Briefings in Bioinformatics, 2022, **23(5)**, 1–12

https://doi.org/10.1093/bib/bbac174

Problem Solving Protocol

ToxinPred2: an improved method for predicting toxicity of proteins

Neelam Sharma, Leimarembi Devi Naorem, Shipra Jain and Gajendra P.S. Raghava

Corresponding author: Gajendra P.S. Raghava, Head of Department, Department of Computational Biology, Indraprastha Institute of Information Technology, Okhla Phase 3, New Delhi-110020, India. Tel.: +91-11-26907444; E-mail: raghava@iiitd.ac.in

Abstract

Proteins/peptides have shown to be promising therapeutic agents for a variety of diseases. However, toxicity is one of the obstacles in protein/peptide-based therapy. The current study describes a web-based tool, ToxinPred2, developed for predicting the toxicity of proteins. This is an update of ToxinPred developed mainly for predicting toxicity of peptides and small proteins. The method has been trained, tested and evaluated on three datasets curated from the recent release of the SwissProt. To provide unbiased evaluation, we performed internal validation on 80% of the data and external validation on the remaining 20% of data. We have implemented the following techniques for predicting protein toxicity; (i) Basic Local Alignment Search Tool-based similarity, (ii) Motif-EmeRging and with Classes-Identification-based motif search and (iii) Prediction models. Similarity and motif-based techniques achieved a high probability of correct prediction with poor sensitivity/coverage, whereas models based on machine-learning techniques achieved balance sensitivity and specificity with reasonably high accuracy. Finally, we developed a hybrid method that combined all three approaches and achieved a maximum area under receiver operating characteristic curve around 0.99 with Matthews correlation coefficient 0.91 on the validation dataset. In addition, we developed models on alternate and realistic datasets. The best machine learning models have been implemented in the web server named 'ToxinPred2', which is available at https://webs.iiitd.edu.in/raghava/toxinpred2/ and a standalone version at https://github.com/raghavagps/toxinpred2. This is a general method developed for predicting the toxicity of proteins regardless of their source of origin.

Keywords: toxins, toxicity, machine learning, prediction, BLAST, motifs, proteins

Introduction

Proteins and peptides are naturally occurring molecules that play various functions and processes in the body that are essential to sustain cellular mechanisms [1]. Their aberrant activity has been involved in various disease conditions, including cancer, neurodegenerative disorders and diabetes [2]. Thus, using them as therapeutic agents is regarded as a promising way to fight against a variety of diseases. In recent years, they have shown the potential to revolutionize medical therapy. They are preferred over small molecules and antibodies due to their high target specificity, tissue penetration, high biological activity and inexpensive [3]. Various peptides and proteins have been approved by the U.S. Food and Drug Administration authority for clinical use as therapeutics [4, 5]. However, there are certain prime concerns in the protein/peptide-based drug discovery and development, such as toxicity, immunogenicity and stability [6]. Due to this, the assessment of toxic properties of protein/peptide is of great necessity to take this forward as a drug target.

Toxins are substances that have the potential to cause deleterious effects on living organisms. They are naturally present in plants or can be produced by animals (snakes, spiders, cone snails) and different types of microbes (such as bacteria, fungi to enhance their pathogenicity) [7]. Toxins, either natural or synthetic, can cause adverse health effects whenever the individual is exposed to them. Toxins originated from certain bacterial species such as *Clostridium botulinum*, *Vibrio cholerae*, *Clostridium tetani* cause deadly diseases like botulism, cholera and tetanus, respectively [7]. A variety of toxins from certain animals can lead to several lethal effects, such as scorpion venom can overstimulate neuronal signalling leading to paralysis [8]. Furthermore, snake venom could be neurotoxic, causing neuromuscular

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

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

Shipra Jain is pursuing her PhD 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.

Received: January 5, 2022. **Revised:** March 31, 2022. **Accepted:** April 18, 2022

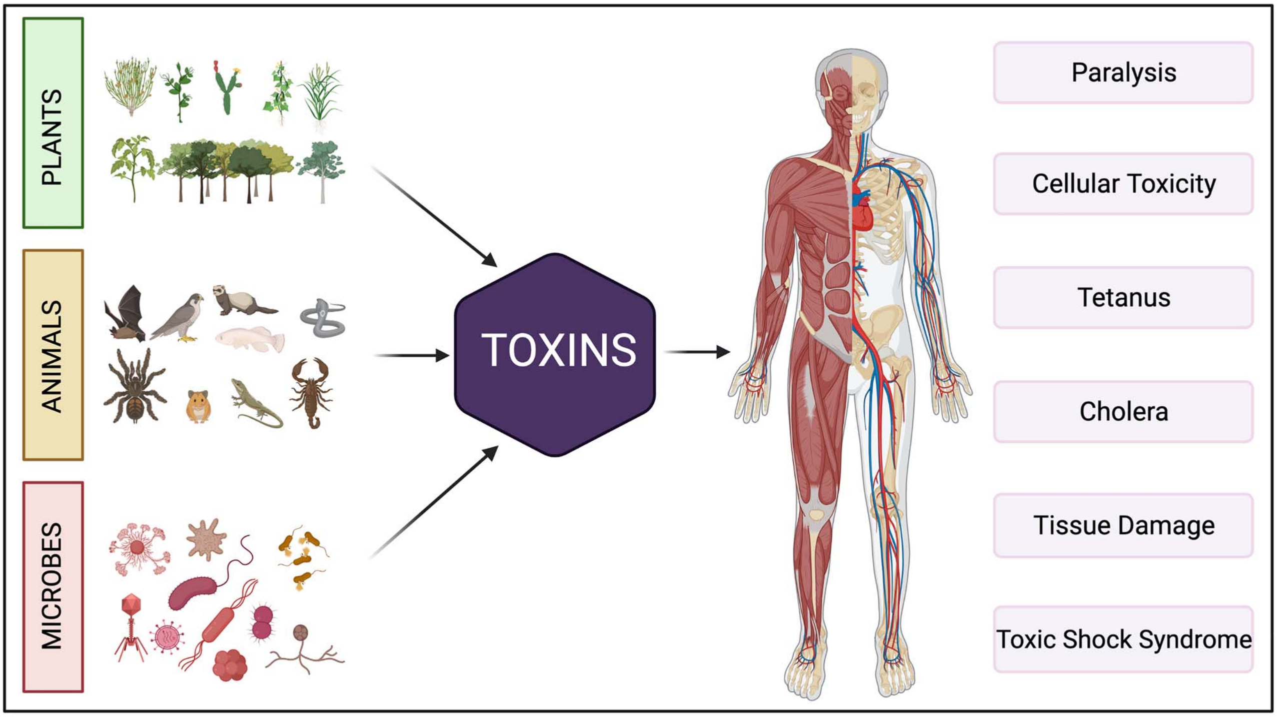

Figure 1. Various effects of toxins on human from different sources.

paralysis as well as haemotoxic damaging the circulatory system leading to acute tissue damage [7, 9, 10]. The various effects of toxins from different sources are depicted in Figure 1. Toxins are generally macromolecules like proteins, peptides and small molecules such as chemicals. The conventional experimental techniques are used to evaluate the toxicity of unknown proteins, peptides and chemical compounds. However, these approaches are laborious, cost-intensive and involve animal testing for *in vivo* assessment. All these impediments lead to an inclination towards the applicability of *in silico* techniques [11]. With the advent of highly accurate and cost-effective methods, the scientific community has adopted data-driven computational methods, such as machine learning techniques, to predict the toxicity of molecules [12].

Over the years, a plethora of research work has been published to study the toxicity of chemicals, namely DeepTox [13], ProTox-II [14] and eToxPred [15]. The limited attempts have been made for predicting the toxicity of proteins and peptides. The majority of available tools are extensively specialized for toxins of certain animal origins; for instance, in 2007, prediction methods named BTXpred [16] and NTXpred [17] were developed for the classification of bacterial toxins and neurotoxins, respectively. In 2009, ClanTox, developed by Naamati *et al.*, is a classifier of animal toxins from their primary protein sequences [18]. Another machine learning-based method, SpiderP, has been developed to predict the propeptide cleavage sites in spider toxins [19]. Similarly, ToxClassifier was developed by Gacesa *et al.*, in 2016 to identify venom toxins [20]. Deep learning-based methods such as TOXIFY [21] and ToxDL [22] were developed in 2019 and 2020, respectively. TOXIFY can be used to classify animal venom proteins from non-toxic proteins, whereas ToxDL can be used to assess the protein toxicity of animal origin.

In 2013, Gupta *et al.*, proposed a general method ToxinPred for predicting toxicity of peptides and proteins irrespective of their source. This method is heavily used by the research community to predict the toxicity of protein/peptides. It is a support vector machine (SVM)-based method that utilizes several features like amino acid composition (AAC), dipeptide composition and finding toxic motifs/ regions derived from the sequences [23]. NNTox is a machine learning method to detect the toxicity of protein based on several gene ontology annotations [24]. Recently, a deep learning-based tools such as ATSE [25] and ToxIBLT [26] were developed by Wei *et al.*, for the prediction of protein/peptide toxicity using structural, evolutionary and physicochemical properties of the sequences. In addition, a number of methods have been developed for predicting the specific type of toxicity, like haemotoxicity. In Table 1, we provide a comprehensive list of toxicity prediction methods.

In this study, we have made an attempt to develop a highly accurate method for predicting toxicity of large proteins, which will complement our previous method ToxinPred. Although, ToxinPred is highly accurate and used widely by the scientific community, but there are several constraints that necessitate improvement. ToxinPred was trained on 1805 toxic peptides where the maximum length was 35 amino acids. Thus, ToxinPred is suitable only for peptides or small peptides of length up to 50 amino acids but not suitable for large proteins. To address these limitations, we have proposed the

Table 1. List of computational tools developed for predicting wide range of toxicity of peptide, proteins and small molecules

Tools (year)	Description (link)	Reference
Protein/Peptide Toxicity Prediction Tools		
BTXpred (2007)	Classification of bacterial toxins (exotoxins and endotoxins) (https://webs.iiitd.edu.in/raghava/btxpred/)	[16]
NTXpred (2007)	Prediction of neurotoxins (https://webs.iiitd.edu.in/raghava/ntxpred/)	[17]
ClanTox (2009)	Classification of animal toxins from their primary protein sequences (http://www.clantox.cs.huji.ac.il)	[18]
SpiderP (2013)	Prediction of the propeptide cleavage sites in spider toxins (http://www.arachnoserver.org/spiderP.html)	[19]
ToxClassifier (2016)	Prediction of venom toxins from other proteins (http://bioserv7.bioinfo.pbf.hr/ToxClassifier/)	[20]
TOXIFY (2019)	Deep learning approach for the classification of animal venom proteins (https://www.github.com/tijeco/toxify)	[21]
NNTox (2019)	Detection of protein toxicity based on gene ontology annotations (http://www.github.com/kiharalab/NNTox)	[24]
ToxDL (2020)	Prediction of toxic proteins from animal species like snakes and spiders (http://www.csbio.sjtu.edu.cn/bioinf/ToxDL/)	[22]
ATSE (2021)	Prediction of peptide toxicity with their structural and evolutionary information (http://server.malab.cn/ATSE)	[25]
ToxIBLT (2022)	A deep learning approach for the prediction of peptide toxicity using information bottleneck and transfer learning (http://server.wei-group.net/ToxIBTL)	[26]
Other Toxicity Prediction Tools		
ToxiPred (2016)	Prediction of aqueous toxicity of small chemical molecules (https://webs.iiitd.edu.in/raghava/toxipred/)	[27]
HemoPI (2016)	Prediction of hemolytic or hemotoxic nature of peptides (https://webs.iiitd.edu.in/raghava/hemopi/)	[28]
DeepTox (2016)	Prediction of chemical compounds using deep learning (http://www.bioinf.jku.at/research/DeepTox/)	[13]
HemoPred (2017)	Predicting the hemolytic activity of peptides (http://codes.bio/hemopred/)	[29]
ToxiM (2017)	Prediction of toxicity of small molecules (http://metagenomics.iiserb.ac.in/ToxiM/)	[30]
CLC-Pred (2018)	Prediction of the cytotoxicity of a chemical compound (http://way2drug.com/Cell-line/)	[31]
ProTox-II (2018)	Prediction of toxicity of chemicals (http://tox.charite.de/protox_II)	[14]
eToxPred (2019)	To predict the toxicity of drug candidates (https://github.com/pulimeng/etoxpred)	[15]
HLPpred-Fuse (2020)	Prediction of hemolytic peptides and its activity (http://thegleelab.org/HLPpred-Fuse)	[32]

updated method named 'ToxinPred2' to classify the toxic and non-toxic protein sequences, which is trained and evaluated on large proteins/toxins. Models developed in this study have been trained and evaluated on the latest dataset consisting of 8233 toxic sequences. In addition, several features have been integrated into ToxinPred2, which enhance the performance of the model with high precision.

Materials and Methods

Creation and compilation of datasets

The dataset was retrieved from UniProt release 2021_03 (released on 2 June 2021) [33] using different keywords for obtaining toxic and non-toxic proteins. We extracted 9940 toxic proteins using the keyword 'toxin AND reviewed: yes'. All protein sequences comprising 'BJOUXZ', <35 amino acids and non-toxic sequences similar to toxic sequences were discarded. Ultimately, we obtained 8233 toxic sequences, which is referred to as a positive dataset. The compilation of experimentally validated or well-annotated toxic proteins is possible, whereas it is challenging to obtain non-toxic proteins. Therefore, we have extracted the negative dataset from Swiss-Prot [34] using keywords 'NOT toxin NOT allergen AND reviewed: yes' and obtained 554 145 proteins. In this study, we have considered proteins that are reviewed and manually curated. From these data, we have discarded the sequences with length <35 amino acids and with non-standard characters. Hence, we proceeded with 460 257 non-toxic sequences as a negative dataset. The creation of datasets is depicted in Figure 2.

After that, CD-HIT software [35] was applied to both datasets at 40% sequence identity. It leads to a reduced number of sequences for positive and negative datasets. After applying CD-HIT, the positive dataset is reduced to 1924 sequences from 8233, whereas the negative dataset is reduced to 88 263 sequences from 460 257. We have created three datasets based on the number of toxic and non-toxic protein sequences, as described below.

(a) Main Dataset: this dataset contains 8233 toxic (obtained after preprocessing of positive data) and 8233 non-toxic (randomly selected from 88 263 negative data obtained after CD-HIT) protein sequences.

(b) Alternate Dataset: this dataset contains 1924 toxic (obtained after applying CD-HIT on 8233 positive data) and 1924 non-toxic (randomly selected from 88 263 sequences obtained after CD-HIT) non-redundant protein sequences. In this dataset, no two proteins have >40% sequence similarity.

(c) Realistic Dataset (10 times Negative Dataset): this dataset consists of 1924 toxic and 19 240 non-toxic protein sequences. These toxic sequences are same as those used in the alternate dataset, where no two proteins have >40% similarity. The non-toxic protein sequences were randomly selected from non-redundant 88 263 non-toxic sequences obtained after applying CD-HIT.

Basic local alignment search tool for similarity search

Basic Local Alignment Search Tool (BLAST version-2.2.29+) is a program that is extensively used to annotate nucleotide and protein sequences [36]. In this study, we have utilized it for the identification of toxins based on

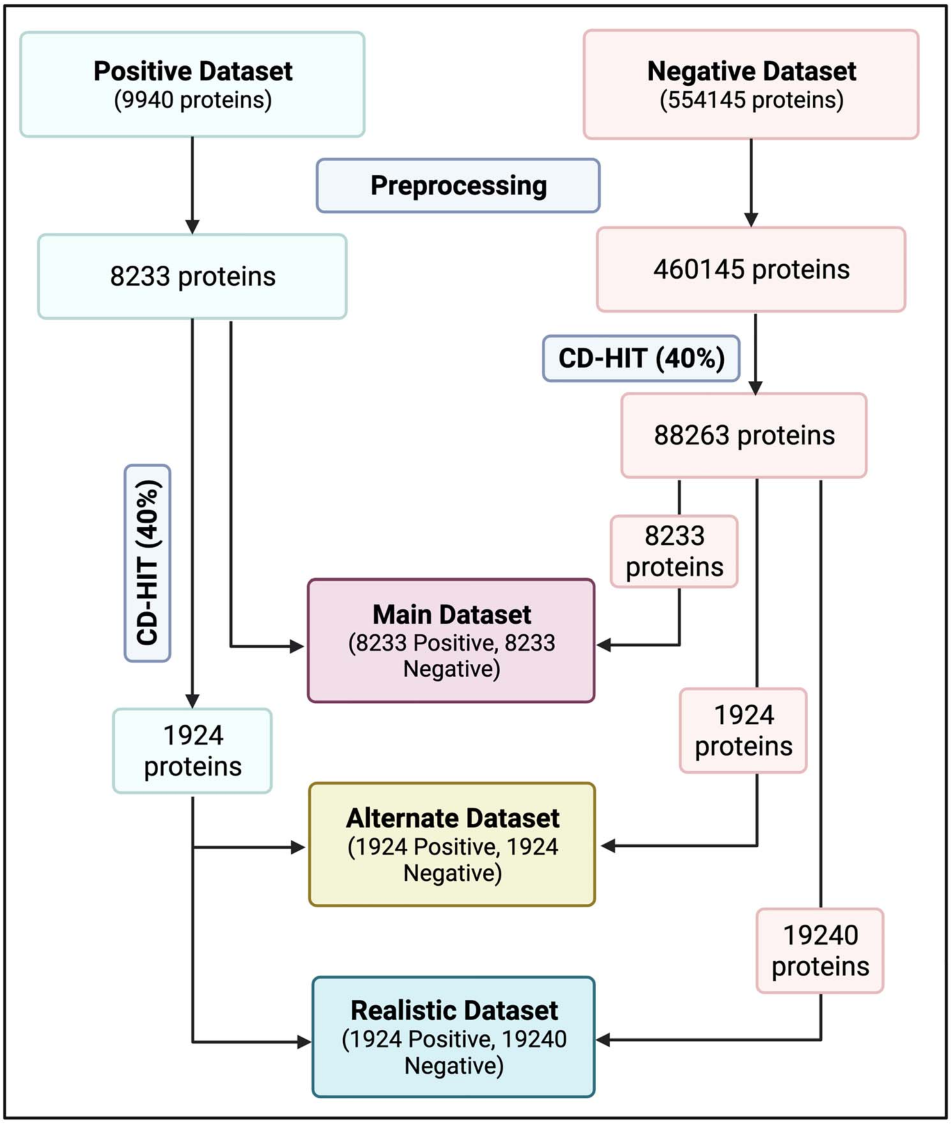

Figure 2. Overview of the creation the datasets.

the similarity of a protein sequence with toxic and non-toxic sequences. Using the protein–protein BLAST, the similarity-based search module was created in which the query sequences were searched against the database of toxins and non-toxins.

For this, two different approaches were used to identify toxins i.e. the top hit and ensemble of top five hits of BLAST at different E-value cutoffs. The sequences are specified as toxins and non-toxins based on the first hit of the query sequence against the database. Furthermore, a voting strategy was adopted to annotate a query protein, which is termed as an ensemble of top five hits. In this, there should be at least or more than five hits for consideration as a hit with respect to a query protein sequence. The query sequence was assigned as a toxin if the top five hits have maximum toxins. A similar approach is used to assign the query protein sequence as non-toxin. The performance of the method was assessed based on the various E-value cutoffs. This methodology has been used and well-annotated in different studies [37, 38].

Motif analysis

The toxic proteins were searched for the motifs by using Motif-EmeRging and with Classes-Identification (MERCI) tool, a program to locate motifs in any set of sequences [39]. Motif analysis provides information related to recurring patterns present in the toxic sequences. The software uses Perl script to locate motifs in the files using the default parameters.

Feature generation

There is a need to extract a set of relevant features for each protein/peptide sequence to develop any prediction model. Recent studies such as iLBE [40], ProIn-Fuse [41] and HLPpred-Fuse [32] have used different feature encoding techniques to represent the sequence into a feature vector. In this study, we have used a standalone tool, Pfeature, to generate a wide range of features such as composition and evolutionary information-based features of the protein sequences [42].

Composition-based features

Using a composition-based feature module of Pfeature [42], a vector of 9163 features was calculated against each sequence for all three datasets. The detailed information of each feature, along with the length of the vector, is tabulated in Supplementary Table S1.

Evolutionary information-based features

It has been shown in the past that evolutionary information provides more information about a protein than its primary sequence [38, 44]. To extract the evolutionary information for a given protein, position-specific scoring matrix (PSSM) profile was calculated using Position-Specific Iterated BLAST [43]. A 20 × 20 compositional matrix (PSSM-400) was created for each protein sequence from the PSSM profile of a protein with a size 20 × length of protein sequence [44]. For generating PSSM-400 from PSSM profile, the following steps are involved. Firstly, PSSM values are normalized in the range of 0–1. Secondly, composition of occurrences of each type of amino acid corresponding to each type of amino acid in protein sequence is computed. It means for each column we will have 20 values instead of one. Hence, we will have a vector of dimension 20 × 20 for PSSM matrix. To generate this PSSM-400 matrix, we used Pfeature software, which generates a normalized matrix PSSM-400 for each protein with vector dimension 20 × 20 [42].

Feature selection and ranking

Previous studies have shown that all the features are not important. Thus, it is a major challenge to select the appropriate features from a larger set of features. In this study, we have used the support vector classifier (SVC)-L1-based feature selection technique (Scikit-learn package) to select the significant features from the high-dimensional feature set. This method is based on the SVC with linear kernel, penalized with L1 regularization [45].

Using this method, we have listed important features for all three datasets from the pool of 9163 features. Out of that, 129 features were selected for the main dataset, 32 for the alternate dataset and 52 for the realistic dataset. Furthermore, the feature-selector tool was utilized for ranking the top key features. It uses a decision tree (DT)-based algorithm called the Light Gradient Boosting Machine for ranking the feature that is frequently used to split the data across all trees [46]. The obtained top-ranked features that were used to build the different machine learning prediction models in all three datasets are shown in Supplementary Table S2.

Machine learning (ML)-based classifiers

Several machine learning techniques have been used to discriminate toxic from non-toxic proteins. Random Forest (RF) [47], Logistic Regression (LR) [48], Gaussian Naive Bayes (GNB) [49], DT [50], k-nearest neighbours (KNNs) [51], XGBoost (XGB) [52] and SVC [53] were implemented to develop the classification models. These classifiers were optimized using various hyperparameters, and the best results were included. The complete workflow of ToxinPred2 is depicted in Figure 3.

Cross-validation and performance metrics

In this study, all the three datasets were divided into 80:20 ratio, respectively, where 80% constitute the training and 20% validation datasets. Here, 5-fold cross-validation (CV) was applied on 80% training data for evaluating the machine learning models [38, 54]. For internal validation, the training data is split into five-folds, where four-folds are used for training and the remaining fifth for testing purposes. The same procedure is iterated five times so that each of the 5-fold is used for testing at least once. The performance of several machine learning models was assessed using the standard evaluation metrics, including threshold dependent and independent parameters. The area under receiver operating characteristic curve (AUC) is a threshold-independent parameter, and sensitivity, specificity, accuracy and Matthews correlation coefficient (MCC) are threshold-dependent parameters. These metrics are well annotated in the previous studies [37, 55].

$$Sensitivity = \frac{TP}{TP + FN} \times 100 \tag{1}$$

$$Specificity = \frac{TN}{TN + FP} \times 100 \tag{2}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \times 100 \tag{3}$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)\,(TP + FN)\,(TN + FP)\,(TN + FN)}} \tag{4}$$

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

Hybrid approach

In this study, we have also implemented a hybrid approach to enhance the prediction of the model.

The hybrid approach is the weighted scoring method, in which the score is computed by integrating three different methods (i) similarity-based approach using BLAST, (ii) motif-based approach using MERCI and (iii) ML-based technique. First, the given protein sequence was classified using BLAST at the E-value of 10^{-6}. We assigned the weight of '+0.5' for the positive predictions (toxic proteins), '−0.5' for negative predictions (non-toxic proteins) and '0' for no hits. Second, the same protein sequence was classified using MERCI. We assigned the score of '+0.5' if the motifs were found and '0' if the motifs were not found. In the case of a hybrid approach, scores obtained from three methods (i.e. BLAST, MERCI and ML scores) were combined to compute the overall score. Based on the overall score at different threshold values, the protein sequence is categorized as toxic and non-toxic. This hybrid approach has been extensively employed in several studies [23, 37, 38, 54].

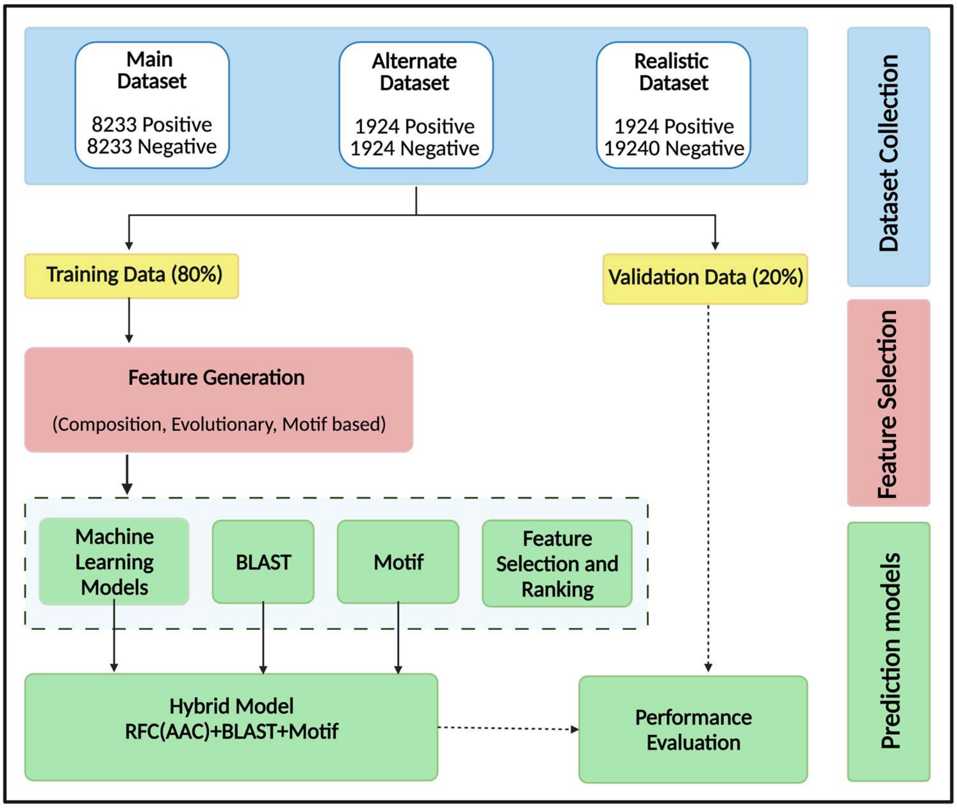

Figure 3. Flowchart depicting the overall architecture of ToxinPred2.

Results

Compositional analysis

In the study, AAC for both toxic and non-toxic proteins was computed. We found that the average AAC of amino acid residues such as cysteine, glycine, lysine and tryptophan is abundant in the toxic sequences, whereas alanine, glutamate, isoleucine, leucine and serine are higher in the non-toxic sequences. Also, we have compared the average AAC between the peptides and proteins of ToxinPred and ToxinPred2, respectively. It was observed that peptides of ToxinPred are exceptionally rich in cysteine and proline. In contrast, the proteins of ToxinPred2 are rich in lysine and valine. The comparison of average AAC between ToxinPred and ToxinPred2 is depicted in Figure 4.

Performance of BLAST models

BLAST is used extensively to annotate or assign the function to a query protein sequence based on similarity search. In this study, BLAST was used to classify the protein as toxin or non-toxin. A 5-fold CV is used for the evaluation of the performance of BLAST. First, the sequences in four-folds are used to construct the BLAST database and the sequences of the fifth fold are searched against the respective database. This process has been iterated five times. For evaluating the performance of BLAST on the validation dataset, BLAST database was created on the whole training set, and then each sequence in the validation dataset was searched against it. For this, we have used BLAST in two ways: top hit and ensemble of top five hits. The top hit BLAST is a standard method in which the proteins were assigned their class based on the first hit. However, it was noticed that this approach produces many false positives (data not shown). To reduce false predictions, we have used an ensemble of top five hits. It has been observed that after implying an ensemble of top five hits, the false prediction reduced significantly. For the main dataset, the number of correct hits (sensitivity) rises from 35% to 38.79% for the training dataset and from 36.44% to 40.23% for the validation dataset, with the E-value varying from 10^{-6} to 10^{-1}. This also leads to an increase in the number of wrong hits (error), as shown in Table 2. It indicates that the BLAST method alone is not efficient in discriminating between toxins and non-toxins. A similar approach has been employed on both alternate and realistic datasets. The results for the same are shown in Supplementary Tables S3 and S4.

Machine learning-based models

More than 10 types of composition-based features constituting a total of 9163 features were calculated for each protein using Pfeature. Apart from this, evolutionary-based features were also generated. These calculated features were used to develop different machine learning based models using the Scikit-learn package [45].

Composition-based features

The features, including AAC of toxins and non-toxins, were computed to develop several machine learning

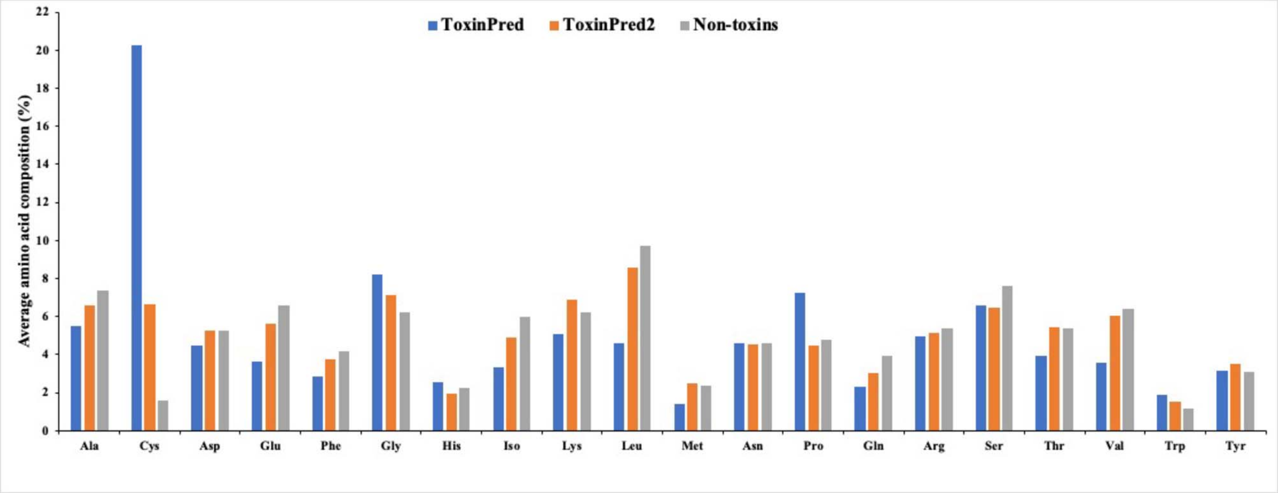

Figure 4. Comparison of average AAC between ToxinPred and ToxinPred2.

Table 2. The performance of BLAST-based approach on main dataset at different e-values

E-value	Training				Validation			
	Toxins		Non-toxins		Toxins		Non-toxins	
	Chits [a] (Sens)	Whits (error)	Chits (Spec)	Whits (error)	Chits (Sens)	Whits (error)	Chits (Spec)	Whits (error)
10^{-6}	4610 (35%)	68 (0.52%)	610 (4.63%)	137 (1.04%)	1201 (36.44%)	16 (0.49%)	182 (5.52%)	42 (1.27%)
10^{-5}	4681 (35.54%)	71 (0.54%)	662 (5.03%)	146 (1.11%)	1224 (37.14%)	17 (0.52%)	198 (6.01%)	45 (1.37%)
10^{-4}	4762 (36.16%)	77 (0.58%)	735 (5.58%)	157 (1.19%)	1251 (37.96%)	19 (0.58%)	212 (6.43%)	51 (1.55%)
10^{-3}	4869 (36.97%)	87 (0.66%)	812 (6.17%)	174 (1.32%)	1271 (38.56%)	24 (0.73%)	248 (7.52%)	55 (1.67%)
10^{-2}	4976 (37.78%)	102 (0.77%)	900 (6.83%)	192 (1.46%)	1293 (39.23%)	28 (0.85%)	271 (8.22%)	61 (1.85%)
10^{-1}	5109 (38.79%)	124 (0.94%)	1034 (7.85%)	214 (1.62%)	1326 (40.23%)	35 (1.06%)	319 (9.68%)	72 (2.18%)

[a] Chits: Correct hits; Whits: Wrong hits; Sens: Sensitivity; Spec: Specificity

models. For the main dataset, it was observed that RF-based models performed quite well when compared with other models and achieved a maximum AUC of 0.93 and 0.92 on training and validation datasets, respectively. For the alternate dataset, RF-based model attained the AUC of 0.76 and 0.75 on the training and validation dataset, respectively. In the case of a realistic dataset, it was found that the model based on RF obtained the AUC of 0.78 and 0.77 for the training and validation dataset, respectively. The performance of the main dataset for composition based features is shown in Table 3, whereas for other datasets, it is tabulated in Supplementary Tables S5 and S6.

Evolutionary information-based features

The PSSM profiles based on evolutionary information were also generated for protein sequences and used to develop ML-based models. We found that XGB achieved the AUC of 0.94 on training and 0.93 on validation for the main dataset. Furthermore, for an alternate dataset, RF-based model attained the AUC of 0.80 on training and 0.79 on the validation dataset. For the realistic dataset, XGB performed better and obtained the maximum AUC of 0.80 on training and 0.79 on validation. The performance of the PSSM-based model for all three datasets is shown in Supplementary Tables S7–S9.

Selected features

As mentioned above in the ‘Feature Selection and Ranking’ section, a total of 9163 features were assessed for all three datasets. These features were then reduced to 129 (main dataset), 32 (alternate dataset) and 52 (realistic dataset) using the SVC-L1 method. These reduced features were used to develop different classification models, where the RF-based model performed better for all three datasets. For the main dataset, it achieved a maximum AUC of 0.94 on training and 0.93 on the validation dataset. Likewise, the AUC of 0.78 on training and 0.77 on the validation dataset was obtained in the case of the alternate dataset. For a realistic dataset, the attained AUC for the training dataset is 0.80 and 0.79 for validation dataset. The performance of selected features developed on alternate and realistic datasets is demonstrated in Supplementary Tables S10–S12.

Top-ranked features

The feature-selector tool ranks the feature based on their normalized and cumulative importance. The main objective is to determine the best minimal feature set that can distinguish toxins and non-toxins with maximum accuracy and AUC. Thus, the different prediction models were developed based on top selected features (5, 10, 20, 30, ...) for all three datasets. The complete

Table 3. The performance of machine learning-based models developed using AAC on the main dataset

Main Dataset										
ML	Training					Validation				
	Sens	Spec	Acc	AUC	MCC	Sens	Spec	Acc	AUC	MCC
RF	**86.59 ± 0.85**	**86.15 ± 0.71**	**86.37 ± 0.18**	**0.93 ± 0.01**	**0.73 ± 0.01**	**86.47**	**84.10**	**85.29**	**0.92**	**0.71**
SVC	84.39 ± 0.64	84.63 ± 0.66	84.51 ± 0.27	0.92 ± 0.01	0.69 ± 0.01	83.62	82.16	82.89	0.91	0.66
XGB	83.55 ± 0.54	83.98 ± 0.29	83.77 ± 0.23	0.91 ± 0.01	0.68 ± 0.01	82.34	82.77	82.56	0.91	0.65
KNN	82.60 ± 0.2	81.72 ± 1.24	82.16 ± 0.58	0.91 ± 0.00	0.64 ± 0.01	82.40	82.16	82.28	0.90	0.65
DT	77.19 ± 0.69	77.90 ± 1.18	77.55 ± 0.53	0.85 ± 0.00	0.55 ± 0.01	75.73	80.52	78.13	0.85	0.56
LR	75.54 ± 0.78	75.11 ± 0.92	75.32 ± 0.47	0.84 ± 0.01	0.51 ± 0.01	75.61	73.30	74.45	0.84	0.49
GNB	75.22 ± 1.63	74.49 ± 0.9	74.85 ± 1.03	0.83 ± 0.01	0.50 ± 0.02	74.76	73.60	74.18	0.82	0.48

Sens = sensitivity; Spec = specificity; Acc = accuracy; AUC = area under receiver operating characteristic; MCC = Matthews correlation coefficient

Table 4. The performance of motif-based approach on main dataset when combined with machine learning-based models developed using AAC

Main Dataset										
	Training					Validation				
ML	Sens	Spec	Acc	AUC	MCC	Sens	Spec	Acc	AUC	MCC
RF	**84.60 ± 0.97**	**89.36 ± 0.56**	**86.98 ± 0.34**	**0.94 ± 0.01**	**0.74 ± 0.01**	**84.59**	**87.93**	**86.26**	**0.93**	**0.73**
SVC	85.24 ± 0.64	82.20 ± 0.65	83.72 ± 0.33	0.92 ± 0.00	0.68 ± 0.01	84.59	80.40	82.49	0.91	0.65
XGB	83.28 ± 0.97	84.53 ± 0.51	83.90 ± 0.49	0.91 ± 0.01	0.68 ± 0.01	81.80	83.50	82.65	0.91	0.65
KNN	81.72 ± 0.54	83.40 ± 0.6	82.56 ± 0.54	0.91 ± 0.00	0.65 ± 0.01	81.74	84.16	82.95	0.90	0.66
DT	75.46 ± 1.37	80.90 ± 1.03	78.18 ± 0.71	0.85 ± 0.01	0.56 ± 0.01	74.15	83.13	78.64	0.85	0.58
LR	74.02 ± 1.08	78.76 ± 0.51	76.389 ± 0.63	0.84 ± 0.01	0.53 ± 0.01	74.21	77.00	75.61	0.84	0.51
GNB	72.12 ± 1.68	79.56 ± 0.47	75.84 ± 0.94	0.83 ± 0.01	0.52 ± 0.02	72.76	79.37	76.06	0.82	0.52

Sens = sensitivity; Spec = specificity; Acc = accuracy; AUC = area under receiver operating characteristic; MCC = Matthews correlation coefficient

results for all the top-ranked features are provided in Supplementary Tables S13–S15.

Motif approach

MERCI software has been used to identify the motifs exclusively present in toxic proteins of the main dataset. The motifs such as 'GCYCG, MKTLL, TLLLTL and LLLTLV' are solely found in toxic proteins. Composition-based models (AAC) built using different ML techniques were integrated with this approach. For the main dataset, RF-based model attains the AUC of 0.94 on training and 0.93 on validation dataset, whereas for alternate dataset (GNB-based model) and realistic dataset (RF-based model) achieved the AUC of 0.72, 0.71 and 0.79, 0.77 on training and validation dataset, respectively. The performance of the combined approach (ML + MERCI) for the main dataset is shown in Table 4, whereas for other datasets, it is shown in Supplementary Tables S16 and S17.

BLAST search

To build an enhanced method, the similarity search approach BLAST and ML-based models were synergized. The BLAST search was initially implemented for a query sequence; if a BLAST hit was obtained, the query sequence was assigned as toxin and non-toxin based on the BLAST result. If no hit is obtained, then the composition-based model is utilized to predict the same sequence. For the main dataset, the performance of RF-based model increases from AUC 0.93 to 0.98 on training and 0.92 to 0.99 on validation dataset, as shown in Table 5. In the case of an alternate dataset, it is found that the AUC increases from 0.76 to 0.89 on training and 0.75 to 0.89 on validation using an RF-based model. For realistic dataset, LR-based model performed better and improved significantly with AUC of 0.75 to 0.94 and 0.74 to 0.94 for the training and validation dataset. The results for alternate and realistic datasets are shown in Supplementary Tables S18 and S19.

Hybrid approach

Ultimately, multiple approaches were integrated to overcome the limitations of individual methods. These approaches were developed to detect the toxins with better precision. In this, a composition-based model is combined with BLAST and MERCI-based approaches. Initially, proteins were classified using ensemble BLAST at an E-value of 10^{-6} led by the MERCI approach. An ML-based model then predicts the protein sequences that are not predicted by these two approaches. The hybrid method significantly enhanced the coverage and accuracy, which is not feasible by using all these methods individually. The performance of the hybrid method has improved by combining all these methods, as shown in Table 6. RF-based model performed the best for all three datasets on training and validation datasets. They

Table 5. The performance of BLAST-based approach on main dataset when combined with machine learning-based models using AAC

Main Dataset

ML	Training					Validation				
	Sens	Spec	Acc	AUC	MCC	Sens	Spec	Acc	AUC	MCC
RF	**93.85 ± 0.42**	**96.14 ± 0.39**	**95.12 ± 0.32**	**0.98 ± 0.01**	**0.9 ± 0.01**	**94.36**	**95.75**	**95.05**	**0.99**	**0.9**
KNN	93.53 ± 0.55	95.22 ± 0.57	94.37 ± 0.54	0.98 ± 0.00	0.89 ± 0.01	94.05	95.57	94.81	0.98	0.9
XGB	94.73 ± 0.46	93.35 ± 0.21	94.04 ± 0.31	0.98 ± 0.00	0.88 ± 0.01	95.02	93.14	94.08	0.98	0.88
LR	93.65 ± 0.44	92.44 ± 0.37	93.04 ± 0.36	0.98 ± 0.00	0.86 ± 0.01	94.72	92.6	93.66	0.98	0.87
GNB	94.08 ± 0.56	91.19 ± 0.83	92.63 ± 0.6	0.96 ± 0.00	0.85 ± 0.01	94.9	91.57	93.23	0.96	0.87
SVC	91.69 ± 0.46	91.47 ± 0.48	91.58 ± 0.17	0.98 ± 0.00	0.83 ± 0.00	91.14	90.78	90.96	0.98	0.82
DT	88.63 ± 0.76	91.77 ± 0.62	90.2 ± 0.54	0.97 ± 0.00	0.8 ± 0.01	88.53	90.59	89.56	0.97	0.79

Sens = sensitivity; Spec = specificity; Acc = accuracy; AUC = area under receiver operating characteristic; MCC = Matthews correlation coefficient

achieved an AUC of 0.98 and 0.99 (main dataset), AUC of 0.90 and 0.90 (alternate dataset) and AUC of 0.95 and 0.96 (realistic dataset) on training and validation dataset.

Design and implementation of a web server

A web server, ToxinPred2 (https://webs.iiitd.edu.in/raghava/toxinpred2/), has been developed for predicting toxic proteins. We have executed our two best performing models i.e. Model-1 (AAC-based RF approach) and Model-2 (hybrid approach). Both the models are trained on the main dataset for predicting toxins. The major modules such as (a) prediction, (b) motif scan, (c) BLAST search and (d) Download are integrated into the web server. The 'prediction module' permits the user to submit the single as well as multiple protein sequences in FASTA format. This module can efficiently classify toxic and non-toxic proteins. The 'motif scan module' uses MERCI software to identify the exclusively present motifs in the toxic protein sequences. It also maps or scans the motifs in the query protein sequence given by the user and distinguishes them as toxin and non-toxin. The 'BLAST search module' aids the user to carry out a similarity-based search using BLAST against toxins and non-toxins database. The web server is built with a responsive HTML template and browser compatibility for various operating systems. To facilitate the users to predict toxins at the genome scale, we have also developed a python-based standalone package of ToxinPred2, which can be accessed from 'Download' module of the web server.

Comparison with other methods

It is important to compare the performance of the proposed method with existing methods to justify the development of new method. We have shown the comparison of the performance of proposed method, ToxinPred2 and other existing methods as reported in the literature in Table 7. ToxinPred2 outperforms other existing methods, as shown in Table 7 under the heading 'Performance of existing methods reported in the literature'. To provide an unbiased comparison, we have computed the performance of ToxinPred2 on the validation dataset used in existing methods. As shown in Table 7, under the heading 'Performance of ToxinPred2 on validation dataset of existing methods', ToxinPred2 obtained AUC of 0.96, 0.99 and 0.99 for protein datasets used in ToxClassifer, TOXIFY and ToxDL, respectively. Our proposed method achieves AUC of 0.94 on peptide datasets used in ToxinPred and ATSE, which are lower than their original performance. It is because ToxinPred2 is developed/-trained for proteins not for peptides. We also attempted to evaluate the performance of existing methods on validation data (main dataset) of ToxinPred2. Since the dataset has larger protein sequences, the peptide toxicity prediction methods (ToxinPred, ToxIBLT and ATSE) cannot be implemented. Moreover, we were unable to predict the toxicity of proteins in the validation dataset of ToxinPred2 using ToxDL and ToxClassifier due to the limitations of their web services (one is non-functional, another allows a maximum of 10 sequences per submission for prediction). We used a standalone version of TOXIFY to predict toxicity of proteins in validation dataset of ToxinPred2. It has predicted the toxicity of only 2617 out of 3296 protein sequences, which have length up to 500 amino acids. As shown in Table 7, under the heading 'Performance of existing methods on validation dataset of ToxinPred2', it achieved AUC of 0.88, which is lower than the performance reported by ToxinPred2 on the same dataset. This comparison demonstrates the importance of the newly proposed method in the field of toxicity prediction.

Discussion

One of the major challenges in the field of protein/peptide based therapeutics is to identify toxic regions in a protein. There is a dire need to determine the toxic potential of newly synthesized proteins. Experimental techniques for determining toxicity proteins are costly and time-consuming. Thus, there is a need to develop computer-aided techniques for predicting toxicity of proteins/peptides with high precision. To facilitate the scientific community, our group developed a method, ToxinPred, for predicting and designing toxic peptides. It is heavily used by the scientific community in the field of therapeutic peptides. This tool has been developed mainly for peptides as models have been

Table 6. The performance of hybrid method on main dataset that combines BLAST, MERCI and machine learning-based approaches

ML	Training					Validation				
	Sens	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
				Main Dataset						
RF	**92.95 ± 0.40**	**97.65 ± 0.42**	**95.30 ± 0.36**	**0.98 ± 0.01**	**0.91 ± 0.01**	**93.69**	**97.39**	**95.54**	**0.99**	**0.91**
KNN	93.53 ± 0.55	95.22 ± 0.57	94.37 ± 0.54	0.98 ± 0.01	0.89 ± 0.03	94.05	95.57	94.81	0.98	0.9
LR	92.26 ± 0.44	95.99 ± 0.26	94.12 ± 0.29	0.98 ± 0.01	0.88 ± 0.05	93.33	95.81	94.57	0.98	0.89
XGB	94.73 ± 0.46	93.35 ± 0.21	94.04 ± 0.31	0.98 ± 0.01	0.88 ± 0.05	95.02	93.14	94.08	0.98	0.88
SVC	91.69 ± 0.46	91.63 ± 0.46	91.66 ± 0.18	0.98 ± 0.01	0.83 ± 0.05	91.14	90.96	91.05	0.98	0.82
GNB	94.08 ± 0.56	91.19 ± 0.83	92.63 ± 0.60	0.96 ± 0.01	0.85 ± 0.04	94.90	91.57	93.23	0.96	0.87
DT	88.64 ± 0.78	91.81 ± 0.62	90.23 ± 0.54	0.97 ± 0.01	0.80 ± 0.04	88.59	90.59	89.59	0.97	0.79
				Alternate Dataset						
RF	**79.94 ± 0.70**	**80.52 ± 3.87**	**80.23 ± 1.76**	**0.9 ± 0.01**	**0.60 ± 0.03**	**76.82**	**82.55**	**79.69**	**0.90**	**0.59**
XGB	78.90 ± 2.41	82.53 ± 3.77	80.71 ± 2.56	0.89 ± 0.02	0.61 ± 0.05	77.08	81.25	79.17	0.90	0.58
SVC	77.08 ± 1.82	79.55 ± 3.65	78.31 ± 2.64	0.89 ± 0.02	0.57 ± 0.05	76.30	83.59	79.95	0.89	0.60
KNN	73.96 ± 2.79	82.66 ± 1.27	78.31 ± 1.34	0.88 ± 0.01	0.57 ± 0.03	73.96	81.51	77.73	0.89	0.56
LR	76.10 ± 2.02	77.34 ± 2.44	76.72 ± 2.21	0.87 ± 0.02	0.53 ± 0.04	76.04	79.43	77.73	0.88	0.56
GNB	75.39 ± 1.69	79.16 ± 2.73	77.27 ± 1.94	0.85 ± 0.02	0.55 ± 0.04	76.04	80.73	78.39	0.86	0.57
DT	74.42 ± 1.68	78.31 ± 0.87	76.36 ± 0.82	0.84 ± 0.01	0.53 ± 0.02	73.70	78.12	75.91	0.85	0.52
				Realistic Dataset						
RF	**83.12 ± 2.35**	**90.16 ± 0.40**	**89.52 ± 0.22**	**0.95 ± 0.00**	**0.57 ± 0.01**	**81.77**	**91.22**	**90.36**	**0.96**	**0.58**
SVC	69.94 ± 3.69	97.10 ± 0.26	94.63 ± 0.43	0.95 ± 0.01	0.67 ± 0.03	71.88	97.17	94.87	0.95	0.69
XGB	80.45 ± 2.85	90.68 ± 0.39	89.75 ± 0.33	0.94 ± 0.01	0.56 ± 0.02	80.47	91.42	90.43	0.95	0.58
KNN	77.08 ± 2.48	90.92 ± 0.67	89.66 ± 0.52	0.94 ± 0.01	0.54 ± 0.01	78.65	90.77	89.67	0.95	0.55
LR	81.1 ± 0.27	86.73 ± 0.82	86.22 ± 0.73	0.94 ± 0.01	0.49 ± 0.01	82.03	87.37	86.89	0.94	0.51
GNB	78.44 ± 2.58	89.24 ± 0.49	88.26 ± 0.54	0.93 ± 0.01	0.52 ± 0.02	77.6	89.5	88.42	0.93	0.52
DT	83.25 ± 2.47	77.44 ± 0.43	77.97 ± 0.28	0.91 ± 0.01	0.39 ± 0.01	87.24	78.56	79.35	0.92	0.42

Sens = sensitivity; Spec = specificity; Acc = accuracy; AUC = area under receiver operating characteristic; MCC = Matthews correlation coefficient

Table 7. Comparison of proposed method ToxinPred2 with existing methods

Method	Type of dataset used	Sensitivity	Specificity	Accuracy	AUC	MCC
ToxinPred2	All types of toxins (Proteins)	93.69	97.39	95.54	0.99	0.91
			Performance of existing methods reported in the literature			
ToxinPred	All types of toxins (Peptides)	93.80	94.85	94.50	0.98	0.88
ToxClassifier	Animal venom toxins (Proteins)	96.70	99.80	99.70	NA	0.89
TOXIFY	Animal venom toxins (Proteins)	96.00	76.00	86.00	NA	0.74
ToxDL	Animal toxins (Proteins)	NA	NA	NA	0.98	0.79
ATSE	Toxic peptides	96.50	94.00	95.20	0.97	0.90
			Performance of ToxinPred2 on validation dataset of existing methods			
ToxinPred	Toxic peptides	97.73	45.73	63.12	0.94	0.44
ToxClassifier	Toxic proteins	97.15	77.54	87.38	0.96	0.76
TOXIFY	Toxic proteins	96.48	92.71	94.59	0.99	0.89
ToxDL	Toxic proteins	100	88.81	89.74	0.99	0.63
ATSE	Toxic peptides	96.65	58.21	76.03	0.94	0.58
			Performance of existing methods on validation dataset of ToxinPred2			
TOXIFY	Toxic proteins	68.94	97.94	81.85	0.88	0.68

trained on peptides having length up to 35 amino acids. To complement ToxinPred, we proposed a new method, ToxinPred2, for predicting toxicity of proteins. In this study, three datasets were created, namely main, alternate and realistic datasets curated from SwissProt. The main dataset consists of 8233 toxic and non-toxic proteins, alternate dataset contains 1924 non-redundant toxic and non-toxic proteins. Realistic dataset was generated to create realistic conditions in which negative data is multiple folds than positive data. Thus, 1924 toxic and 19 240 non-toxic proteins were used in realistic dataset.

Various features for the protein sequences were computed using Pfeature tool. The relevant set of features was further selected and ranked using SVC-L1 and Feature Selector tool, respectively. Our compositional analysis exhibited that cysteine, glycine, lysine and tryptophan are dominant in toxic proteins in comparison to non-toxic proteins. It is noteworthy that the composition-based features are among the top selected features. This suggests that these features can be used to distinguish between toxic and non-toxic proteins. Furthermore, we have implemented the BLAST, a widely used tool to annotate any query protein sequence. If the query protein sequence shows high similarity with a known protein function, then it designates the same function to the query protein. As shown in Table 2, BLAST has identified some toxins correctly with the probability of correct prediction of >40%, with a very low error rate. Thereby, it can be inferred that BLAST is generating a large number of no hits; hence, it fails when the unknown protein has no similarity with toxins and non-toxins. To overcome this limitation, hybrid models were developed using ML models (composition-based features like AAC), BLAST and MERCI. We have achieved the highest performance with balanced sensitivity and specificity and higher accuracy, as shown in Table 6.

In this study, we have provided a comprehensive platform where users can classify toxic and non-toxic proteins/peptides. We anticipate that our research will be beneficial to scientists working in the field of protein or peptide therapeutics. To facilitate the scientific community and to promote widespread usage of the proposed prediction method, we have provided a freely accessible web server and a standalone package of ToxinPred2. In the webserver, we have incorporated the best performing model for correctly predicting the toxins and non-toxins. However, one of the limitations of our method is that it can classify toxins and non-toxins regardless of their source of origin. We hope that the researchers would use our prediction method extensively for designing improved and accurate protein/peptide-based therapeutics against various diseases.

Key Points

- Upgraded version of ToxinPred has been developed for predicting toxicity of proteins.
- A wide range of features have been generated for developing prediction models.
- Cutting-edge techniques have been employed for feature engineering.
- Several machine learning techniques have been deployed to build prediction models.
- Similarity and motif-based techniques have been incorporated.

Data Availability Statement

All the datasets generated for this study are available at the 'ToxinPred2' webserver https://webs.iiitd.edu.in/raghava/toxinpred2/stand.html. The source code is hosted on GitHub and can be found at https://github.com/raghavagps/toxinpred2.

Conflict of Interest Statement

The authors declare that they have no conflict of interest.

Author Contributions

N.S. collected, compiled and processed the data sets. N.S. developed computer programs. N.S. implemented the algorithms and prediction models. N.S. created the back-end of the webserver and S.J. and N.L.D. front-end user interface. N.S. and N.L.D. analysed the results. N.S., N.L.D., S.J. and G.P.S.R. wrote the manuscript. G.P.S.R. conceived and coordinated the project and provided overall supervision of the project. All authors have read and approved the final manuscript.

Supplementary data

Supplementary data are available online at https://academic.oup.com/bib.

Acknowledgements

N.S. is thankful to the Department of Science and Technology (DST-INSPIRE) for providing Senior Research Fellowship. N.L.D. is thankful to DBT-RA program in Biotechnology and Life Sciences for providing Research Associate fellowship. N.S., N.L.D. and S.J. 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.

References

1. Deb PK, Al-Attraqchi O, Chandrasekaran B, *et al.* Protein/peptide drug delivery systems. Basic fundam. *Drug Deliv* 2019;651–84.
2. Keservani RK, Sharma AK, Jarouliya U. Protein and peptide in drug targeting and its therapeutic approach. *Ars Pharm* 2015;**56**: 165–77.
3. Bruno BJ, Miller GD, Lim CS. Basics and recent advances in peptide and protein drug delivery. *Ther Deliv* 2013;**4**:1443–67.
4. Fosgerau K, Hoffmann T. Peptide therapeutics: current status and future directions. *Drug Discov Today* 2015;**20**:122–8.
5. Usmani SS, Bedi G, Samuel JS, *et al.* THPdb: database of FDA-approved peptide and protein therapeutics. *PLoS One* 2017;**12**:e0181748.
6. Otvos L, Wade JD. Current challenges in peptide-based drug discovery. *Front Chem* 2014;**2**:62.
7. Clark GC, Casewell NR, Elliott CT, *et al.* Friends or foes? Emerging impacts of biological toxins. *Trends Biochem Sci* 2019;**44**:365–79.
8. Petricevich VL. Scorpion venom and the inflammatory response. *Mediators Inflamm* 2010;**2010**:903295.
9. Casewell NR, Jackson TNW, Laustsen AH, *et al.* Causes and consequences of snake venom variation. *Trends Pharmacol Sci* 2020;**41**:570–81.
10. Slagboom J, Kool J, Harrison RA, *et al.* Haemotoxic snake venoms: their functional activity, impact on snakebite victims and pharmaceutical promise. *Br J Haematol* 2017;**177**:947–59.
11. Sharma N, Naorem LD, Gupta S, *et al.* Computational resources in healthcare. *WIREs Data Min. Knowl Discov* 2021;e1437.
12. Pérez Santín E, Rodríguez Solana R, González García M, *et al.* Toxicity prediction based on artificial intelligence: a multidisciplinary overview. *WIREs Comput Mol Sci* 2021;e1516.
13. Mayr A, Klambauer G, Unterthiner T, *et al.* DeepTox: toxicity prediction using deep learning. *Front Environ Sci* 2016;**3**:80.
14. Banerjee P, Eckert AO, Schrey AK, *et al.* ProTox-II: a webserver for the prediction of toxicity of chemicals. *Nucleic Acids Res* 2018;**46**:W257–63.

15. Pu L, Naderi M, Liu T, *et al.* eToxPred: a machine learning-based approach to estimate the toxicity of drug candidates. *BMC Pharmacol Toxicol* 2019;**20**:2.
16. Saha S, Raghava GPS. BTXpred: prediction of bacterial toxins. *In Silico Biol* 2007;**7**:405–12.
17. Saha S, Raghava GPS. Prediction of neurotoxins based on their function and source. *In Silico Biol* 2007;**7**:369–87.
18. Naamati G, Askenazi M, Linial M. ClanTox: a classifier of short animal toxins. *Nucleic Acids Res* 2009;**37**:W363–8.
19. Wong ESW, Hardy MC, Wood D, *et al.* SVM-based prediction of propeptide cleavage sites in spider toxins identifies toxin innovation in an Australian tarantula. *PLoS One* 2013;**8**:e66279.
20. Gacesa R, Barlow DJ, Long PF. Machine learning can differentiate venom toxins from other proteins having non-toxic physiological functions. *PeerJ Comput Sci* 2016;**2**:e90.
21. Cole TJ, Brewer MS. TOXIFY: a deep learning approach to classify animal venom proteins. *PeerJ* 2019;**7**:e7200.
22. Pan X, Zuallaert J, Wang X, *et al.* ToxDL: deep learning using primary structure and domain embeddings for assessing protein toxicity. *Bioinformatics* 2021;**36**:5159–68.
23. Gupta S, Kapoor P, Chaudhary K, *et al.* In silico approach for predicting toxicity of peptides and proteins. *PLoS One* 2013;**8**:e73957.
24. Jain A, Kihara D. NNTox: gene ontology-based protein toxicity prediction using neural network. *Sci Rep* 2019;**9**:17923.
25. Wei L, Ye X, Xue Y, *et al.* ATSE: a peptide toxicity predictor by exploiting structural and evolutionary information based on graph neural network and attention mechanism. *Brief Bioinform* 2021;**5**:bbab041.
26. Wei L, Ye X, Sakurai T, *et al.* ToxIBTL: prediction of peptide toxicity based on information bottleneck and transfer learning. *Bioinformatics* 2022;**6**:1514–24.
27. Mishra NK, Singla D, Agarwal S, *et al.* ToxiPred: a server for prediction of aqueous toxicity of small chemical molecules in *T. Pyriformis*. *J Transl Toxicol* 2014;**1**:21–7.
28. Chaudhary K, Kumar R, Singh S, *et al.* A web server and mobile app for computing hemolytic potency of peptides. *Sci Rep* 2016;**6**:22843.
29. Win TS, Malik AA, Prachayasittikul V, *et al.* HemoPred: a web server for predicting the hemolytic activity of peptides. *Future Med Chem* 2017;**9**:275–91.
30. Sharma AK, Srivastava GN, Roy A, *et al.* ToxiM: a toxicity prediction tool for small molecules developed using machine learning and chemoinformatics approaches. *Front Pharmacol* 2017;**8**:880.
31. Lagunin AA, Dubovskaja VI, Rudik AV, *et al.* CLC-Pred: a freely available web-service for in silico prediction of human cell line cytotoxicity for drug-like compounds. *PLoS One* 2018;**13**:e0191838.
32. Hasan MM, Schaduangrat N, Basith S, *et al.* HLPpred-Fuse: improved and robust prediction of hemolytic peptide and its activity by fusing multiple feature representation. *Bioinformatics* 2020;**36**:3350–6.
33. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* 2021;**49**:D480–9.
34. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* 2000;**28**:45–8.
35. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 2006;**22**:1658–9.
36. Altschul SF, Gish W, Miller W, *et al.* Basic local alignment search tool. *J Mol Biol* 1990;**215**:403–10.
37. Saha S, Raghava GPS. AlgPred: prediction of allergenic proteins and mapping of IgE epitopes. *Nucleic Acids Res* 2006;**34**:W202–9.
38. Sharma N, Patiyal S, Dhall A, *et al.* AlgPred 2.0: an improved method for predicting allergenic proteins and mapping of IgE epitopes. *Brief Bioinform* 2021;**22**:bbaa294.
39. Vens C, Rosso M-N, Danchin EGJ. Identifying discriminative classification-based motifs in biological sequences. *Bioinformatics* 2011;**27**:1231–8.
40. Hasan MM, Khatun MS, Kurata H. iLBE for computational identification of linear B-cell epitopes by integrating sequence and evolutionary features. *Genom Proteom Bioinform* 2020;**18**:593–600.
41. Khatun MS, Hasan MM, Shoombuatong W, *et al.* ProIn-Fuse: improved and robust prediction of proinflammatory peptides by fusing of multiple feature representations. *J Comput Aided Mol Des* 2020;**34**:1229–36.
42. Pande A, Patiyal S, Lathwal A, *et al.* Computing wide range of protein/peptide features from their sequence and structure. *bioRxiv* 2019;599126.
43. Altschul SF, Madden TL, Schäffer AA, *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**:3389–402.
44. Kumar M, Gromiha MM, Raghava GPS. Identification of DNA-binding proteins using support vector machines and evolutionary profiles. *BMC Bioinformatics* 2007;**8**:463.
45. Pedregosa F, Varoquaux G, Gramfort A, *et al.* Scikit-learn: machine learning in python. *J Mach Learn Res* 2011;**12**:2825–30.
46. Ke G, Meng Q, Finley T, *et al.* Lightgbm: a highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst* 2017;**30**:3146–54.
47. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. *Mach Learn* 2006;**63**:3–42.
48. Tolles J, Meurer WJ. Logistic regression: relating patient characteristics to outcomes. *J Am Med Assoc* 2016;**316**:533–4.
49. Zhang H. Exploring conditions for the optimality of Naive Bayes. *Int J Pattern Recognit Artif Intell* 2005;**19**:183–98.
50. Fürnkranz J. Decision tree. *Encycl Mach Learn* 2011;**63**:263–7.
51. Mucherino A, Papajorgji PJ, Pardalos PM. *k-nearest neighbor classification. In: Data Mining in Agriculture*, Vol. **34**. New York, NY: Springer Optimization and Its Applications, 2009, 83–106.
52. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Proc. 22nd ACM SIGKDD. *Int Conf Knowl Discov Data Min* 2016; 785–94.
53. Zhang C, Shao X, Li D. Knowledge-based support vector classification based on C-SVC. *Proc Comput Sci* 2013;**17**:1083–90.
54. Agrawal P, Bhagat D, Mahalwal M, *et al.* AntiCP 2.0: an updated model for predicting anticancer peptides. *Brief Bioinform* 2021;**22**:bbaa153.
55. Sharma N, Patiyal S, Dhall A, *et al.* ChAlPred: a web server for prediction of allergenicity of chemical compounds. *Comput Biol Med* 2021;**136**:104746.