.1998.275.3.C623.

[9] S. Principe, C. Porsbjerg, S. Bolm Ditlev, D. Kjaersgaard Klein, K. Golebski, N. Dyhre-Petersen, Y.E. van Dijk, J.J.M.H. van Bragt, L.L.H. Dankelman, S.-E. Dahlen, C.E. Brightling, S.J.H. Vijverberg, A.H. Maitland-van der Zee, Treating severe asthma: targeting the IL-5 pathway, Clin. Exp. Allergy 51 (2021) 992–1005, https://doi.org/10.1111/cea.13885.

[10] G. Varricchi, G.W. Canonica, The role of interleukin 5 in asthma, Expet Rev. Clin. Immunol. 12 (2016) 903–905, https://doi.org/10.1080/1744666X.2016.1208564.

[11] S. Gopalakrishnan, S. Sen, J.S. Adhikari, P.K. Chugh, T. Sekhri, S. Rajan, The role of T-lymphocyte subsets and interleukin-5 blood levels among Indian subjects with autoimmune thyroid disease., Hormones (Basel). 9 (n.d.) 76–81. https://doi.org/10.14310/horm.2002.1256..

[12] G. Garcia, C. Taillé, P. Laveneziana, A. Bourdin, P. Chanez, M. Humbert, Anti-interleukin-5 therapy in severe asthma, Eur. Respir. Rev. 22 (2013) 251–257, https://doi.org/10.1183/09059180.00004013.

[13] A.E. Yuzhalin, A.G. Kutikhin, Interleukin-3, interleukin-5, and cancer, in: Interleukins Cancer Biol., Elsevier, 2015, pp. 91–116, https://doi.org/10.1016/B978-0-12-801121-8.00004-X.

[14] S.A. Gorski, Y.S. Hahn, T.J. Braciale, Group 2 innate lymphoid cell production of IL-5 is regulated by NKT cells during influenza virus infection, PLoS Pathog. 9 (2013), e1003615, https://doi.org/10.1371/journal.ppat.1003615.

[15] L.V. Rizzo, Differential modulatory effect of IL-5 on MHC class II expression by macrophages and B cells, Brazilian J. Med. Biol. Res. = Rev. Bras. Pesqui. Medicas e Biol. 25 (1992) 509–513. http://www.ncbi.nlm.nih.gov/pubmed/1342227.

[16] A. Le Moine, M. Surquin, F.X. Demoor, J.C. Noël, M.A. Nahori, M. Pretolani, V. Flamand, M.Y. Braun, M. Goldman, D. Abramowicz, IL-5 mediates eosinophilic rejection of MHC class II-disparate skin allografts in mice, J. Immunol. 163 (1999) 3778–3784. http://www.ncbi.nlm.nih.gov/pubmed/10490975.

[17] M. Selvaraja, V.K. Chin, M. Abdullah, M. Arip, S. Amin-Nordin, HLA-DRB1*04 as a risk allele to systemic lupus erythematosus and lupus nephritis in the Malay population of Malaysia, Front. Med. 7 (2020), 598665, https://doi.org/10.3389/fmed.2020.598665.

[18] H. Singh, G.P.S. Raghava, ProPred1: prediction of promiscuous MHC Class-I binding sites, Bioinformatics 19 (2003) 1009–1014, https://doi.org/10.1093/bioinformatics/btg108.

[19] S. Lata, M. Bhasin, G.P.S. Raghava, Application of machine learning techniques in predicting MHC binders, Methods Mol. Biol. 409 (2007) 201–215, https://doi.org/10.1007/978-1-60327-118-9_14.

[20] G. Liu, D. Li, Z. Li, S. Qiu, W. Li, C.-C. Chao, N. Yang, H. Li, Z. Cheng, X. Song, L. Cheng, X. Zhang, J. Wang, H. Yang, K. Ma, Y. Hou, B. Li, PSSMHCpan: a novel PSSM-based software for predicting class I peptide-HLA binding affinity, GigaScience 6 (2017) 1–11, https://doi.org/10.1093/gigascience/gix017.

[21] M. Nielsen, C. Lundegaard, T. Blicher, K. Lamberth, M. Harndahl, S. Justesen, G. Røder, B. Peters, A. Sette, O. Lund, S. Buus, NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence, PLoS One 2 (2007) e796, https://doi.org/10.1371/journal.pone.0000796.

[22] C. Lundegaard, K. Lamberth, M. Harndahl, S. Buus, O. Lund, M. Nielsen, NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11, Nucleic Acids Res. 36 (2008) W509–W512, https://doi.org/10.1093/nar/gkn202.

[23] 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.

[24] 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.

[25] M. Nielsen, O. Lund, Nn-align, An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction, BMC Bioinf. 10 (2009) 296, https://doi.org/10.1186/1471-2105-10-296.

[26] M. Andreatta, E. Karosiene, M. Rasmussen, A. Stryhn, S. Buus, M. Nielsen, Accurate pan-specific prediction of peptide-MHC class II binding affinity with improved binding core identification, Immunogenetics 67 (2015) 641–650, https://doi.org/10.1007/s00251-015-0873-y.

[27] A. Dhall, S. Patiyal, G.P.S. Raghava, HLAncPred: a method for predicting promiscuous non-classical HLA binding sites, Briefings Bioinf. (2022), https://doi.org/10.1093/bib/bbac192.

[28] S. Lata, G.P.S. Raghava, CytoPred: a server for prediction and classification of cytokines, Protein Eng. Des. Sel. 21 (2008) 279–282, https://doi.org/10.1093/protein/gzn006.

[29] S.K. Dhanda, S. Gupta, P. Vir, G.P.S. Raghava, Prediction of IL4 inducing peptides, Clin. Dev. Immunol. 2013 (2013), 263952, https://doi.org/10.1155/2013/263952.

[30] 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, https://doi.org/10.1093/bib/bbaa259.

[31] G. Nagpal, S.S. Usmani, S.K. Dhanda, H. Kaur, S. Singh, M. Sharma, G.P.S. Raghava, Computer-aided designing of immunosuppressive peptides based on IL-10 inducing potential, Sci. Rep. 7 (2017), 42851, https://doi.org/10.1038/srep42851.

[32] O. Singh, W.-L. Hsu, E.C.-Y. Su, ILeukin10Pred: a computational approach for predicting IL-10-inducing immunosuppressive peptides using combinations of amino acid global features, Biology 11 (2021) 5, https://doi.org/10.3390/biology11010005.

[33] S. Gupta, P. Mittal, M.K. Madhu, V.K. Sharma, IL17eScan: a tool for the identification of peptides inducing IL-17 response, Front. Immunol. 8 (2017) 1430, https://doi.org/10.3389/fimmu.2017.01430.

[34] 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, https://doi.org/10.1016/j.compbiomed.2022.105297.

[35] B. Manavalan, T.H. Shin, M.O. Kim, G. Lee, PIP-EL: a new ensemble learning method for improved proinflammatory peptide predictions, Front. Immunol. 9 (2018) 1783, https://doi.org/10.3389/fimmu.2018.01783.

[36] S. Gupta, M.K. Madhu, A.K. Sharma, V.K. Sharma, ProInflam: a webserver for the prediction of proinflammatory antigenicity of peptides and proteins, J. Transl. Med. 14 (2016) 178, https://doi.org/10.1186/s12967-016-0928-3.

[37] R. Vita, S. Mahajan, J.A. Overton, S.K. Dhanda, S. Martini, J.R. Cantrell, D. K. Wheeler, A. Sette, B. Peters, 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, G. Mishra, H. Kaur, N. Sharma, S. Jain, Computing wide range of protein/peptide features from their sequence and structure, bioRxiv (2019), 599126.

[39] UniProt Consortium, UniProt: the universal protein knowledgebase in 2021, Nucleic Acids Res. 49 (2021) D480–D489, https://doi.org/10.1093/nar/gkaa1100.

[40] O. Wagih, ggseqlogo: a versatile R package for drawing sequence logos, Bioinformatics 33 (2017) 3645–3647, https://doi.org/10.1093/bioinformatics/btx469.

[41] 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.

[42] J. Hong, Y. Luo, M. Mou, J. Fu, Y. Zhang, W. Xue, T. Xie, L. Tao, Y. Lou, F. Zhu, Convolutional neural network-based annotation of bacterial type IV secretion system effectors with enhanced accuracy and reduced false discovery, Briefings Bioinf. 21 (2020) 1825–1836, https://doi.org/10.1093/bib/bbz120.

[43] J. Hong, Y. Luo, Y. Zhang, J. Ying, W. Xue, T. Xie, L. Tao, F. Zhu, Protein functional annotation of simultaneously improved stability, accuracy and false discovery rate achieved by a sequence-based deep learning, Briefings Bioinf. 21 (2020) 1437–1447, https://doi.org/10.1093/bib/bbz081.

[44] W. Xia, L. Zheng, J. Fang, F. Li, Y. Zhou, Z. Zeng, B. Zhang, Z. Li, H. Li, F. Zhu, PFmulDL: a novel strategy enabling multi-class and multi-label protein function annotation by integrating diverse deep learning methods, Comput. Biol. Med. 145 (2022), 105465, https://doi.org/10.1016/j.compbiomed.2022.105465.

[45] Y.H. Li, J.Y. Xu, L. Tao, X.F. Li, S. Li, X. Zeng, S.Y. Chen, P. Zhang, C. Qin, C. Zhang, Z. Chen, F. Zhu, Y.Z. Chen, SVM-prot 2016: a web-server for machine learning prediction of protein functional families from sequence irrespective of similarity, PLoS One 11 (2016), e0155290, https://doi.org/10.1371/journal.pone.0155290.

[46] S. Gupta, P. Kapoor, K. Chaudhary, A. Gautam, R. Kumar, Open Source Drug Discovery Consortium, G.P.S. Raghava, in silico approach for predicting toxicity of peptides and proteins, PLoS One 8 (2013), e73957, https://doi.org/10.1371/journal.pone.0073957.

[47] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[48] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, C.-J. Lin, LIBLINEAR: a library for large linear classification, J. Mach. Learn. Res. 9 (2008) 1871–1874.

[49] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.Y. Liu, Lightgbm: a highly efficient gra- dient boosting decision tree, Adv. Neural Inf. Process. Syst. 30 (2017) 3146–3154.

[50] P. Agrawal, D. Bhagat, M. Mahalwal, N. Sharma, G.P.S. Raghava, AntiCP 2.0: an updated model for predicting anticancer peptides, Briefings Bioinf. 22 (2021), https://doi.org/10.1093/bib/bbaa153.

[51] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410, https://doi.org/10.1016/S0022-2836(05)80360-2.

[52] S. Solis-Reyes, M. Avino, A. Poon, L. Kari, An open-source k-mer based machine learning tool for fast and accurate subtyping of HIV-1 genomes, PLoS One 13 (2018), e0206409, https://doi.org/10.1371/journal.pone.0206409.

[53] G.S. Randhawa, M.P.M. Soltysiak, H. El Roz, C.P.E. de Souza, K.A. Hill, L. Kari, Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study, PLoS One 15 (2020), e0232391, https://doi.org/10.1371/journal.pone.0232391.

[54] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, Mach. Learn. 63 (2006) 3–42, https://doi.org/10.1007/s10994-006-6226-1.

[55] A. Mucherino, P.J. Papajorgji, P.M. Pardalos, in: K-Nearest Neighbor Classification, 2009, pp. 83–106, https://doi.org/10.1007/978-0-387-88615-2_4.
[56] J. Fürnkranz, in: C. Sammut, G.I. Webb (Eds.), Decision Tree BT - Encyclopedia of Machine Learning, Springer US, Boston, MA, 2010, pp. 263–267, https://doi.org/10.1007/978-0-387-30164-8_204.
[57] H. Zhang, Exploring conditions for the optimality of naïve bayes, Int. J. Pattern Recogn. Artif. Intell. 19 (2005) 183–198, https://doi.org/10.1142/S0218001405003983.
[58] T. Chen, C. Guestrin, XGBoost, in: Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min, ACM, New York, NY, USA, 2016, pp. 785–794, https://doi.org/10.1145/2939672.2939785.
[59] J. Tolles, W.J. Meurer, Logistic regression: relating patient characteristics to outcomes, JAMA 316 (2016) 533–534, https://doi.org/10.1001/jama.2016.7653.
[60] 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), https://doi.org/10.1093/bib/bbaa294.
[61] S. Saha, G.P.S. Raghava, AlgPred: prediction of allergenic proteins and mapping of IgE epitopes, Nucleic Acids Res. 34 (2006) W202–W209, https://doi.org/10.1093/nar/gkl343.
[62] 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, https://doi.org/10.1016/j.compbiomed.2021.104746.
[63] N. Sharma, L.D. Naorem, S. Jain, G.P.S. Raghava, ToxinPred2: an improved method for predicting toxicity of proteins, Briefings Bioinf. (2022), https://doi.org/10.1093/bib/bbac174.
[64] M. Bhasin, G.P.S. Raghava, Analysis and prediction of affinity of TAP binding peptides using cascade SVM, Protein Sci. 13 (2004) 596–607, https://doi.org/10.1110/ps.03373104.
[65] A. Matucci, A. Vultaggio, E. Maggi, I. Kasujee, Is IgE or eosinophils the key player in allergic asthma pathogenesis? Are we asking the right question? Respir. Res. 19 (2018) 113, https://doi.org/10.1186/s12931-018-0813-0.
[66] T. Lalani, R.K. Simmons, A.R. Ahmed, Biology of IL-5 in health and disease, Ann. Allergy Asthma Immunol. 82 (1999) 317–332, https://doi.org/10.1016/S1081-1206(10)63281-4, quiz 332–3.
[67] C. Lucas, P. Wong, J. Klein, T.B.R. Castro, J. Silva, M. Sundaram, M.K. Ellingson, T. Mao, J.E. Oh, B. Israelow, T. Takahashi, M. Tokuyama, P. Lu, A. Venkataraman, A. Park, S. Mohanty, H. Wang, A.L. Wyllie, C.B.F. Vogels, R. Earnest, S. Lapidus, I. M. Ott, A.J. Moore, M.C. Muenker, J.B. Fournier, M. Campbell, C.D. Odio, A. Casanovas-Massana, , Yale IMPACT Team, R. Herbst, A.C. Shaw, R. Medzhitov, W.L. Schulz, N.D. Grubaugh, C. Dela Cruz, S. Farhadian, A.I. Ko, S.B. Omer, A. Iwasaki, Longitudinal analyses reveal immunological misfiring in severe COVID-19, Nature 584 (2020) 463–469, https://doi.org/10.1038/s41586-020-2588-y.
[68] F. Roufosse, Targeting the interleukin-5 pathway for treatment of eosinophilic conditions other than asthma, Front. Med. 5 (2018) 49, https://doi.org/10.3389/fmed.2018.00049.
[69] M. Ikutani, T. Yanagibashi, M. Ogasawara, K. Tsuneyama, S. Yamamoto, Y. Hattori, T. Kouro, A. Itakura, Y. Nagai, S. Takaki, K. Takatsu, Identification of innate IL-5–producing cells and their role in lung eosinophil regulation and antitumor immunity, J. Immunol. 188 (2012) 703–713, https://doi.org/10.4049/jimmunol.1101270.
[70] D. Bagnasco, M. Ferrando, G. Varricchi, F. Puggioni, G. Passalacqua, G. W. Canonica, Anti-Interleukin 5 (IL-5) and IL-5Ra biological drugs: efficacy, safety, and future perspectives in severe eosinophilic asthma, Front. Med. 4 (2017) 135, https://doi.org/10.3389/fmed.2017.00135.
[71] H. Nagase, S. Ueki, S. Fujieda, The roles of IL-5 and anti-IL-5 treatment in eosinophilic diseases: asthma, eosinophilic granulomatosis with polyangiitis, and eosinophilic chronic rhinosinusitis, Allergol. Int. 69 (2020) 178–186, https://doi.org/10.1016/j.alit.2020.02.002.
[72] A. Harish, S.A. Schwartz, Targeted anti-IL-5 therapies and future therapeutics for hypereosinophilic syndrome and rare eosinophilic conditions, Clin. Rev. Allergy Immunol. 59 (2020) 231–247, https://doi.org/10.1007/s12016-019-08775-4.

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

Neelam Sharma is pursuing her Ph.D. in Computational Biology from the Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India.

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