scientific reports

OPEN

A hybrid method for discovering interferon-gamma inducing peptides in human and mouse

Anjali Dhall[1,2], Sumeet Patiyal[1,2] & Gajendra P. S. Raghava[1✉]

Interferon-gamma (IFN-γ) is a versatile pleiotropic cytokine essential for both innate and adaptive immune responses. It exhibits both pro-inflammatory and anti-inflammatory properties, making it a promising therapeutic candidate for treating various infectious diseases and cancers. We present IFNepitope2, a host-specific technique to annotate IFN-γ inducing peptides, it is an updated version of IFNepitope introduced by Dhanda et al. In this study, dataset used for developing prediction method contain experimentally validated 25,492 and 7983 IFN-γ inducing peptides in human and mouse host, respectively. In initial phase, machine learning techniques have been exploited to develop classification model using wide range of peptide features. Further, to improve machine learning based models or alignment free models, we explore potential of similarity-based technique BLAST. Finally, a hybrid model has been developed that combine best machine learning based model with BLAST. In most of the case, models based on extra tree perform better than other machine learning techniques. In case of peptide features, compositional feature particularly dipeptide composition performs better than one-hot encoding or binary profile. Our best machine learning based models achieved AUROC 0.89 and 0.83 for human and mouse host, respectively. The hybrid model achieved the AUROC 0.90 and 0.85 for human and mouse host, respectively. All models have been evaluated on an independent/validation dataset not used for training or testing these models. Newly developed method performs better than existing method on independent dataset. The major objective of this study is to predict, design and scan IFN-γ inducing peptides, thus server/software have been developed (https://webs.iiitd.edu.in/raghava/ifnepitope2/). This method is also available as standalone at https://github.com/raghavagps/ifnepitope2 and python package index at https://pypi.org/project/ifnepitope2/.

Type II interferons or IFN-γ is a multifunctional pleiotropic cytokine play an essential role in innate and acquired immune responses. IFN-γ is majorly secreted by helper T cells, cytotoxic T cells, natural killer cells, B cells, macrophages[1–5]. Various studies report that, IFN-γ possess both pro-inflammatory and anti-inflammatory properties for example it upregulates MHC-I/II protein expression, modulates the production of T helper cells (Th1), activates tumoricidal activity and inhibit the expression of immunosuppressive cytokine (IL-17)[6–8]. Moreover, it is involved in intracellular communication, tumour cell identification and eradication[1,9]. Interferons have emerged as the potential therapeutic candidates for a number of infectious diseases, including cancer[10,11]. For example, Yan et al., shows that IFN-γ acts as a prognostic marker for PD-1/PD-L1 immune checkpoint inhibitor cancer therapy and associated with the better response to immune checkpoint blockade (ICB) therapy in cancer patients[12,13].

In the recent years, peptide-based cancer immunotherapies have great potential to treat a variety of malignant tumours[14]. Peptides acts as versatile immune candidates that stimulate both innate and adaptive arms of our immune system by interacting with various immune cells such as dendritic cells, NK cells, T cells and B cells[15–18]. In order to generate anti-tumor or anti-viral immune response, the potential peptide-based vaccine candidate should activate the production of Th1 associated cytokines such as IFN-γ, IL-2, IL-12, and TNF-α, and inhibit the production of Th2 cytokines including IL-4, IL-6 and IL-13[19]. Studies also report that production of IFN-γ cytokine is essential to activate the anti-viral and anti-tumor immune responses[3,20,21]. However, the major challenges while designing subunit/peptide-based vaccine is to identify a precise antigenic region or

[1]Department of Computational Biology, Indraprastha Institute of Information Technology, Okhla Industrial Estate, Phase III, (Near Govind Puri Metro Station), New Delhi 110020, India. [2]These authors contributed equally: Anjali Dhall and Sumeet Patiyal. ✉email: raghava@iiitd.ac.in

peptide which can activate a specific arm of immune system[22–25]. The ideal scenario would be to experimentally evaluate the immune response to every potential fragment or peptide of the pathogen proteome. However, the experimental approaches are highly expensive and time-consuming.

Therefore, in order to design subunit vaccines with precision it is important to develop computational methods for the prediction of potential peptides/subunit candidates which can activate the immune responses and production of cytokines[26–29]. Recently, a number of in-silico tools have been developed for the prediction of cytokine inducing peptides such as IFNepitope[30], TNFepitope[31], IL2Pred[32], IL10Pred[33], IL6Pred[34], IL13Pred[35], IL17eScan[36], StackIL6[37], ILeukin10Pred[38], IL5Pred[39], MVIL6[40], Meta-IL4[41]. These prediction methods are utilized by researchers and experimental biologist for the prediction and designing of peptide-based vaccine candidates. In order to complement the existing tools, we have made a systematic attempt to improve our previous method IFNepitope[30]. In the current method, we have proposed a novel host-specific tool for the accurate prediction of IFN-γ inducing or non-inducing peptides in human and mouse hosts. These hosts exhibits variable IFN-γ inducing peptides and have distinct residue compositions. Moreover, we have used a number of composition features, a bigger experimentally validated dataset, hybrid approach to merge alignment-free (machine-learning) and alignment-based method (BLAST) which significantly improve the performance of new method IFNepitope2.

Materials and methods
Overall design
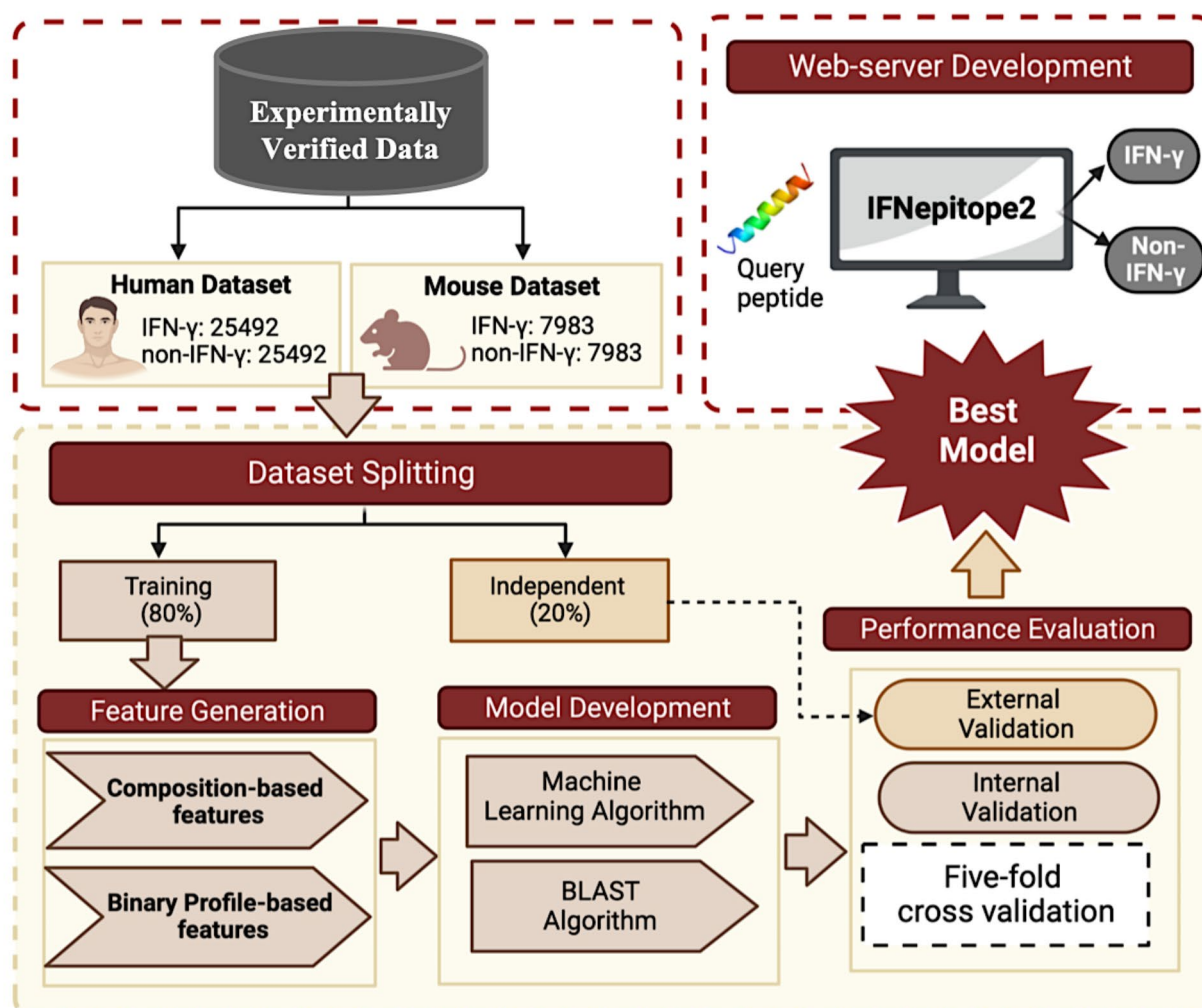The workflow adapted in this study is depicted in Fig. 1.

Fig. 1. The overall architecture of the IFNepitope2, including dataset collection, feature generation, model development.

Dataset preparation and preprocessing

In the current study, we have collected the experimentally validated IFN-γ inducing and non-inducing peptides from the Immune Epitope Database & Tools (IEDB) https://www.iedb.org[42]. At first, we preprocessed the downloaded file by removing the peptides having non-natural amino acids. After conducting the host-based analysis, we found that most of the sequences are related to either the mouse or human hosts. Therefore, we considered sequences belong to human and mouse host for further analysis. Then, we did the analysis based on length and observed that most of the sequences were lying in the length range of 8–20 amino acids, hence, we selected sequences with length range 8–20 and discarded the rest. Further, we removed the redundant IFN-γ inducing and non-inducing peptides from the independent host-specific dataset. Eventually, we were left with 25,492 IFN-γ inducing and 61,680 non-inducing peptides belong to human host, and 7983 IFN-γ inducing and 27,837 non-inducing peptides from mouse host. On further analysis, we have found out that there is high similarity in the IFN-γ inducing and non-inducing sequences in human as well as in mouse host. In order to handle that, we have removed sequences from the non-inducing peptides which either contains IFN-γ inducing peptides or differing with only one or two amino acids. We were left with 25,492 IFN-γ inducing and 41,092 non-inducing peptides in case of human host, and 7983 IFN-γ inducing and 16,121 non-inducing peptides in mouse host. Finally, to make the balanced datasets for further consideration, we have randomly selected the equal number of sequences as of number IFN-γ inducing peptides, from the set of non-inducing peptides for each host organism. The final dataset consists of 25,492 IFN-γ inducing and 25,492 non-inducing peptides from human host, and 7983 IFN-γ inducing and 7983 non-inducing peptide sequences belong to mouse host.

Training and independent dataset

For fair evaluation of the generated models, we have followed the standard procedure used in several studies[43–48] and divided the entire datasets in 80:20 ratio. The 80% of the dataset was used for the training purpose or internal validation and the rest 20% was kept aside for the external validation. In case of human host, the training dataset was comprised of 20,394 IFN-γ inducing and non-inducing peptides where independent dataset contained 5098 IFN-γ inducing peptides and equal number of non-inducing peptides. Similarly, for mouse host, 6387 IFN-γ inducing and 6387 non-inducing peptide sequences were used in the training dataset, and 1596 IFN-γ inducing and non-inducing peptides were used as the independent dataset. Moreover, we tried to maintain the distribution of length of peptides in training and independent dataset.

Composition analysis

To understand the preference of amino acids in the IFN-γ inducing peptides as compare to non-inducing peptides in both the hosts, we have calculated the amino acid composition of each peptide in IFN-γ inducing and non-inducing datasets. Further, we compared the average composition of each residue in IFN-γ inducing and non-inducing dataset for human as well as mouse host. We have implemented the composition-based module of Pfeature[49] to compute the amino acid composition of each peptide.

Positional preference analysis

The composition analysis provides the overall preference of each amino acid in the IFN-γ inducing and non-inducing dataset, but fails to capture the preference of residues at a particular position in the peptides. In order to achieve that, we have generated the sequence logos using the two-sample logo software[50]. One of the major challenges with TSL tool was that it does not takes variable length peptides as input, therefore, it was required to fix the peptide length. The minimum length of the peptides in the dataset was 8, therefore, to generate the fixed length peptides, we have considered the 8 amino acids from the N-terminal and 8 residues from the C-terminal. Further, we joined them to generate peptides of length 16, which were used as the input to generate the sequence logos for both human and mouse datasets.

Feature generation

To build the prediction models with the ability to discriminate the IFN-γ inducing peptides from the non-inducing peptides, it is necessary to represent the sequences using numerical vectors. Thus, we have implemented the composition- and binary profile-based modules of Pfeature to calculate different types of features. We have computed composition-based such as amino acid composition, amphiphilic pseudo amino acid composition, atomic composition, bond composition, composition-enhanced transition distribution, conjoint triad composition, distance distribution of residues, dipeptide composition, pseudo amino acid composition, physicochemical properties composition, quasi-sequence order, residue repeats information, sequence order coupling number, Shannon-entropy of physicochemical properties, and combination of all the features. Additionally, we have also calculated four binary profiles such as binary profile eight residues from N-terminal (N8) as eight was the minimum length of the peptides, binary profile of eight residues from C-terminal (C8), combination of N8 and C8 (N8C8), and binary profile using 20 residues (NC20) where dummy residue "X" was added to make up the length 20 where peptide length was less than 20[43]. We have used all the above-mentioned features to develop the various classification models to classify the IFN-γ inducing peptides.

Model development

In order to build the prediction models to segregate IFN-γ inducing peptides from the non-inducing peptides, we have implemented various machine learning classifiers using scikit-learn library of Python. These classifiers include decision tree (DT), random forest (RF), logistic regression (LR), K-nearest neighbor (KNN), Gaussian naïve Bayes (GNB), extra trees (ET), and support vector classifier (SVC). We have trained different classifiers using the training dataset and evaluated them on the independent dataset.

We have also performed the hyperparameter tuning using grid search approach to identify the best performing parameters for each classifier.

BLAST approach

After that, we have implemented a similarity search approach using BLAST[51], where we categories the epitopes as IFN-inducing or non-inducing based on the similarity of the sequences. In this case, we created a custom database using the makeblastdb suite of NCBI-BLAST + version 2.2.29 and performed similarity searches using blastp suite. Using the training dataset, we built a unique database, and validation dataset sequences were checked against it. We categories the hits as IFN-inducer or non-inducer based on their similarity to the customized database. At the moment, we take only the top-hit of BLAST (i.e., if the top-hit of BLAST is against the IFN-inducer peptide, the query sequence was classified as IFN-inducing peptide or vice-versa. We run BLAST at a variety of e-value cutoffs ranging from 1e-6 to 1e + 3 in order to determine the optimal e-value value score.

Hybrid model

We have used a hybrid strategy that combines BLAST and machine learning-based prediction approach in order to enhance the performance of prediction models. Here, we initially categories the peptide or epitope in accordance with the BLAST query. Then, we integrate '-0.5' score for the incorrect or negative predictions, '+0.5' score for the correct or positive predictions, and '0' value if no hits were discovered. Finally, we integrate the BLAST score and the machine learning prediction score to make the final predictions.

Cross-validation technique

To avoid the overfitting and biasness of the generated prediction models, we have implemented the five-fold cross-validation technique on the training dataset to perform the internal validation. In this technique, the entire training dataset was first divided into five parts out of which four were used to train the model where the remaining one part was used for the testing. The same procedure was repeated five times so that each part gets the chance to be utilized for testing the model[52–54]. The final performance is evaluated by taking the mean of the performances from each iteration.

Parameter optimization

To optimize the performance of the classifiers, we conducted the hyperparameter tuning using grid search with 5-fold stratified cross-validation using scikit-learn library. This process was applied across multiple classifiers, including DT, RF, LR, XGB, KNN, ET, and SVC. For each model, we constructed a tailored parameter grid based on key hyperparameters specific to each classifier. The grid search explored various combinations of these parameters, and for each configuration, 5-fold stratified cross-validation was used to evaluate the model performance. The objective was to identify the best set of hyperparameters that maximized the Area Under the Receiver Operating Characteristic Curve (AUROC). Furthermore, parameters were dynamically adjusted where necessary to ensure compatibility with the data. The complete code used for hyperparameter optimization is available on the GitHub repository at https://github.com/raghavagps/ifnepitope2.

Evaluation metrices

In order to evaluate the performance of the models developed using various machine learning classifiers, we have used the standard performance evaluation parameters. We have computed the threshold-dependent and threshold-independent parameters. Threshold-dependent parameters included sensitivity, specificity, accuracy, F1-score, Kappa, and Matthews correlation coefficients. The equations for the threshold-dependent metrices are represented in Eqs. 1–5. On the other hand, area-under the receiver operating characteristics curve was used as threshold-independent parameter.

$$Sensitivity = \frac{P_{tr}}{P_{tr} + N_{fa}} \tag{1}$$

$$Specificity = \frac{N_{tr}}{N_{tr} + P_{fa}} \tag{2}$$

$$Accuracy = \frac{P_{tr} + N_{fa}}{P_{tr} + N_{tr} + P_{fa} + N_{fa}} \tag{3}$$

$$F1 - score = \frac{2P_{tr}}{2P_{tr} + P_{fa} + N_{fa}} \tag{4}$$

$$MCC = \frac{(P_{tr} * N_{tr}) - (P_{fa} * N_{fa})}{\sqrt{(P_{tr} + P_{fa})(P_{tr} + N_{fa})(N_{tr} + P_{fa})(N_{tr} + N_{fa})}} \tag{5}$$

Where, P_{tr} is true positive, N_{tr} is true negative, P_{fa} is false positive, and N_{fa} is false negative.

Results

Composition and conservation analysis

At first, we computed amino acid composition of IFN-inducing and non-inducing peptides using the human and mouse datasets. Figure 2A and C shows the average amino acid for human and mouse dataset. As shown in Fig. 2A, the average amino acid composition of residues (K, P, N and Q) is significantly higher in IFN-γ
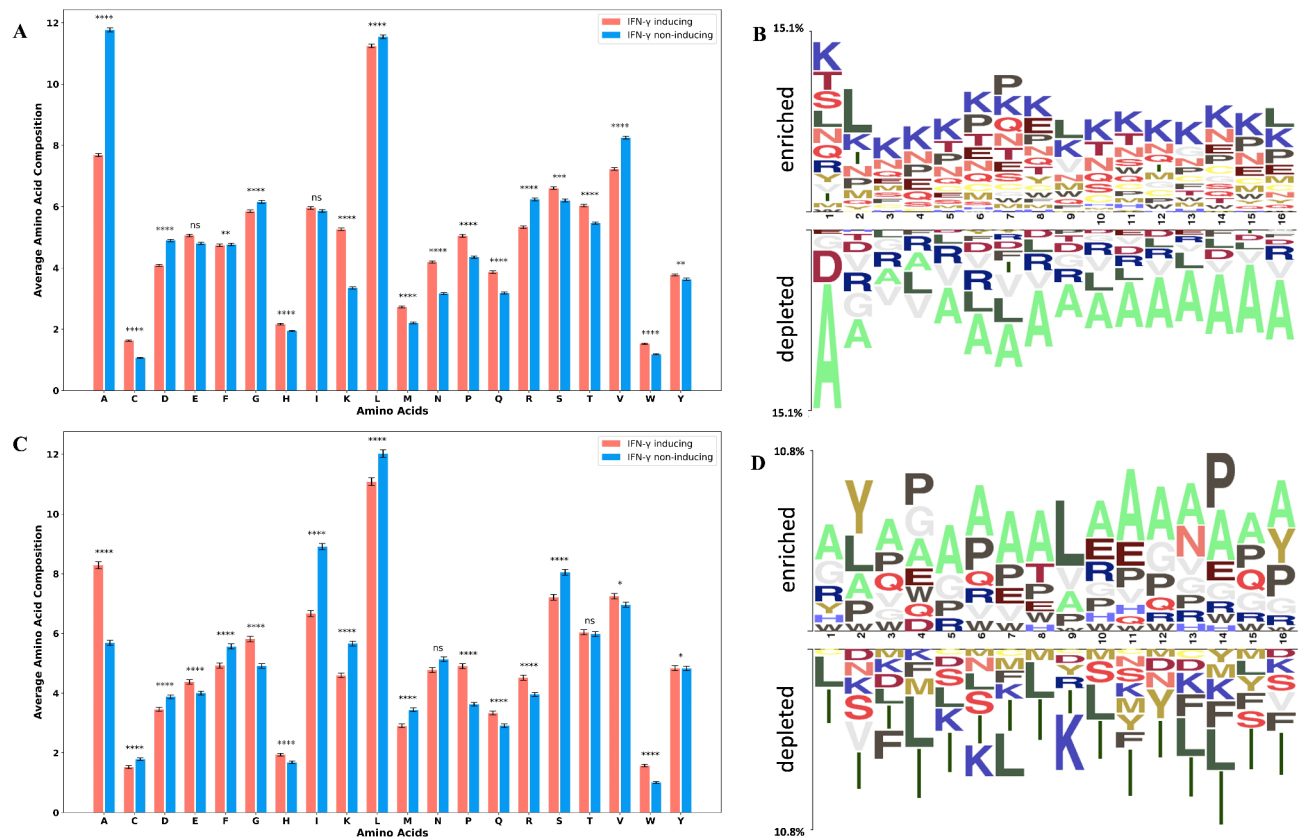
Fig. 2. Average amino-acid composition and two sample logo of IFN-γ inducing and non-inducing peptide, (**A** and **B**) Human dataset (**C** and **D**) Mouse dataset.

inducing peptides in the human host dataset. However, the average composition of residues such as A, G, P and R is significantly higher in mouse IFN-inducing peptides (See Fig. 2C). In the case of human host IFN-inducing epitopes, residue 'K' are highly conserved at the majority of positions, whereas 'P' is preferred at the 6th, 7th, 15th and 16th positions (See Fig. 2B). In IFN-inducing epitopes of mouse host, residues 'A' and 'P' highly dominated at most of the positions (See Fig. 2D).

Motif based analysis

We utilized MERCI software[55] with default parameters for the identification of conserved patterns in the IFN-γ inducing peptides in comparison to the non-inducing peptides and vice-versa. We have performed Motif based analysis independently on human and mouse host datasets. We reported the motifs along with the coverage in the human and mouse host datasets in Supplementary Table S1. The analysis provided a comparative insights into the motif patterns observed in peptides that are either inducing or non-inducing for IFN- γ in both host species. We observed that the conserved patterns have shown the notable differences in human and mouse IFN-γ inducing peptides, indicating that precise and accurate results require predictions on host-specific cytokines. The most frequently occurred conserved patterns in human host are "PQQPF", "PQPQQ" and "QPQLP", which corresponds with the previous version of the IFNepitope2. Whereas, in the mouse host, the most frequently occurred motif patterns in the IFN-γ inducing peptides are "NRDF", "HAM", and "QGPK".

Performance of machine learning models

Various prediction models were developed using seven classifiers such as DT, RF, ET, GNB, KNN, LR and SVC for both human and mouse datasets. We have used Pfeature standalone package to compute binary and composition based features. Then we evaluated the performance on different features as well as combining all the features on training and independent datasets.

Performance of composition-based models

We have developed various composition-based features using human dataset. We observed that ET classifier achieved highest performance for most of the feature types. In Table 1, we have provided the performance of all the features obtained using ET classifier. The complete results of all the classifiers for the different features are given in Supplementary Table S2. As shown in Table 1, the di-peptide composition-based feature achieved highest AUROC of (0.90 and 0.89) and MCC (0.63 and 0.61) on training and independent dataset. ALL_COMB based-features achieved comparable performance in terms of AUROC (0.89 and 0.88) and MCC (0.61 and 0.60) on training and independent datasets. Similarly, AAC, APAAC and PAAC able to achieve equivalent performances

	Training dataset						Independent dataset					
Human dataset												
Feature type	Sens	Spec	ACC	AUROC	F1	MCC	Sens	Spec	ACC	AUROC	F1	MCC
AAC	80.35	79.43	79.89	0.88	0.8	0.6	79.25	78.93	79.09	0.87	0.79	0.58
DPC	80.73	82.29	81.51	0.9	0.82	0.63	81.56	79.38	80.47	0.89	0.81	0.61
ALL_COMB	80.64	80.71	80.67	0.89	0.81	0.61	79.87	80.05	79.96	0.88	0.8	0.6
Mouse dataset												
Feature type	Sens	Spec	ACC	AUROC	F1	MCC	Sens	Spec	ACC	AUROC	F1	MCC
AAC	72.23	73.18	72.7	0.8	0.73	0.45	71.24	72.37	71.81	0.79	0.72	0.44
DPC	79.77	80.21	79.99	0.88	0.8	0.6	76	75.13	75.56	0.83	0.76	0.51
ALL_COMB	73.82	75.48	74.65	0.83	0.74	0.49	73.94	73.87	73.9	0.82	0.74	0.48

Table 1. Performance metrices of top three composition-based features using extra tree (ET) classifier on training and independent dataset of human host and mouse host. *Sens: Sensitivity; Spec: Specificity; Acc: Accuracy; AUROC: Area Under the Receiver Operating Characteristics curve; MCC: Matthews Correlation Coefficient; AAC: Amino Acid Composition; DPC: Dipeptide Composition; ALL_COMP: Combination of All Composition based features.

with an AUROC of 0.88 on training and 0.79 on independent dataset, with balanced sensitivity and specificity. PCP, DDR, QSO, SPC and CETD features perform quite well with an AUROC > 0.80 on training and validation datasets. As shown in Supplementary Table S2, RF and SVC classifier also achieved comparable performances on independent dataset. We found RF achieved maximum AUROC of 0.87 using ALL_COMP features, and SVC achieved maximum AUROC of 0.8 on DPC-based features on training and validation datasets, respectively (See Supplementary Table S2).

In order to identify IFN-inducing epitopes in mouse host, we have developed similar prediction models for mouse datasets. We observed that in case of mouse host, ET based classifier perform well on most of the feature, hence we have provided performance of ET in Table 1. The overall results of other classifiers are provided in Supplementary Table S3. As shown in Table 1, di-peptide based features performed best among the other feature type. We obtained a highest AUROC of 0.88 and 0.83; MCC of 0.60 and 0.51 on training and independent datasets. ALL_COMB features also performed well on both training and validation datasets with an AUROC of 0.83 and 0.82; MCC of 0.49 and 0.48, respectively. Whereas, AAC, APAAC and PAAC performed equivalent with an AUROC of 0.80 and 0.79 on training and validation datasets. Similarly, we observed that RF based models also perform well on independent dataset. As shown in Supplementary Table S3, for AAC and DPC based features RF achieved maximum AUROC of (0.78 and 0.85) and AUROC of (0.77 and 0.83) on training and independent dataset.

Performance of binary profile-based models

In addition, we have generated binary profile-based features and computed performances on binary profile of eight residues from N-terminal (N8), binary profile of eight residues from C-terminal (C8), combination of N8 and C8 (N8C8), and binary profile using 20 residues (NC20). As shown in the Table 2, ET classifier perform best among the other classifiers using NC20 binary profile-based features. It achieves maximum AUROC of 0.86 and MCC of 0.56 on both training and independent datasets. Similarly, N8C8 type features also perform well with an AUROC of 0.80 training and 0.79 independent datasets (See Supplementary Table S4). On the other side, in the case of mouse dataset we were able to achieve maximum AUROC of 0.76 and MCC of 0.37 on independent dataset using NC20 binary profile-based features (See Table 3). In addition, we have obtained comparable performance of AUROC 0.70 and 0.71 using SVC classifier and N8C8 features, on training and independent dataset. The complete results of all the other classifiers for human and mouse hosts is provided in Supplementary Table S4. Of note, we observed that for both human and mouse hosts the di-peptide composition-based features performed best on independent datasets (See Table 1).

Performance of hybrid models

In order to predict IFN-inducing and non-inducing peptides, we have built a hybrid model by combining BLAST and machine learning approach. Here, we have combined di-peptide based predictions and BLAST similarity scores. As we observed di-peptide composition-based features performed best among the other features, in case of both human and mouse hosts (See Table 1). So, we generated hybrid models using di-peptide based features and BLAST alignment score. As shown in Table 2, we have calculated the performances at various e-value cutoffs. In the case of human models, we achieve maximum performance at e-value (e-value (1.00E-06) with AUC of 0.92 and MCC of 0.70 on training datasets and AUC of 0.90 and MCC of 0.66 on independent datasets. In case of mouse dataset, we obtained maximum performance at e-value (1.00E-06) with AUC of 0.89 and MCC of 0.62 on training datasets and AUC 0.85 and MCC 0.55 on independent datasets (See Table 2). The complete results of hybrid models are available in Supplementary Table S5. Moreover, to further validate the robustness of our model, we created an additional dataset, referred to as the "Independent Dataset 2". This dataset was constructed using IFN-γ inducing peptides from the independent dataset and non-inducing peptides that were not previously used in the training and validation stages. For the human host, the dataset comprises 5098 IFN-γ inducing peptides and 33946 non-inducing peptides, while for the mouse host, it includes 1596 IFN-γ inducing peptides and 8139 non-inducing peptides. The model was evaluated on the independent dataset 2 to assess its generalization capabilities. The results demonstrated that the model performs equivalently on the independent dataset 2, achieving an AUC of 0.88 for the human host and 0.82 for the mouse host. These findings further reinforce the reliability and effectiveness of our approach across different datasets and species. All performance measures and best performing parameters for human and mouse hosts were reported in Supplementary Table S6, S7.

Web-server development

For the prediction of IFN-inducing and non-inducing epitopes using sequence information, we have created a web service called "IFNepitope2". The server is developed using HTML, JAVA, and PHP scripts and is compatible with a variety of gadgets, including laptops, iPhones, phones, etc. We have incorporated the hybrid models (DPC + BLAST) at the back-end of the webserver. IFNepitope2 includes three major modules: 'Predict', 'Scan', and 'Design'. The 'Predict' module classify the IFN-inducing and non-inducing peptides; and allow the user to provide peptide sequences in FASTA format. 'Scan' module provides the facility to the user to map/scan the IFN-γ inducing fragments in the protein sequences. 'Design' module facilitates the user to generate the possible mutants of the query sequence and predict the mutated sequence have the capacity to secrete IFN-γ inducing/ non-inducing peptides. Moreover, the training and validation datasets were available on the download page of the webserver. The webserver is available at https://webs.iiitd.edu.in/raghava/ifnepitope2/ link.

Comparison with exiting tools

It is crucial to compare newly developed methods with existing ones to accurately evaluate their respective advantages and disadvantages. Over the years, several cytokine-inducing peptide prediction tools have been

Human dataset

E-value	Training dataset						Independent dataset					
	Sens	Spec	ACC	AUC	F1	MCC	Sens	Spec	ACC	AUC	F1	MCC
1.00E-06	84.54	85.24	84.89	0.92	0.85	0.70	83.31	82.88	83.09	0.90	0.83	0.66
1.00E-05	84.29	85.22	84.76	0.92	0.85	0.70	83.35	82.68	83.01	0.90	0.83	0.66
1.00E-04	84.26	85.10	84.68	0.91	0.85	0.69	83.41	82.62	83.01	0.89	0.83	0.66
1.00E-03	84.15	84.90	84.53	0.91	0.85	0.69	83.54	82.23	82.89	0.89	0.83	0.66
1.00E-02	84.37	84.75	84.56	0.90	0.85	0.69	83.54	82.29	82.92	0.88	0.83	0.66
1.00E-01	84.52	83.91	84.22	0.90	0.84	0.68	84.13	81.76	82.94	0.88	0.83	0.66
1.00E+00	84.51	83.22	83.87	0.89	0.84	0.68	83.76	81.40	82.58	0.88	0.83	0.65
1.00E+01	83.05	83.29	83.17	0.89	0.83	0.66	82.68	82.07	82.38	0.88	0.82	0.65
1.00E+02	75.40	75.04	75.22	0.88	0.75	0.50	76.44	76.34	76.39	0.87	0.76	0.53
1.00E+03	77.71	77.34	77.53	0.89	0.78	0.55	77.76	77.64	77.70	0.88	0.78	0.55

Mouse dataset

E-value	Training dataset						Independent dataset					
	Sens	Spec	ACC	AUC	F1	MCC	Sens	Spec	ACC	AUC	F1	MCC
1.00E-06	80.40	81.31	80.85	0.89	0.81	0.62	76.75	77.82	77.29	0.85	0.77	0.55
1.00E-05	80.41	81.04	80.73	0.89	0.81	0.62	76.38	77.44	76.91	0.85	0.77	0.54
1.00E-04	81.82	79.38	80.60	0.88	0.81	0.61	77.95	76.19	77.07	0.84	0.77	0.54
1.00E-03	80.48	81.15	80.81	0.88	0.81	0.62	75.50	78.20	76.85	0.84	0.77	0.54
1.00E-02	79.80	80.85	80.33	0.88	0.80	0.61	76.57	78.51	77.54	0.84	0.77	0.55
1.00E-01	81.43	80.04	80.73	0.88	0.81	0.62	76.88	77.19	77.04	0.84	0.77	0.54
1.00E+00	80.74	80.21	80.48	0.88	0.81	0.61	76.94	77.26	77.10	0.84	0.77	0.54
1.00E+01	80.77	79.85	80.31	0.87	0.80	0.61	76.94	77.51	77.22	0.85	0.77	0.54
1.00E+02	75.92	76.47	76.19	0.87	0.76	0.52	81.08	74.37	77.73	0.77	0.78	0.56
1.00E+03	75.89	75.31	75.60	0.81	0.76	0.51	73.00	80.70	76.85	0.85	0.76	0.54

Table 2. Performance of hybrid model on different E-values, build by integrating BLAST and machine learning approach on training and independent datasets for human and mouse host. *Sens: Sensitivity; Spec: Specificity; Acc: Accuracy; AUROC: Area Under the Receiver Operating Characteristics curve; MCC: Matthews Correlation Coefficient;

	Training dataset						Independent dataset					
Human dataset												
Feature type	Sens	Spec	ACC	AUROC	F1	MCC	Sens	Spec	ACC	AUROC	F1	MCC
N8	63.52	64.30	63.91	0.70	0.64	0.28	62.91	64.10	63.51	0.70	0.63	0.27
C8	63.84	62.13	62.98	0.69	0.63	0.26	63.85	61.22	62.53	0.68	0.63	0.25
N8C8	71.37	69.60	70.49	0.78	0.71	0.41	72.17	69.32	70.74	0.78	0.71	0.42
NC20	77.73	78.20	77.96	0.86	0.78	0.56	78.01	78.07	78.04	0.86	0.78	0.56
Mouse dataset												
Feature type	Sens	Spec	ACC	AUROC	F1	MCC	Sens	Spec	ACC	AUROC	F1	MCC
N8	62.85	60.87	61.86	0.68	0.62	0.24	61.40	62.03	61.72	0.68	0.62	0.23
C8	63.10	62.30	62.70	0.69	0.63	0.25	64.04	61.40	62.72	0.68	0.63	0.25
N8C8	64.30	65.67	64.98	0.71	0.65	0.30	64.04	65.29	64.66	0.71	0.64	0.29
NC20	69.88	68.95	69.41	0.76	0.70	0.39	68.11	68.92	68.52	0.76	0.68	0.37

Table 3. Performance of extra tree (ET) classifier-based models build using binary profile-based features on training and independent dataset of human and mouse host. *Sens: Sensitivity; Spec: Specificity; Acc: Accuracy; AUROC: Area Under the Receiver Operating Characteristics curve; MCC: Matthews Correlation Coefficient;

developed, we have conducted a comprehensive comparison of the various available methods (See Supplementary Table S8). However, most of these methods fails to consider host-specific cytokine-inducing peptides. Since, most of these tools rely on amino acid composition or specific residues at the N-terminus or C-terminus. It is important to develop host-specific tools for cytokines prediction, because different host have variable compositions of amino-acid residues. Our analysis revealed that distinct amino acid preferences and compositions between human and mouse host models. For instance, lysine (K) is highly conserved in human IFN-γ inducing peptides, while alanine (A) is specific to mouse IFN-γ inducing peptides. Whereas in the previous tool (IFNepitope)[30] this import feature is missing. We have compared the performance of IFNepitiope and IFNepitope2 tools on the same independent dataset to provide unbiased evaluation. AUROC plots show that IFNepitope2 outperforms the existing method on both human and mouse hosts (See Fig. 3). Additionally, to evaluate the significance of host-specific models, we conducted a dataset switch, in which the performance of human model is assessed on mouse independent dataset and vice versa. Our analysis clearly demonstrates a decline in the performance of the models as the datasets exchanged (See Supplementary Table S9). Therefore, a host-specific model is essential for accurately predicting or designing IFN-γ inducing peptides. The performance measures such as sensitivity,
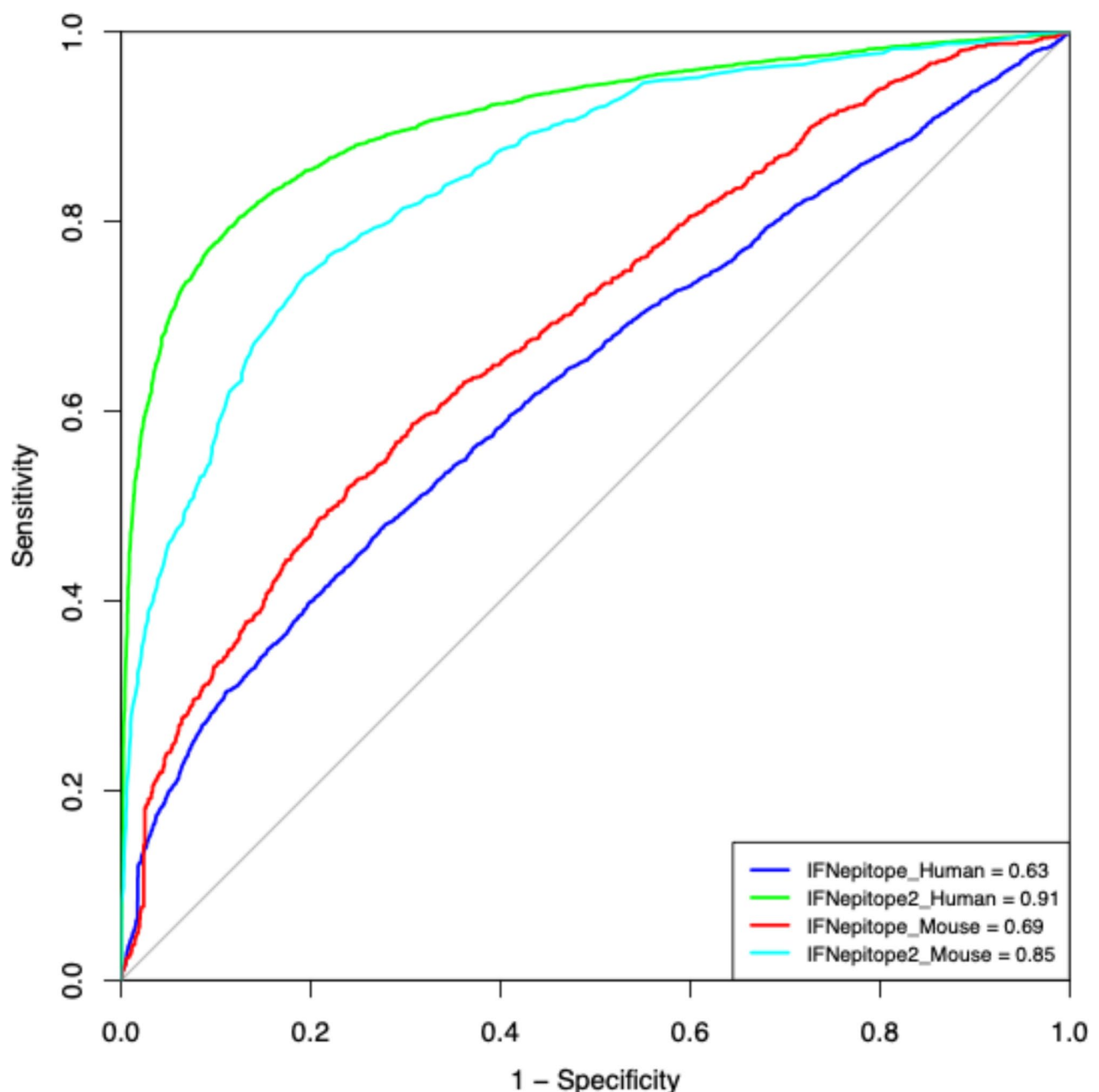

Fig. 3. Comparison of performance in terms of AUROC between IFNepitope and IFNepitope2 on the independent dataset for human and mouse dataset.

specificity, AUROC, accuracy, F1, and MCC were computed to compare the performance of both the methods (See Supplementary Table S9).

Discussion and conclusion

Traditional vaccinations are made up of whole pathogens that have been destroyed or weekend so they can no longer spread illness. Whole-pathogen vaccines can stimulate protective immune responses against pathogens and infections[56,57]. However, it is not necessary the entire protein is effective in targeting the pathogen or eliciting the desired immune response. Hence, subunit/peptide-based vaccines are developed in which a small portion of pathogen protein is utilized in order to generate the vaccine candidates against[24,58–63]. While designing a subunit-vaccines it is necessary to check the peptide/antigens have the capacity to effectively stimulate the immune system and generate specific type of cytokines to send signal to the immune system to perform inflammatory responses against the pathogens and cancerous cells. IFN-γ is one of most important cytokines which generate anti-tumor, anti-viral and proinflammatory responses against pathogens[64–66]. Hence, it is necessary to check the IFN-inducing potential subunit vaccine candidates. In order to help the researchers and experimental biologist we have developed a highly accurate tool for the prediction and designing of IFN-γ inducing peptides.

In the current study, we have used experimentally validated IFN-γ inducing and non-inducing peptides from IEDB resource. We have developed host-specific prediction models using human and mouse datasets. At first, we performed composition and positional preference analysis using human and mouse datasets. We observed that the average composition of 'Lysine', 'Proline' and 'Asparagine' amino acid is very high for human IFN-γ inducing peptides. While the composition of 'Alanine' and 'Proline' is higher in the case of mouse IFN-γ inducing peptides. Similarly, TSL plot reveals that 'Lysine' residue is highly preferred at most of the position in case of human and 'Alanine' is preferred in case of mouse dataset. We generated features binary-profile and composition-based features using Pfeature standalone package. We have developed various machine learning classifiers using number of features for both human and mouse hosts. Our hybrid model is able to achieve a maximum AUROC of (0.92 and 0.90) for human host and AUROC of (0.89 and 0.85) for mouse host on training and independent datasets, respectively. We have used the best models for both the host to generate the web-server (IFNepitope2) and standalone package. Moreover, we also compared the performance of our models with the existing tool. Our tool 'IFNepitope2' outperform the existing tool on independent validation for both human and mouse hosts. We anticipate this newly developed method is validated on huge amount of experimentally validated dataset and highly accurate then the existing method.

Data availability

The datasets used in this study is available at https://webs.iiitd.edu.in/raghava/ifnepitope2/download.php.

References

1. Castro, F., Cardoso, A. P., Goncalves, R. M., Serre, K. & Oliveira, M. J. Interferon-gamma at the crossroads of tumor immune surveillance or evasion. *Front. Immunol.* **9**, 847. https://doi.org/10.3389/fimmu.2018.00847 (2018).
2. Zaidi, M. R. & Merlino, G. The two faces of interferon-gamma in cancer. *Clin. Cancer Res.* **17**, 6118–6124. https://doi.org/10.1158/1078-0432.ccr-11-0482 (2011).
3. Tau, G. & Rothman, P. Biologic functions of the IFN-gamma receptors. *Allergy.* **54**, 1233–1251. https://doi.org/10.1034/j.1398-9995.1999.00099.x (1999).
4. Pestka, S., Krause, C. D. & Walter, M. R. Interferons, interferon-like cytokines, and their receptors. *Immunol. Rev.* **202**, 8–32. https://doi.org/10.1111/j.0105-2896.2004.00204.x (2004).
5. Schroder, K., Hertzog, P. J., Ravasi, T. & Hume, D. A. Interferon-gamma: an overview of signals, mechanisms and functions. *J. Leukoc. Biol.* **75**, 163–189. https://doi.org/10.1189/jlb.0603252 (2004).
6. Muhl, H. & Pfeilschifter, J. Anti-inflammatory properties of pro-inflammatory interferon-gamma. *Int. Immunopharmacol.* **3**, 1247–1255. https://doi.org/10.1016/s1567-5769(03)00131-0 (2003).
7. Miller, N. M., Wang, J., Tan, Y. & Dittel, B. N. Anti-inflammatory mechanisms of IFN-gamma studied in experimental autoimmune encephalomyelitis reveal neutrophils as a potential target in multiple sclerosis. *Front. Neurosci.* **9**, 287. https://doi.org/10.3389/fnins.2015.00287 (2015).
8. Yu, J. et al. Pro- and antiinflammatory cytokine signaling: reciprocal antagonism regulates interferon-gamma production by human natural killer cells. *Immunity.* **24**, 575–590. https://doi.org/10.1016/j.immuni.2006.03.016 (2006).
9. Gocher, A. M., Workman, C. J. & Vignali, D. A. A. Interferon-gamma: teammate or opponent in the tumour microenvironment? *Nat. Rev. Immunol.* **22**, 158–172. https://doi.org/10.1038/s41577-021-00566-3 (2022).
10. Bergman, S. J., Ferguson, M. C. & Santanello, C. Interferons as therapeutic agents for infectious diseases. *Infect. Dis. Clin. North. Am.* **25**, 819–834. https://doi.org/10.1016/j.idc.2011.07.008 (2011).
11. Strannegard, O. Interferons and their therapeutic applications. *EJIFCC.* **11**, 52–58 (1999).
12. Akinleye, A. & Rasool, Z. Immune checkpoint inhibitors of PD-L1 as cancer therapeutics. *J. Hematol. Oncol.* **12**, 92. https://doi.org/10.1186/s13045-019-0779-5 (2019).
13. Qureshi, S. et al. Immune checkpoint inhibitors in triple negative breast cancer: the search for the optimal biomarker. *Biomark. Insights.* **17**, 11772719221078774. https://doi.org/10.1177/11772719221078774 (2022).
14. Zhang, L. et al. Peptide-based materials for cancer immunotherapy. *Theranostics.* **9**, 7807–7825. https://doi.org/10.7150/thno.37194 (2019).
15. Kuai, R., Ochyl, L. J., Bahjat, K. S., Schwendeman, A. & Moon, J. J. Designer vaccine nanodiscs for personalized cancer immunotherapy. *Nat. Mater.* **16**, 489–496. https://doi.org/10.1038/nmat4822 (2017).
16. Gjertsen, M. K. et al. Vaccination with mutant ras peptides and induction of T-cell responsiveness in pancreatic carcinoma patients carrying the corresponding RAS mutation. *Lancet.* **346**, 1399–1400. https://doi.org/10.1016/s0140-6736(95)92408-6 (1995).
17. Qin, H., Ding, Y., Mujeeb, A., Zhao, Y. & Nie, G. Tumor microenvironment targeting and responsive peptide-based nanoformulations for improved tumor therapy. *Mol. Pharmacol.* **92**, 219–231. https://doi.org/10.1124/mol.116.108084 (2017).

18. Porkka, K., Laakkonen, P., Hoffman, J. A., Bernasconi, M. & Ruoslahti, E. A fragment of the HMGN2 protein homes to the nuclei of tumor cells and tumor endothelial cells in vivo. *Proc. Natl. Acad. Sci. U S A.* **99**, 7444–7449. https://doi.org/10.1073/pnas.062189599 (2002).

19. Mahaki, H. et al. The role of peptide-based tumor vaccines on cytokines of adaptive immunity: a review. *Int. J. Pept. Res. Ther.* **27**, 2527–2542. https://doi.org/10.1007/s10989-021-10270-4 (2021).

20. Miller, C. H., Maher, S. G. & Young, H. A. Clinical use of interferon-gamma. *Ann. N Y Acad. Sci.* **1182**, 69–79. https://doi.org/10.1111/j.1749-6632.2009.05069.x (2009).

21. Bhat, P., Leggatt, G., Waterhouse, N. & Frazer, I. H. Interferon-gamma derived from cytotoxic lymphocytes directly enhances their motility and cytotoxicity. *Cell. Death Dis.* **8**, e2836. https://doi.org/10.1038/cddis.2017.67 (2017).

22. Purcell, A. W., McCluskey, J. & Rossjohn, J. More than one reason to rethink the use of peptides in vaccine design. *Nat. Rev. Drug Discov.* **6**, 404–414. https://doi.org/10.1038/nrd2224 (2007).

23. Li, W., Joshi, M. D., Singhania, S., Ramsey, K. H. & Murthy, A. K. Peptide vaccine: progress and challenges. *Vaccines (Basel).* **2**, 515–536. https://doi.org/10.3390/vaccines2030515 (2014).

24. Gong, W. et al. Peptide-based vaccines for tuberculosis. *Front. Immunol.* **13**, 830497. https://doi.org/10.3389/fimmu.2022.830497 (2022).

25. Dhanda, S. K. et al. Novel in silico tools for designing peptide-based subunit vaccines and immunotherapeutics. *Brief. Bioinform.* **18**, 467–478. https://doi.org/10.1093/bib/bbw025 (2017).

26. Black, M., Trent, A., Tirrell, M. & Olive, C. Advances in the design and delivery of peptide subunit vaccines with a focus on toll-like receptor agonists. *Expert Rev. Vaccines.* **9**, 157–173. https://doi.org/10.1586/erv.09.160 (2010).

27. Oli, A. N. et al. Immunoinformatics and vaccine development: an overview. *Immunotargets Ther.* **9**, 13–30. https://doi.org/10.2147/itt.s241064 (2020).

28. Tahir Ul Qamar, M. et al. Multiepitope-based subunit vaccine design and evaluation against respiratory syncytial virus using reverse vaccinology approach. *Vaccines (Basel).* **8**(2), 288. https://doi.org/10.3390/vaccines8020288 (2020).

29. Umar, A. et al. Development of a candidate multi-epitope subunit vaccine against Klebsiella aerogenes: subtractive proteomics and immuno-informatics approach. *Vaccines (Basel).* **9**(11), 1373. https://doi.org/10.3390/vaccines9111373 (2021).

30. Dhanda, S. K., Vir, P. & Raghava, G. P. Designing of interferon-gamma inducing MHC class-II binders. *Biol. Direct.* **8**, 30. https://doi.org/10.1186/1745-6150-8-30 (2013).

31. Patiyal, S. et al. TNFepitope: A webserver for the prediction of TNF-α inducing epitopes. *Comput. Biol. Med.* **160**, 106929. https://doi.org/10.1016/j.compbiomed.2023.106929 (2023).

32. Kumar, R., Lathwal, A., Kaur, D. & Raghava, G. P. S. In silico model for predicting IL-2 inducing peptides in human. *bioRxiv*. https://doi.org/10.1101/2021.06.20.449146 (2021).

33. Nagpal, G. et al. Computer-aided designing of immunosuppressive peptides based on IL-10 inducing potential. *Sci. Rep.* **7**, 42851. https://doi.org/10.1038/srep42851 (2017).

34. Dhall, A., Patiyal, S., Sharma, N., Usmani, S. S. & Raghava, G. P. S. Computer-aided prediction and design of IL-6 inducing peptides: IL-6 plays a crucial role in COVID-19. *Brief Bioinform.* **22**, 936–945. https://doi.org/10.1093/bib/bbaa259 (2021)

35. Jain, S., Dhall, A., Patiyal, S. & Raghava, G. P. S. IL13Pred: A method for predicting immunoregulatory cytokine IL-13 inducing peptides. *Comput. Biol. Med.* **143**, 105297. https://doi.org/10.1016/j.compbiomed.2022.105297 (2022).

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

37. Charoenkwan, P. et al. StackIL6: a stacking ensemble model for improving the prediction of IL-6 inducing peptides. *Brief. Bioinform.* **22**(6), bbab172. https://doi.org/10.1093/bib/bbab172 (2021).

38. Singh, O., Hsu, W. L. & Su, E. C. ILeukin10Pred: a computational approach for predicting IL-10-inducing immunosuppressive peptides using combinations of amino acid global features. *Biology.* **11**(1), 5. https://doi.org/10.3390/biology11010005 (2021).

39. Naorem, L. D., Sharma, N. & Raghava, G. P. S. A web server for predicting and scanning of IL-5 inducing peptides using alignment-free and alignment-based method. *Comput. Biol. Med.* **158**, 106864. https://doi.org/10.1016/j.compbiomed.2023.106864 (2023).

40. Wang, R. et al. MVIL6: Accurate identification of IL-6-induced peptides using multi-view feature learning. *Int. J. Biol. Macromol.* **246**, 125412. https://doi.org/10.1016/j.ijbiomac.2023.125412 (2023).

41. Hassan, M. T., Tayara, H. & Chong, K. T. Meta-IL4: An ensemble learning approach for IL-4-inducing peptide prediction. *Methods.* **217**, 4956. https://doi.org/10.1016/j.ymeth.2023.07.002 (2023).

42. Vita, R. et al. The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res.* **47**, D339–D343. https://doi.org/10.1093/nar/gky1006 (2019).

43. Dhall, A., Patiyal, S. & Raghava, G. P. S. HLAncPred: a method for predicting promiscuous non-classical HLA binding sites. *Brief. Bioinform.* **23**(5), bbac192. https://doi.org/10.1093/bib/bbac192 (2022).

44. Sharma, N. et al. AlgPred 2.0: an improved method for predicting allergenic proteins and mapping of IgE epitopes. *Brief. Bioinform.* **22**(4), bbaa294. https://doi.org/10.1093/bib/bbaa294 (2021).

45. Patiyal, S., Dhall, A. & Raghava, G. P. S. A deep learning-based method for the prediction of DNA interacting residues in a protein. *Brief. Bioinform.* **23**(5), bbac322. https://doi.org/10.1093/bib/bbac322 (2022).

46. Patiyal, S. et al. NAGbinder: An approach for identifying N-acetylglucosamine interacting residues of a protein from its primary sequence. *Protein Sci.* **29**, 201–210. https://doi.org/10.1002/pro.3761 (2020).

47. Dhall, A., Patiyal, S., Sharma, N., Devi, N. L. & Raghava, G. P. S. Computer-aided prediction of inhibitors against STAT3 for managing COVID-19 associated cytokine storm. *Comput. Biol. Med.* **137**, 104780. https://doi.org/10.1016/j.compbiomed.2021.104780 (2021).

48. Patiyal, S., Dhall, A., Bajaj, K., Sahu, H. & Raghava, G. P. S. Prediction of RNA-interacting residues in a protein using CNN and evolutionary profile. *Brief. Bioinform.* **24**, bbac538. https://doi.org/10.1093/bib/bbac538 (2023).

49. Pande, A. et al. Pfeature: a tool for computing wide range of protein features and building prediction models. *J. Comput. Biol.* **30**(2), 204–222. https://doi.org/10.1089/cmb.2022.0241 (2023).

50. Vacic, V., Iakoucheva, L. M. & Radivojac, P. Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. *Bioinformatics.* **22**, 1536–1537. https://doi.org/10.1093/bioinformatics/btl151 (2006).

51. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410. https://doi.org/10.1016/s0022-2836(05)80360-2 (1990).

52. Dhall, A. Computing skin cutaneous melanoma outcome from the HLA-alleles and clinical characteristics. *Front. Genet.* **11**, 221. https://doi.org/10.3389/fgene.2020.00221 (2020).

53. Kumar, N. et al. DMPPred: a tool for identification of antigenic regions responsible for inducing type 1 diabetes mellitus. *Brief. Bioinform.* **24**, bbac525. https://doi.org/10.1093/bib/bbac525 (2023).

54. Kumar, V., Patiyal, S., Dhall, A., Sharma, N. & Raghava, G. P. S. B3Pred: a random-forest-based method for predicting and designing blood-brain barrier penetrating peptides. *Pharmaceutics.* **13**, 1237. https://doi.org/10.3390/pharmaceutics13081237 (2021).

55. Vens, C., Rosso, M. N. & Danchin, E. G. Identifying discriminative classification-based motifs in biological sequences. *Bioinformatics.* **27**(9), 1231–1238. https://doi.org/10.1016/j.ymeth.2023.07.002 (2011).

56. Pulendran, B. & Ahmed, R. Immunological mechanisms of vaccination. *Nat. Immunol.* **12**, 12:509–517. https://doi.org/10.1038/ni.2039 (2011).

57. Iwasaki, A. & Omer, S. B. Why and how vaccines work. *Cell.* **183**, 290–295. https://doi.org/10.1016/j.cell.2020.09.040 (2020).

58. Vassilakopoulou, V., Karachaliou, C. E., Evangelou, A., Zikos, C. & Livaniou, E. Peptide-based vaccines for neurodegenerative diseases: recent endeavors and future perspectives. *Vaccines (Basel).* **9**, 1278. https://doi.org/10.3390/vaccines9111278 (2021).

59. Liu, W. et al. Peptide-based therapeutic cancer vaccine: Current trends in clinical application. *Cell. Prolif.* **54**, e13025. https://doi.org/10.1111/cpr.13025 (2021).

60. Abd-Aziz, N. & Poh, C. L. Development of peptide-based vaccines for cancer. *J. Oncol.* **2022**, 9749363. https://doi.org/10.1155/2022/9749363 (2022).

61. Zhang, J. et al. Peptide-based nanovaccines in the treatment of cervical cancer: a review of recent advances. *Int. J. Nanomed.* **17**, 869–900. https://doi.org/10.2147/ijn.s269986 (2022).

62. Heitmann, J. S. et al. A COVID-19 peptide vaccine for the induction of SARS-CoV-2 T cell immunity. *Nature.* **601**, 617–622. https://doi.org/10.1038/s41586-021-04232-5 (2022).

63. Yang, H. et al. Developing an effective peptide-based vaccine for COVID-19: preliminary studies in mice models. *Viruses.* **14**, 449. https://doi.org/10.3390/v14030449 (2022).

64. Ivashkiv, L. B. IFNgamma: signalling, epigenetics and roles in immunity, metabolism, disease and cancer immunotherapy. *Nat. Rev. Immunol.* **18**, 545–558. https://doi.org/10.1038/s41577-018-0029-z (2018).

65. Kak, G., Raza, M. & Tiwari, B. K. Interferon-gamma (IFN-gamma): Exploring its implications in infectious diseases. *Biomol. Concepts.* **9**, 64–79. https://doi.org/10.1515/bmc-2018-0007 (2018).

66. McNab, F., Mayer-Barber, K., Sher, A., Wack, A. & O'Garra, A. Type I interferons in infectious disease. *Nat. Rev. Immunol.* **15**, 87–103. https://doi.org/10.1038/nri3787 (2015).

Acknowledgements

Author contributions

AD and GPSR collected and processed the datasets. AD, SP and GPSR implemented the algorithms and developed the prediction models. AD, SP and GPSR analysed the results. AD and SP created the web server. AD, SP and GPSR penned the manuscript. GPSR conceived and coordinated the project. All authors have read and approved the final manuscript.

Funding

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-77957-8.

Correspondence and requests for materials should be addressed to G.P.S.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.