SOFTWARE **Open Access**

MultiToxPred 1.0: a novel comprehensive tool for predicting 27 classes of protein toxins using an ensemble machine learning approach

Jorge F. Beltrán[1*], Lisandra Herrera-Belén[2], Fernanda Parraguez-Contreras[1], Jorge G. Farías[1], Jorge Machuca-Sepúlveda[1] and Stefania Short[1]

*Correspondence:
beltran.lissabet.jf@gmail.com

[1] Department of Chemical Engineering, Faculty of Engineering and Science, Universidad de La Frontera, Ave. Francisco Salazar, 01145 Temuco, Chile
[2] Departamento de Ciencias Básicas, Facultad de Ciencias, Universidad Santo Tomas, Temuco, Chile

Abstract

Protein toxins are defense mechanisms and adaptations found in various organisms and microorganisms, and their use in scientific research as therapeutic candidates is gaining relevance due to their effectiveness and specificity against cellular targets. However, discovering these toxins is time-consuming and expensive. In silico tools, particularly those based on machine learning and deep learning, have emerged as valuable resources to address this challenge. Existing tools primarily focus on binary classification, determining whether a protein is a toxin or not, and occasionally identifying specific types of toxins. For the first time, we propose a novel approach capable of classifying protein toxins into 27 distinct categories based on their mode of action within cells. To accomplish this, we assessed multiple machine learning techniques and found that an ensemble model incorporating the Light Gradient Boosting Machine and Quadratic Discriminant Analysis algorithms exhibited the best performance. During the tenfold cross-validation on the training dataset, our model exhibited notable metrics: 0.840 accuracy, 0.827 F1 score, 0.836 precision, 0.840 sensitivity, and 0.989 AUC. In the testing stage, using an independent dataset, the model achieved 0.846 accuracy, 0.838 F1 score, 0.847 precision, 0.849 sensitivity, and 0.991 AUC. These results present a powerful next-generation tool called MultiToxPred 1.0, accessible through a web application. We believe that MultiToxPred 1.0 has the potential to become an indispensable resource for researchers, facilitating the efficient identification of protein toxins. By leveraging this tool, scientists can accelerate their search for these toxins and advance their understanding of their therapeutic potential.

Keywords: Protein toxin, Machine learning, Prediction, Therapeutic

Introduction

Over the course of evolution, many organisms and microorganisms have developed the ability to express different types of protein toxins (PT) as part of their defense mechanisms and adaptations to the environment [1]. These proteins can be found in animals [2] and poisonous plants [3–6], as well as in pathogenic bacteria [5, 6]. PTs have a wide range of molecular targets, which has allowed them to be extensively studied as therapeutic candidates for the treatment of various diseases, generally, such as pain [7, 8], cancer [9–14], autoimmune diseases [15], cardiovascular diseases [16, 17], neurodegenerative diseases [18], viral [14, 19] and bacterial [14] infections, among others Currently, there are different proposals to classify PTs, and one of these is their classification into three main groups, (1) toxins that hinder or interfere with cellular processes through their enzymatic activity, (2) toxins that cause harm to cells by compromising the integrity of their membranes, and (3) toxins that interfere with the regular electrical functioning of the nervous system in an intoxicated organism [1]. However, the fact that PTs have a wide variety of molecular targets makes a more specific classification of these not entirely clear at present. In this regard, it has been reported that PTs can act on various molecular targets among which we find the cell membrane [20–22], voltage-gated sodium channel [23], voltage-gated calcium channel [24, 25], voltage-gated potassium channel [26, 27], acetylcholine receptor [28, 29], G-protein coupled receptor [30], and bradykinin receptor [31], among many others.

In recent years, the study of protein toxins has increased due to the great potential they represent as therapeutic drugs. In this regard, various in vitro, in vivo [32], and in silico [33] methodologies have been evaluated for their study. Among the in silico methodologies, the use of bioinformatics tools [34–37] and, more recently, machine learning (ML) [33], has gained greater relevance as it allows for the acceleration and reduction of costs of resources allocated to the search for PTs. Particularly, ML constitutes a robust and modern strategy for the discovery of pharmaceutical candidates [38, 39], with PTs being no exception in this context. Currently, there are several works based on machine learning and deep learning that generally, following a binary classification approach, allow discrimination between PTs and non-PTs. These tools are NTXpred [40], Yang and Li's method [41], Jayaraman et al.'s method [42], Kumar et al.'s method [43], NNTox [44], TOXIFY [45], ClanTox [46], ToxClassifier [47], ToxinPred2 [33], SpiderP [48], BTXpred [49], ToxDL [50], ATSE [51], ToxIBTL [52], ToxinMI [53], Toxicity-vib [54], and CSM-Toxin [55], which have been of great utility in the field of PT study. These tools undoubtedly greatly aid in the discovery of new toxins; however, they follow a binary classification approach where the output only informs if a protein is a PT or not. Taking into account the wide variety of molecular targets that PTs act upon, it would be interesting to approach a more specific prediction method that would allow us to elucidate more specific cellular targets. Following this idea, for the first time in this work, the development of ML models for the multiple classification of 27 different classes of PTs with different modes of cellular action was evaluated.

Methods

Data sets

The amino acid sequences of toxins used in this work were obtained from the Universal Protein Resource (UniProt) [56]. These sequences were only selected based on the following criteria, (1) the sequence must be reviewed, (2) the sequence has at least one scientific publication demonstrating the respective PT activity, and (3) the sequence must be complete. In this regard, a certain number of PT sequences were identified considering the reported target for each of these. Below are the number of identified PT sequences (sn) with their respective cellular targets or mode of action in the cell:

acetylcholine receptor inhibiting toxin (sn = 493), blood coagulation cascade activating toxin (sn = 107), blood coagulation cascade inhibiting toxin (ns = 133), bradykinin receptor impairing toxin (sn = 35), calcium-activated potassium channel impairing toxin (sn = 72), cell adhesion impairing toxin (sn = 228), chloride channel impairing toxin (sn = 20), complement system impairing toxin (sn = 27), dermonecrotic toxin (sn = 216), enterotoxin (ns = 101), fibrinogenolytic toxin (sn = 102), fibrinolytic toxin (sn = 41), G-protein coupled acetylcholine receptor impairing toxin (sn = 29), G-protein coupled receptor impairing toxin (sn = 228), hemorrhagic toxin (sn = 62), hemostasis impairing toxin (sn = 942), platelet aggregation activating toxin (sn = 74), platelet aggregation inhibiting toxin (sn = 350), potassium channel impairing toxin (sn = 664), proton-gated sodium channel impairing toxin (sn = 25), ryanodine-sensitive calcium-release channel impairing toxin (sn = 27), target cell cytoplasm (sn = 16), target cell membrane (sn = 418), voltage-gated calcium channel impairing toxin (sn = 247), voltage-gated chloride channel impairing toxin (sn = 18), voltage-gated potassium channel impairing toxin (sn = 508), and voltage-gated sodium channel impairing toxin (sn = 840) (Additional files 1–8). On the other hand, 600 random amino acid sequences of different lengths were generated, which were considered non-PT.

Calculation of molecular descriptors and balancing of the data set

From all the sequences, the calculation of two types of molecular descriptors widely used in the development of predictive models from primary protein structures was carried out: pseudo amino acid composition (PAAC, lamda = 5, weight = 0.05) [57], and dipeptide composition descriptors (DPC) [58]. Both molecular descriptors were computed with the Python propy3 package (https://pypi.org/project/propy3/) was used for the calculation of these molecular descriptors.

Subsequently, the resulting data set was labeled for later balancing and evaluation with classification algorithms. Because the data set contains labeled classes (PT and non-PT) with an imbalanced numerical proportion, its balance was carried out through the synthetic minority over-sampling technique (SMOTE). Imbalanced data sets can cause a bias in predictive models, and in this sense, SMOTE is a data preprocessing technique used to deal with the class imbalance problem in machine learning data sets. In this technique, synthetic examples of minority classes are generated. This is done by taking examples from minority classes and creating similar but slightly modified examples, "oversampling" the minority classes to balance the data set [59]. The Python imbalanced-learn package (https://pypi.org/project/imbalanced-learn/) was used to balance the data set with SMOTE.

Training, cross-validation, and testing

In this study, nine machine learning classification algorithms were evaluated: Random Forest (RF), Multi-layer Perceptron (MLP), eXtreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), Logistic Regression (LR), Naïve Bayes (NB), *k*-nearest neighbors (*k*-NN), and Quadratic Discriminant Analysis (QDA). Training with all the classifiers was conducted on 80% of the complete dataset, which underwent tenfold cross-validation. The remaining 20% of the data (independent dataset) was used to evaluate the performance of the trained models. The mentioned analyses were carried out using the libraries scikit-learn (https://pypi.org/project/scikit-learn/), XGBoost (https://pypi.org/project/xgboost/), and Microsoft LightGBM (https://pypi.org/project/lightgbm/). In this study, we evaluated the StackingClassifier, which is a meta-ensembling technique that leverages the strengths of diverse base learners by stacking their predictions as input for a final estimator. This method effectively combines multiple classification models, each of which may capture different patterns within the data. The justification for employing a StackingClassifier lies in its ability to blend various predictive models, potentially leading to better generalization on unseen data. By using predictions of base learners as features, the meta-learner can learn to correct the individual classifier mistakes, thereby improving overall accuracy. This approach is supported by empirical studies demonstrating its superiority over individual classifiers and even other ensemble methods when carefully implemented. The mean of the performance measures used to evaluate the models in cases of multiple classifications, both in the training stage through cross-validation and in the testing stage, were the following:

$$Sensitivity\,(TPR) = TP/(TP + FN) \tag{1}$$

$$Accuracy\,(ACC) = TP + TN/(TP + FP + FN + TN) \tag{2}$$

$$Precision\,(PPV) = TP/(TP + FP) \tag{3}$$

$$F1\,score\,(F1) = 2TP/(2TP + FP + FN) \tag{4}$$

In this research, we also assessed the effectiveness of the predictive models using the area under the curve (AUC) of the receiver operating characteristic (ROC) plot. A modern web application was developed using the Python 3.11 programming language (https://www.python.org/) for making predictions of PT. The first version of the web application, named MultiToxPred 1.0, scores the outputs with a probability from 0 to 1. Figure 1 shows the working architecture used in this study.

Results

The tenfold cross-validation analysis on the training data showed that the RF, XGBoost, and LightGBM algorithms displayed the best performance in the classification of PTs using the PAAC molecular descriptor (Table 1). On the other hand, when evaluating the DPC molecular descriptor, it was observed that LightGBM again showed good performance, as did the MLP and QDA algorithms (Table 2). The LR algorithm showed good performance with the use of DPC, however, low performance measures were obtained with PAAC and NB, with the latter algorithm having the worst performance with both evaluated descriptors (Tables 1 and 2). In the testing stage (Tables 3 and 4), in general, there was a consistent increase in the evaluated performance measures, which is indicative that the models are efficient at predicting PTs on independent data sets.

Considering the performance of the best algorithms in this study, both in the training and testing stages, we proceeded to evaluate the development of predictive models of PTs using an ensemble approach. In this direction, for the case of the PAAC molecular descriptor, an ensemble of RF and LightGBM was generated. For

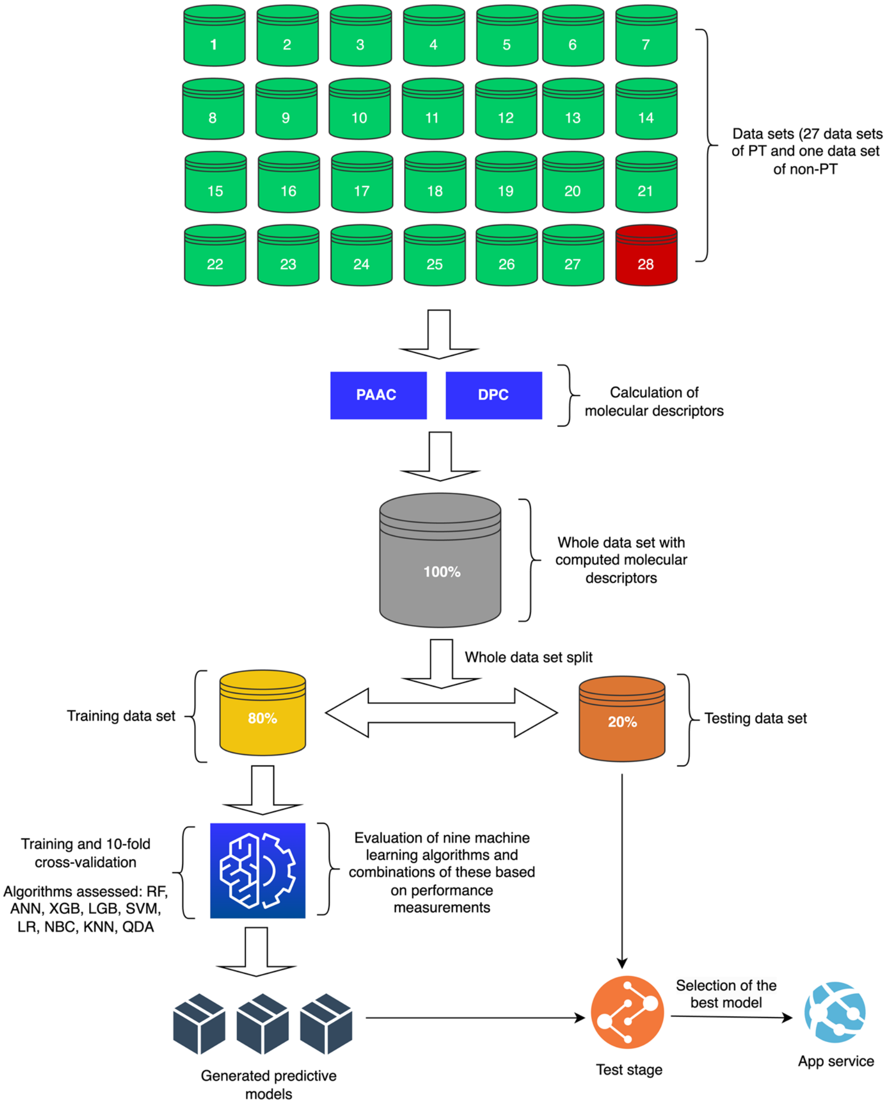
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
Data sets (27 data sets of PT and one data set of non-PT
PAAC
DPC
Calculation of molecular descriptors
100%
Whole data set with computed molecular descriptors
Whole data set split
Training data set
80%
20%
Testing data set
Training and 10-fold cross-validation
Algorithms assessed: RF, ANN, XGB, LGB, SVM, LR, NBC, KNN, QDA
Evaluation of nine machine learning algorithms and combinations of these based on performance measurements
Generated predictive models
Test stage
Selection of the best model
App service

Fig. 1 From the total dataset of amino acid sequences corresponding to different types of protein toxins with different modes of action in the cell (n = 27) and non-toxins (n = 1) randomly generated, the molecular descriptors PAAC and DPC were calculated. Subsequently, eight machine learning algorithms were evaluated, first on a training dataset (80%) which was subjected to tenfold cross-validation. Then, the generated models were evaluated on a test dataset (20%) (independent dataset). The final stage consisted of selecting the best predictive model for its incorporation into a web application called MultiToxPred 1.0

Table 1 Ten-fold cross-validation on the training dataset using the PAAC molecular descriptor

Algorithm	ACC	F1	PPV	TPR	AUC
RF	0.760	0.753	0.750	0.758	0.97
MLP	0.745	0.731	0.732	0.744	0.98
XGBoost	0.754	0.749	0.748	0.752	0.98
LightGBM	0.754	0.750	0.749	0.752	0.98
LR	0.530	0.504	0.501	0.529	0.93
NB	0.451	0.414	0.447	0.450	0.90
k-NN	0.753	0.735	0.743	0.752	0.95
QDA	0.700	0.673	0.671	0.699	0.95

Table 2 Ten-fold cross-validation on the training dataset using the DPC molecular descriptor

Algorithm	ACC	F1	PPV	TPR	AUC
RF	0.774	0.767	0.765	0.773	0.97
MLP	0.791	0.782	0.784	0.789	0.98
XGBoost	0.780	0.776	0.775	0.779	0.98
LightGBM	0.781	0.777	0.777	0.780	0.98
LR	0.792	0.777	0.779	0.791	0.98
NB	0.675	0.649	0.686	0.674	0.95
k-NN	0.749	0.727	0.744	0.748	0.95
QDA	0.806	0.791	0.810	0.805	0.92

Table 3 Performance on the testing dataset using the PAAC molecular descriptor

Algorithm	ACC	F1	PPV	TPR	AUC
RF	0.764	0.761	0.755	0.768	0.97
MLP	0.742	0.733	0.730	0.748	0.98
XGBoost	0.757	0.756	0.753	0.761	0.98
LightGBM	0.757	0.757	0.754	0.761	0.98
LR	0.525	0.506	0.497	0.532	0.94
NB	0.445	0.411	0.444	0.448	0.90
k-NN	0.753	0.741	0.751	0.759	0.96
QDA	0.701	0.678	0.674	0.704	0.95

Table 4 Performance on the testing dataset using the DPC molecular descriptor

Algorithm	ACC	F1	PPV	TPR	AUC
RF	0.779	0.777	0.773	0.783	0.97
MLP	0.794	0.791	0.787	0.799	0.98
XGBoost	0.784	0.784	0.782	0.788	0.99
LightGBM	0.787	0.788	0.787	0.791	0.99
LR	0.793	0.781	0.783	0.797	0.98
NB	0.667	0.645	0.682	0.670	0.95
k-NN	0.745	0.730	0.754	0.752	0.95
QDA	0.814	0.802	0.826	0.818	0.93

the DPC molecular descriptor, three ensembles were evaluated: MLP + LightGBM, MLP + QDA, and LightGBM + QDA. It is important to note that, regardless of the descriptor evaluated, the ensemble-based strategy allowed for better performance measures compared to the individual algorithms, both in the training and testing stages (Table 5).

In the case of DPC, it was observed that these performance measures increased significantly, to a degree > 0.8, which indicates the robustness of this approach using this molecular descriptor and the algorithms used in the ensemble technique (Table 5). In consequence, these results demonstrate that our predictive strategy constitutes a robust approach for the prediction of PTs, taking into account the complexity of

Table 5 Ten-fold cross-validation on the training and testing datasets using the PAAC and DPC molecular descriptors via ensemble algorithms

Ensemble algorithms	ACC	F1	PPV	TPR	AUC
PAAC					
RF + LightGBM $^{\text{Training–CV}}$	0.789	0.781	0.779	0.788	0.98
RF + LightGBM $^{\text{Testing}}$	0.800	0.796	0.793	0.803	0.99
DPC					
MLP + LightGBM $^{\text{Training–CV}}$	0.813	0.801	0.801	0.812	0.99
MLP + LightGBM $^{\text{Testing}}$	0.816	0.806	0.799	0.820	0.99
MLP + QDA$^{\text{Training–CV}}$	0.831	0.817	0.822	0.830	0.99
MLP + QDA$^{\text{Testing}}$	0.844	0.837	0.843	0.847	0.99
LightGBM + QDA$^{\text{Training–CV}}$	0.840*	0.827*	0.836*	0.840*	0.99*
LightGBM + QDA$^{\text{Testing}}$	0.846*	0.838*	0.847*	0.849*	0.99*

CV: cross-validation, *: best performance measurements obtained

the study problem, which involves a high number of classes (27 in total). Of all the ensemble strategies evaluated, we noted that the resulting model from the LightGBM and QDA algorithms performed best during the cross-validation and testing phases (Table 5). In this direction, this model was selected for incorporation into a web application.

The web application developed in this study presents a modern and intuitive user interface, which allows carrying out PTs predictions. The results of the analyses can be downloaded in a csv file and/or can be selected and ranked in the application based on their respective probabilistic score, where scores greater than 0.5 indicate the probability that an unknown amino acid sequence introduced by the user corresponds to one of the 27 proposed classes (PT type and non-PT) in this work. The application, named Multi-ToxPred 1.0, is in its first version and is available for free use at https://www.biochemintelli.com/MultiToxPred-v1.

Discussion

Currently, proteins and peptides (PT) are being extensively studied due to their great potential as therapeutic drugs in the treatment of various diseases, including immunological conditions, metabolic disorders, and neurodegenerative diseases, among others [1, 2, 60, 61]. The diversity in chemical nature and the complexity of PT structures, which are often derived from varied natural sources, make the study of these biomolecules, in most cases, a laborious and costly task. This is reflected in the numerous in vitro and in vivo experimental trials needed to confirm their effectiveness and safety [32]. On the other hand, machine learning techniques represent a robust alternative to rapidly and cost-effectively approach the identification of the functionality of peptides and proteins. These methods can predict the properties and behavior of PT based solely on their primary sequence, which can expedite the drug development process [33].

As mentioned above, numerous studies focusing on the prediction of PT behavior have been conducted. However, to date, no approach has been evaluated for predicting the specific mode of action of these biomolecules within the cell. It is well-documented, for example, that PTs from venomous animals target ion channels, which are in turn classified into several types based on the ions they transport [62, 63]. Predicting a more specific mode of action would not only determine whether a protein or peptide is a toxin but would also allow the elucidation of its modes of action within the cell. In some cases, it may even reveal its molecular target. Certainly, this would have a significant impact on the field of PT study. Considering all the aspects mentioned above, the motivation of this study was focused on an "out of the box" approach. The present study allowed the development of robust strategies that facilitate the prediction of PT in numerous classes, using multiple classification techniques, in contrast to state-of-the-art methods and tools that are based solely on binary classification (PT or non-PT).

Both descriptors used in this study (PAAC and DPC), are widely used in most of the works that apply machine learning techniques for the prediction of the biological functionality of peptides and proteins. In this work, we demonstrate that through the combined use of the LightGBM and QDA algorithms, the best performance measures are obtained with DPC (Table 3). The DPC molecular descriptor is a technique used in bioinformatics that is responsible for representing the properties of proteins or peptides. This descriptor is based on the idea that each dipeptide (a chain of two amino acids) has particular physicochemical properties and its frequency in the protein can provide significant information about its structure and function. In other words, DPC represents the frequency of each possible dipeptide in the total sequence of a protein, thus providing a global view of its composition and, potentially, its biological behavior. It is a tool widely used in the prediction of protein functionality, as it provides a general portrait of the molecular composition of the protein of interest [58]. The DPC has been assessed in various predictive toxin studies using machine learning techniques, proving its efficacy in this domain [40, 41, 43, 49]. This aligns, to a degree, with the findings of our study.

For the first time, we evaluated the development of a predictive model using an ensemble approach with LightGBM and QDA for PT predictions, which allowed us to obtain the best performance measurements (Table 5). The LightGBM is a gradient boosting-based machine learning algorithm that differs from other boosting algorithms in its ability to handle large data sets and its computational efficiency. It uses a leaf-based tree growth approach instead of the traditional depth-based growth, allowing you to focus on the regions of greatest loss and improving model accuracy. These features make LightGBM particularly useful for tasks that require high efficiency and precision [64]. The QDA is a statistical classification technique used in supervised learning. This method is based on Bayesian inference and assumes that each class in the dataset has its own covariance matrix [65]. Both algorithms have also been used in the classification of peptides and proteins, for example, LightGBM has been used in the prediction of anti-cancer peptides [66], protein structural class [67], protein–protein interactions [68], protein-ATP binding residues [69], and ion channels [70], among others. On the other hand, QDA has been used in the prediction of tumor T-cell antigens [71], antimicrobial peptides [72, 73], protein motifs [74], and protein subcellular location [75], among others.

Addressing the challenges inherent in predicting the specific mode of action of PTs in the cell using machine learning techniques will undoubtedly be an important focus for future research. One significant challenge is dealing with imbalanced data, as in many cases, the availability of labeled data for certain classes of PT is limited compared to others. Oversampling methods could be useful, and in this work, we demonstrate that by using SMOTE it is possible to obtain robust predictive models for predicting the molecular targets of PTs. As demonstrated in this study, the SMOTE technique has been used for the augmentation of amino acid sequence data [76], and it is considered the most used oversampling technique due to its fast and good results [77]. However, the exploration of other synthetic data generation techniques for protein and peptides, such as the use of adversarial neural networks [76, 78, 79], could be considered in future work to achieve the same purpose, which could significantly improve the performance of the predictive models.

We believe that this study serves as an initial springboard for the development of machine learning-based predictive tools to predict the specific functionalities of protein toxins. By leveraging sophisticated machine learning algorithms, it is possible to analyze vast amounts of biological data and obtain meaningful insights that would otherwise be too complex or time-consuming to obtain through traditional methods. In this direction, we believe that MultiToxPred 1.0 represents a novel tool that could be key for the study of PTs.

Conclusions

For the first time, this study demonstrated that using a multiple classification approach aided with SMOTE, it is possible to predict the mode of action of a PT in the cell. Of all the machine learning algorithms evaluated, the best performance was observed with the combination of LightGBM and QDA using the DPC molecular descriptor. The model generated with these two combined algorithms was selected for incorporation into the MultiToxPred 1.0 web application, a free resource that facilitates PT predictions. These results highlight the power of machine learning techniques in predicting the functionality of PTs and suggest that MultiToxPred 1.0 may be an important tool in the discovery of these proteins as well as in the therapeutic area.

Availability and requirements

Project name: MultiToxPred-v1.

Project home page: https://www.biochemintelli.com/MultiToxPred-v1

Operating system(s): Platform independent.

Programming language: Python 3.9

Other requirements: Python 3.7 or higher, scikit-learn, biopython, numpy, and pandas.

License: MIT License.

Any restrictions to use by non-academics: None.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-024-05748-z.

Additional file 1. Toxin class information.

Additional file 2. ROC curves in 10-fold cross-validation phase using DPC.

Additional file 3. ROC curves in testing phase using DPC.

Additional file 4. ROC curves in 10-fold cross-validation phase using PAAC.

Additional file 5. ROC curves in testing phase using PAAC.

Additional file 6. ROC curves in 10-fold cross-validation phase using PAAC and DPC.

Additional file 7. ROC curves in testing phase using PAAC and DPC.

Additional file 8. Figure legends S1–S6.

Author contributions

JFB developed the model and application and wrote part of the manuscript. LHB wrote part of the manuscript and provided the main feedback. FPC was responsible for the collection and curation of data. JGF, JMS, and SS provided feedback. All authors reviewed the manuscript.

Funding

None.

Availability of data and materials

All the amino acid sequences as well as the code used in this study are available in the GitHub repository: https://github.com/jfbldevs/MultiToxPred.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 7 August 2023 Accepted: 14 March 2024
Published online: 12 April 2024

References

1. Shapira A, Benhar I. Toxin-based therapeutic approaches. Toxins. 2010;2:2519–83.
2. Chen N, Xu S, Zhang Y, Wang F. Animal protein toxins: origins and therapeutic applications. Biophys Rep. 2018;4:233–42.
3. Kocyigit E, Kocaadam-Bozkurt B, Bozkurt O, Ağagündüz D, Capasso R. Plant toxic proteins: their biological activities, mechanism of action and removal strategies. Toxins (Basel). 2023;15:356.
4. Dang L, Van Damme EJM. Toxic proteins in plants. Phytochemistry. 2015;117:51–64.
5. Sandvig K, Torgersen ML, Engedal N, Skotland T, Iversen T-G. Protein toxins from plants and bacteria: probes for intracellular transport and tools in medicine. FEBS Lett. 2010;584:2626–34.
6. Sandvig K, van Deurs B. Delivery into cells: lessons learned from plant and bacterial toxins. Gene Ther. 2005;12:865–72.
7. Essack M, Bajic VB, Archer JAC. Conotoxins that confer therapeutic possibilities. Mar Drugs. 2012;10:1244–65.
8. Brust A, Palant E, Croker DE, Colless B, Drinkwater R, Patterson B, et al. χ-Conopeptide pharmacophore development: toward a novel class of norepinephrine transporter inhibitor (Xen2174) for pain. J Med Chem. 2009;52:6991–7002.
9. El-Didamony SE, Amer RI, El-Osaily GH. Formulation, characterization and cellular toxicity assessment of a novel bee-venom microsphere in prostate cancer treatment. Sci Rep. 2022;12:13213.
10. Wolf P. Targeted toxins for the treatment of prostate cancer. Biomedicines. 2021;9:986.
11. Antignani A, Ho ECH, Bilotta MT, Qiu R, Sarnvosky R, FitzGerald DJ. Targeting receptors on cancer cells with protein toxins. Biomolecules. 2020;10:1331.
12. Weerakkody LR, Witharana C. The role of bacterial toxins and spores in cancer therapy. Life Sci. 2019;235: 116839.
13. Sharma PC, Sharma D, Sharma A, Bhagat M, Ola M, Thakur VK, et al. Recent advances in microbial toxin-related strategies to combat cancer. Semin Cancer Biol. 2022;86:753–68.
14. Serna N, Sánchez-García L, Unzueta U, Díaz R, Vázquez E, Mangues R, et al. Protein-based therapeutic killing for cancer therapies. Trends Biotechnol. 2018;36:318–35.
15. Madhumathi J, Verma RS. Therapeutic targets and recent advances in protein immunotoxins. Curr Opin Microbiol. 2012;15:300–9.
16. Frangieh J, Rima M, Fajloun Z, Henrion D, Sabatier J-M, Legros C, et al. Snake venom components: tools and cures to target cardiovascular diseases. Molecules. 2021;26:2223.
17. Kini RM, Koh CY. Snake venom three-finger toxins and their potential in drug development targeting cardiovascular diseases. Biochem Pharmacol. 2020;181: 114105.
18. de Souza JM, Goncalves BDC, Gomez MV, Vieira LB, Ribeiro FM. Animal toxins as therapeutic tools to treat neurodegenerative diseases. Front Pharmacol. 2018;9:336857.
19. Utkin Y, Siniavin A, Kasheverov I, Tsetlin V. Antiviral effects of animal toxins: is there a way to drugs? Int J Mol Sci. 2022;23:3634.
20. Peraro MD, van der Goot FG. Pore-forming toxins: ancient, but never really out of fashion. Nat Rev Microbiol. 2016;14:77–92.
21. Gilbert RJC. Pore-forming toxins. Cell Mol Life Sci. 2002;59:832–44.
22. Ulhuq FR, Mariano G. Bacterial pore-forming toxins. Microbiology. 2022;168:001154.
23. Groome JR. Historical perspective of the characterization of conotoxins targeting voltage-gated sodium channels. Mar Drugs. 2023;21:209.

24. Antunes FTT, Campos MM, Carvalho VPR, da Silva Junior CA, Magno LAV, de Souza AH, et al. Current drug development overview: targeting voltage-gated calcium channels for the treatment of pain. Int J Mol Sci. 2023;24:9223.
25. Bourinet E, Zamponi GW. Block of voltage-gated calcium channels by peptide toxins. Neuropharmacology. 2017;127:109–15.
26. Kuzmenkov AI, Gigolaev AM, Pinheiro-Junior EL, Peigneur S, Tytgat J, Vassilevski AA. Methionine-isoleucine dichotomy at a key position in scorpion toxins inhibiting voltage-gated potassium channels. Toxicon. 2023;231: 107181.
27. Wulff H, Castle NA, Pardo LA. Voltage-gated potassium channels as therapeutic targets. Nat Rev Drug Discov. 2009;8:982–1001.
28. Green BT, Welch KD, Panter KE, Lee ST. Plant toxins that affect nicotinic acetylcholine receptors: a review. Chem Res Toxicol. 2013;26:1129–38.
29. Tsetlin VI, Hucho F. Snake and snail toxins acting on nicotinic acetylcholine receptors: fundamental aspects and medical applications. FEBS Lett. 2004;557:9–13.
30. Näreoja K, Näsman J. Selective targeting of G-protein-coupled receptor subtypes with venom peptides. Acta Physiol. 2012;204:186–201.
31. Guido-Patiño JC, Plisson F. Profiling hymenopteran venom toxins: protein families, structural landscape, biological activities, and pharmacological benefits. Toxicon X. 2022;14: 100119.
32. Duracova M, Klimentova J, Fucikova A, Dresler J. Proteomic methods of detection and quantification of protein toxins. Toxins. 2018;10:99.
33. Sharma N, Naorem LD, Jain S, Raghava GPS. ToxinPred2: an improved method for predicting toxicity of proteins. Brief Bioinform. 2022;23:bbac174.
34. Doxey AC, Mansfield MJ, Montecucco C. Discovery of novel bacterial toxins by genomics and computational biology. Toxicon. 2018;147:2–12.
35. Ojeda P, Ramírez D, Alzate-Morales J, Caballero J, Kaas Q, González W. Computational studies of snake venom toxins. Toxins. 2017;10:8.
36. Tan PTJ. Bioinformatics for venom and toxin sciences. Brief Bioinform. 2003;4:53–62.
37. Kaas Q, Craik D. Bioinformatics-aided venomics. Toxins. 2015;7:2159–87.
38. Dara S, Dhamercherla S, Jadav SS, Babu CM, Ahsan MJ. Machine learning in drug discovery: a review. Artif Intell Rev. 2022;55:1947–99.
39. Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 2019;18:463–77.
40. Saha S, Raghava GPS. Prediction of neurotoxins based on their function and source. In Silico Biol. 2007;7:369–87.
41. Yang L, Li Q. Prediction of presynaptic and postsynaptic neurotoxins by the increment of diversity. Toxicol In Vitro. 2009;23:346–8.
42. Bhosale H, Ramakrishnan V, Jayaraman VK. Support vector machine-based prediction of pore-forming toxins (PFT) using distributed representation of reduced alphabets. J Bioinform Comput Biol. 2021;19:2150028.
43. Gupta S, Kapoor P, Chaudhary K, Gautam A, Kumar R, Raghava GPS. In Silico Approach for Predicting Toxicity of Peptides and Proteins. PLoS ONE. 2013;8: e73957.
44. Jain A, Kihara D. NNTox: gene ontology-based protein toxicity prediction using neural network. Sci Rep. 2019;9:17923.
45. Cole TJ, Brewer MS. TOXIFY: a deep learning approach to classify animal venom proteins. PeerJ. 2019;7: e7200.
46. Naamati G, Askenazi M, Linial M. ClanTox: a classifier of short animal toxins. Nucleic Acids Res. 2009;37:W363–8.
47. 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.
48. Wong ESW, Hardy MC, Wood D, Bailey T, King GF. SVM-based prediction of propeptide cleavage sites in spider toxins identifies toxin innovation in an Australian Tarantula. PLoS ONE. 2013;8: e66279.
49. Saha S, Raghava GPS. BTXpred: prediction of bacterial toxins. In Silico Biol. 2007;7:405–12.
50. Pan X, Zuallaert J, Wang X, Shen H-B, Campos EP, Marushchak DO, et al. ToxDL: deep learning using primary structure and domain embeddings for assessing protein toxicity. Bioinformatics. 2021;36:5159–68.
51. Wei L, Ye X, Xue Y, Sakurai T, Wei L. ATSE: a peptide toxicity predictor by exploiting structural and evolutionary information based on graph neural network and attention mechanism. Brief Bioinform. 2021;22:bbab041.
52. Wei L, Ye X, Sakurai T, Mu Z, Wei L. ToxIBTL: prediction of peptide toxicity based on information bottleneck and transfer learning. Bioinformatics. 2022;38:1514–24.
53. Wei L, Ye X, Sakurai T. ToxinMI. In: Proceedings of the Conference on Research in Adaptive and Convergent Systems, New York, NY, USA: ACM; 2022. p. 77–82.
54. Zhao Z, Gui J, Yao A, Le NQK, Chua MCH. Improved prediction model of protein and peptide toxicity by integrating channel attention into a convolutional neural network and gated recurrent units. ACS Omega. 2022;7:40569–77.
55. Morozov V, Rodrigues CHM, Ascher DB. CSM-Toxin: a web-server for predicting protein toxicity. Pharmaceutics. 2023;15:431.
56. Bateman A, Martin M-J, Orchard S, Magrane M, Agivetova R, Ahmad S, et al. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 2021;49:D480–9.
57. Chou K-C. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins Struct Funct Genetics. 2001;43:246–55.
58. Petrilli P. Classification of protein sequences by their dipeptide composition. Bioinformatics. 1993;9:205–9.
59. Elreedy D, Atiya AF. A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance. Inf Sci. 2019;505:32–64.
60. Leader B, Baca QJ, Golan DE. Protein therapeutics: a summary and pharmacological classification. Nat Rev Drug Discov. 2008;7:21–39.
61. Ahn H-J, Park C-S, Cho JJ. Application of therapeutic protein-based fusion toxins. Mol Cell Toxicol. 2019;15:369–81.
62. Kalia J, Milescu M, Salvatierra J, Wagner J, Klint JK, King GF, et al. From foe to friend: using animal toxins to investigate ion channel function. J Mol Biol. 2015;427:158–75.

63. Herzig V, Cristofori-Armstrong B, Israel MR, Nixon SA, Vetter I, King GF. Animal toxins—nature's evolutionary-refined toolkit for basic research and drug discovery. Biochem Pharmacol. 2020;181: 114096.
64. Bentéjac C, Csörgő A, Martínez-Muñoz G. A comparative analysis of gradient boosting algorithms. Artif Intell Rev. 2021;54:1937–67.
65. Qin Y. A review of quadratic discriminant analysis for high-dimensional data. WIREs Comput Stat. 2018;10:1434.
66. Liang X, Li F, Chen J, Li J, Wu H, Li S, et al. Large-scale comparative review and assessment of computational methods for anti-cancer peptide identification. Brief Bioinform. 2021;22:bbaa312.
67. Zhang Y, Gao S, Cai P, Lei Z, Wang Y. Information entropy-based differential evolution with extremely randomized trees and LightGBM for protein structural class prediction. Appl Soft Comput. 2023;136: 110064.
68. Chen C, Zhang Q, Ma Q, Yu B. LightGBM-PPI: Predicting protein-protein interactions through LightGBM with multi-information fusion. Chemom Intell Lab Syst. 2019;191:54–64.
69. Song J, Liu G, Jiang J, Zhang P, Liang Y. Prediction of protein–ATP binding residues based on ensemble of deep convolutional neural networks and LightGBM algorithm. Int J Mol Sci. 2021;22:939.
70. Zhang X. Ion channel prediction Using Lightgbm Model. In: 2020 International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE). IEEE; 2020. p. 349–52.
71. Herrera-Bravo J, Herrera Belén L, Farias JG, Beltrán JF. TAP 1.0: a robust immunoinformatic tool for the prediction of tumor T-cell antigens based on AAindex properties. Comput Biol Chem. 2021;91:107452.
72. Chen W, Luo L. Classification of antimicrobial peptide using diversity measure with quadratic discriminant analysis. J Microbiol Methods. 2009;78:94–6.
73. Feng P, Wang Z, Yu X. Predicting antimicrobial peptides by using increment of diversity with quadratic discriminant analysis method. IEEE/ACM Trans Comput Biol Bioinform. 2019;16:1309–12.
74. YongE F, GaoShan K. Identify beta-hairpin motifs with quadratic discriminant algorithm based on the chemical shifts. PLoS ONE. 2015;10: e0139280.
75. Li F, Zhou H. Predicting protein subcellular location based on improved quadratic discriminant. In: 2011 4th International Conference on Biomedical Engineering and Informatics (BMEI). IEEE; 2011. p. 1989–92.
76. Wan C, Jones DT. Protein function prediction is improved by creating synthetic feature samples with generative adversarial networks. Nat Mach Intell. 2020;2:540–50.
77. Wang Y, Luo X, Zou Q. Effector-GAN: prediction of fungal effector proteins based on pretrained deep representation learning methods and generative adversarial networks. Bioinformatics. 2022;38:3541–8.
78. Lin T-T, Sun Y-Y, Wang C-T, Cheng W-C, Lu I-H, Lin C-Y, et al. AI4AVP: an antiviral peptides predictor in deep learning approach with generative adversarial network data augmentation. Bioinform Adv. 2022;2:vbac080.
79. Lee B, Shin MK, Hwang I-W, Jung J, Shim YJ, Kim GW, et al. A deep learning approach with data augmentation to predict novel spider neurotoxic peptides. Int J Mol Sci. 2021;22:12291.

Publisher's Note