Received: 18 March 2025 | Revised: 27 May 2025 | Accepted: 30 May 2025

DOI: 10.1002/pro.70200

RESEARCH ARTICLE

THE PROTEIN SOCIETY WILEY

A large language model for predicting neurotoxic peptides and neurotoxins

Anand Singh Rathore | **Saloni Jain** | **Shubham Choudhury** | **Gajendra P. S. Raghava**

Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India

Correspondence
Gajendra P. S. Raghava, Department of Computational Biology, Indraprastha Institute of Information Technology, Delhi, Okhla Industrial Estate, Phase III (Near Govind Puri Metro Station), New Delhi 110020, India.
Email: raghava@iiitd.ac.in

Funding information
Department of Biotechnology (DBT), Ministry of Science and Technology, India, Grant/Award Number: BT/PR40158/BTIS/137/24/2021; University Grants Commission (UGC); Council of Scientific and Industrial Research (CSIR); Department of Science and Technology (DST-INSPIRE)

Review Editor: Nir Ben-Tal

Abstract

The accurate prediction of neurotoxicity in peptides and proteins is essential for the safety evaluation of therapeutic proteins and genetically modified (GM) organisms. Existing tools, including our earlier method NTxPred, typically use a single predictive model for both neurotoxic peptides and proteins, despite their structural and functional differences. This lack of specialization may lead to suboptimal performance and limited generalizability. To address this, we developed NTxPred2, distinct, specialized models for predicting neurotoxic peptides and neurotoxins (proteins). Our curated datasets include 877 neurotoxic and 877 non-toxic peptides, and 775 neurotoxic and 775 non-toxic proteins. Certain residues, like cysteine, are prevalent in both but in different magnitudes. Using composition and binary profiles, our machine-learning models achieved an area under the curve (AUC) of 0.97 for peptides and 0.85 for proteins, improving to 0.89 with evolutionary information. Models using protein embeddings reached 0.96 AUC for peptides and 0.94 for proteins, while protein language models achieved 0.98 (esm2-t30) and 0.91 (esm2-t6). All models were validated via five-fold cross-validation, and the final models were evaluated on an independent dataset. We further assessed protein models on the peptide dataset and vice versa, highlighting the necessity of separate models. The proposed models outperform existing methods on independent datasets that are not used for training. Our neurotoxicity prediction models will aid in the safety assessment of GM foods and therapeutic proteins by minimizing the need for animal testing. To support the scientific community, we developed a standalone software and web server NTxPred2 for predicting and scanning neurotoxins (https://webs.iiitd.edu.in/raghava/ntxpred2/, https://github.com/raghavagps/ntxpred2/).

KEYWORDS
embeddings, machine-learning techniques, neurotoxic peptides, neurotoxins, protein language models, WHO guidelines

1 | INTRODUCTION

Over the past 150 years, significant advancements have been made in utilizing proteins for therapeutic applications. A major breakthrough occurred in 1891 with the development of an antibody-based treatment for diphtheria, which led to the awarding of the first Nobel Prize in Medicine (Kaufmann, 2017). This was followed by the approval of insulin, a peptide-based drug, in 1922, providing a life-saving treatment for diabetes. Initially, insulin production relied on animal-derived sources due to limitations in human insulin production. However, the advent of recombinant DNA technology revolutionized this process. In 1982, insulin was successfully produced in *Escherichia coli* using recombinant techniques, making it the first food and

drug administration (FDA)-approved recombinant protein therapy (Johnson, 1983). Today, peptide and protein-based drugs constitute over 10% of the pharmaceutical market, with their share expected to grow in the coming years. The Therapeutic Peptide and Protein Database (THPdb) serves as a comprehensive resource for US FDA-approved protein and peptide therapeutics (Usmani et al., 2017). The first version, published in 2017, documented 239 approved therapeutic peptides and proteins. The latest version, THPdb2, released in 2024, has expanded to include 894 FDA-approved protein-based therapeutics, highlighting the importance of peptides and proteins in modern medicine (Jain et al., 2024). However, this expansion has heightened the need to ensure the safety of novel proteins, especially concerning potential toxicity, a key requirement in both pharmaceuticals and genetically modified (GM) foods (World Health Organization [WHO], 2001).

The WHO has published guidelines for assessing the safety of genetically engineered food proteins, including protocols for identifying toxicity (Oh et al., 2009). While in vivo animal testing is the regulatory gold standard, these methods are expensive, time-consuming, and raise ethical concerns. Therefore, there is an urgent need for computational (in silico) approaches that can accurately predict protein toxicity, particularly neurotoxicity, which affects the central and peripheral nervous systems (PNS) and may be caused by a wide range of molecules such as scorpion toxins, botulinum toxins, and diphtheria toxin (Escoubas et al., 2000; Latham et al., 2023; Pirazzini et al., 2017; Prasad & Rai, 2018; Romey et al., 1976). One of the major challenges in these guidelines is assessing the potential adverse effects, particularly toxicities, of newly discovered proteins. To support experimental researchers, numerous in silico tools have been developed to screen proteins that have any type of toxicity like cytotoxicity, immunotoxicity, hemotoxicity, and neurotoxicity (Dhall et al., 2021; Nagpal et al., 2018; Rathore, Choudhury, et al., 2024; Rathore, Kumar, et al., 2025; Saha & Raghava, 2007a).

In the last decade, a number of methods have been developed for predicting the cytotoxicity of peptides and proteins, including ToxinPred (Gupta et al., 2015; Rathore, Choudhury, et al., 2024; Sharma et al., 2022), ToxClassifier (Gacesa et al., 2016), TOXIFY (Cole & Brewer, 2019), ToxDL (Pan et al., 2021), Attention-based Self-Ensembling (Wei et al., 2021), ToxIBTL (Wei et al., 2022), ToxMVA (Shi et al., 2022), CSM-toxin (Morozov et al., 2023), VISHpred (Mall et al., 2024), and MultiToxPred 1.0 (Beltrán et al., 2024). Similarly, a number of in silico tools have been developed for predicting the allergenicity or immunotoxicity of proteins, including IL6pred (Dhall et al., 2021) and AlgPred (Sharma et al., 2021). Despite these advancements, limited efforts have been made to predict the neurotoxicity of proteins and peptides. Neurotoxicity is a type of toxicity that causes damage to the structure and function of the brain, spinal cord, and PNSs. Neurotoxicity is a significant concern in the field of protein therapeutics and in GM foods. In order to address this issue, Saha & Raghava (2007b) developed a method, NTxPred, in 2007, for identifying neurotoxins based on their functional and source characteristics. This tool has found applications in various forms in the development of therapeutic peptides (Su et al., 2014; Xu et al., 2017). Guang et al. (2010) developed an support vector machine (SVM)-based method using evolutionary information for predicting neurotoxins (Guang et al., 2010). In addition, numerous methods have been developed for the prediction and characterization of spider neurotoxins (Grishin, 1999; Koua & Kuhn-Nentwig, 2017). Neurotoxins are categorized into presynaptic and postsynaptic neurotoxins based on their synaptic location and mechanism of action. A range of machine-learning (ML) techniques has been developed to classify presynaptic and postsynaptic neurotoxins (Chaohong, 2012; Huo et al., 2017; Lee et al., 2021; Li et al., 2020; Mei & Zhao, 2018; Tang et al., 2017; Wan et al., 2023; Yang & Li, 2009). In Table S1, we provide a comprehensive overview of methodologies employed for predicting specific toxicity types in therapeutic peptides and proteins, including hemotoxicity (Ansari & White, 2023; Chaudhary et al., 2016; Guntuboina et al., 2023; Kumar et al., 2020; Rathore, Kumar, et al., 2024; Salem et al., 2022; Timmons & Hewage, 2020; Win et al., 2017).

In this study, we present NTxPred2, an advanced and comprehensive platform for neurotoxicity prediction that introduces several innovations over existing approaches. First, it employs dedicated ML models for peptides and proteins, in contrast to older methods that applied unified models irrespective of molecular type. Second, it utilizes significantly larger and more curated datasets derived from Swiss-Prot, enhancing the robustness and generalizability of predictions. Third, NTxPred2 integrates state-of-the-art (SOTA) protein language models (PLMs) such as Evolutionary Scale Modeling (ESM) 2 and ProtBERT (Devlin et al., 2018), which capture deep contextual, structural, and evolutionary information directly from sequence data. Fourth, we conduct comprehensive cross-dataset validation, which empirically demonstrates the superior performance of separate models compared to hybrid or subtype-specific approaches. While these techniques result in improved predictive accuracy and biological relevance, we acknowledge certain potential limitations, including increased computational complexity due to the use of large PLMs, reliance on high-quality annotated datasets, and the reduced interpretability associated with deep learning (DL)-based models. Nonetheless, these challenges are being actively

addressed by the community, and we believe NTxPred2 provides a strong foundation for future improvements. Overall, NTxPred2 offers a scalable, accurate, and animal-free solution for neurotoxicity prediction, contributing meaningfully to the safe design of therapeutic proteins and the regulatory assessment of engineered food products.

2 | METHODS

2.1 | Data collection and pre-processing

A comprehensive dataset of experimentally validated neurotoxic and non-toxic peptides was curated from the Swiss-Prot. To construct the positive dataset, the keyword "neurotoxin" was queried with the "reviewed_true" filter, retrieving an initial set of 5173 putative neurotoxic peptide sequences. For the negative dataset, sequences were obtained using the keyword "NOT toxin NOT neurotoxin" with the "reviewed_true" filter, yielding an initial set of 566,303 non-toxic peptide sequences. This information has been used to create the following three datasets:

1. Peptide dataset: In this case, we only select peptides having ≤51 amino acids. In addition, we removed peptides containing non-canonical amino acids and redundant peptides. Finally, we got 877 unique neurotoxic peptides called positive dataset of peptides. The same exclusion criteria were applied to non-toxic peptides, yielding 8437 unique non-toxic peptides. To mitigate the class imbalance inherent in the dataset, the CD-HIT software (Li & Godzik, 2006) was employed to cluster the non-toxic peptide sequences at a 60% sequence identity threshold. This clustering significantly reduced the number of redundant sequences within the negative dataset. Resulting in a negative dataset of 877 non-toxic peptides with length distributions matching the positive dataset. Although CD-HIT 40 was also considered, it did not yield a sufficient number of non-toxic peptides to maintain length balance. The minimum peptide length included in the dataset was seven amino acids.
2. Protein dataset: In case of protein dataset, we only select proteins having ≥51 amino acids. The same filtering criteria were applied, removing sequences with non-canonical amino acids, duplicates, and redundant entries, resulting in 3755 neurotoxic and 457,745 non-toxic protein sequences. To further refine the dataset and reduce redundancy, CD-HIT 40 was applied, yielding 775 non-redundant neurotoxic protein sequences. To create a balanced dataset, 775 non-toxic protein sequences of comparable lengths were selected.

To ensure a rigorous and unbiased model evaluation, standard data partitioning strategies were employed. Both the peptide and protein datasets were independently divided into two subsets. The training set (80% of the data) is used for model development, and the independent dataset (20% of the data) is reserved strictly for final model evaluation. It is important that the independent dataset remains completely unseen during training, testing, model selection, and hyperparameter tuning to prevent data leakage and ensure an accurate assessment of model generalizability.

3. Combined dataset: To evaluate the model's ability to generalize across peptide and protein sequences, a combined dataset was created. Instead of simply merging all peptide and protein sequences, this dataset was structured such that the training datasets of peptides and proteins were merged to form a combined training dataset. The independent datasets of peptides and proteins were merged to form a combined independent dataset for model evaluation. This approach ensures a fair and robust model assessment by evaluating performance on a truly independent dataset containing both molecular types.

The complete data curation process is illustrated in Figure 1.

2.2 | Compositional and positional analysis

To gain an initial understanding of the distinguishing features of experimentally validated neurotoxic and non-toxic sequences, we conducted a comprehensive compositional and positional analysis. Our approach involved multiple analytical techniques to assess sequence-based properties across all three datasets. The first stage of analysis focused on amino acid composition within each class. For both neurotoxic and non-toxic sequences, we calculated key statistical metrics, including the mean, median, and standard deviation of amino acid frequencies. To assess the statistical significance of differences between the two classes, we employed independent t-tests (*scipy.stats*) and controlled for false discovery rates using the Benjamini-Hochberg procedure (*statsmodels.stats.multitest*). These statistical corrections ensured robustness by minimizing type I errors. Beyond the neurotoxic and non-toxic classifications, we extended our analysis to compare amino acid compositions across other toxicity categories, including hemotoxicity (Rathore, Kumar, et al., 2024) and cytotoxicity (Rathore, Choudhury, et al., 2024), as well as a general genome-wide amino acid composition. This broader comparison provided insights into the shared and unique sequence characteristics of different toxicity classes. To further investigate sequence-specific preferences, we applied the Two Sample Logo (TSL) method, enabling the

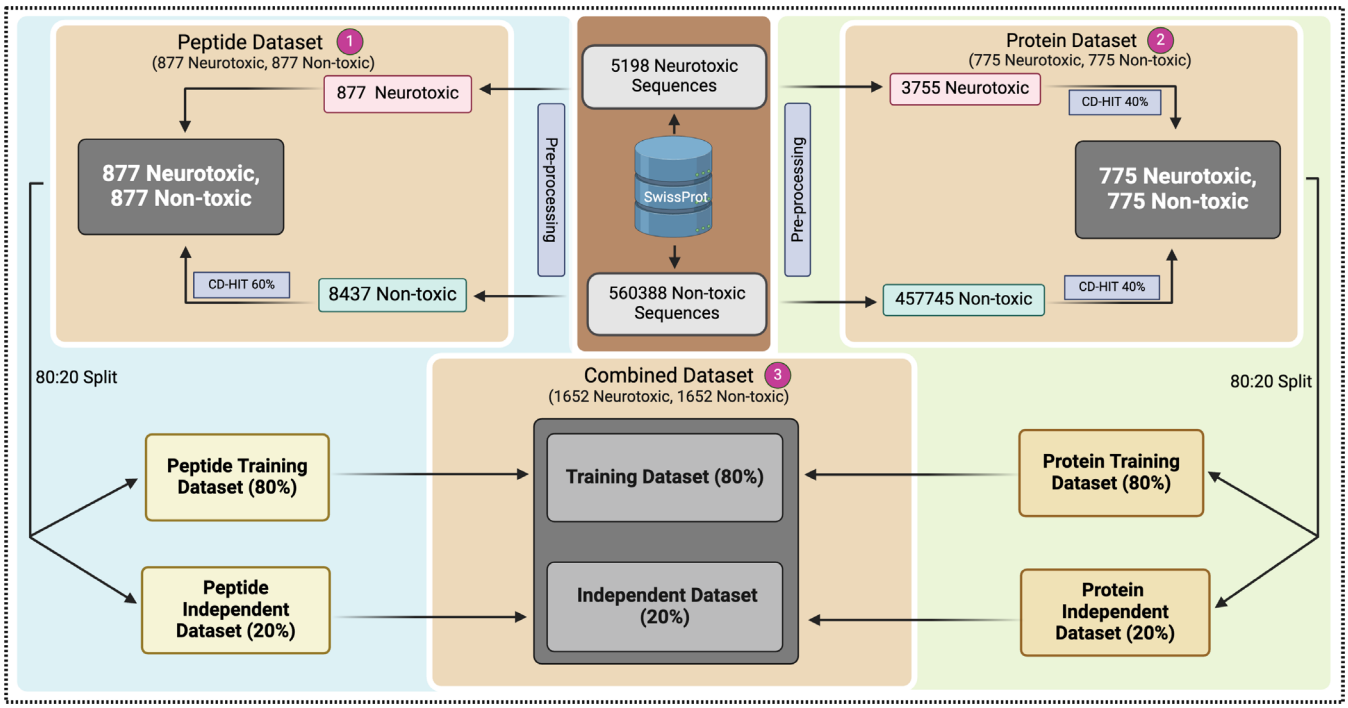

FIGURE 1 A schematic representation of NTxPred2 dataset creation.

identification of distinct positional preferences for amino acid residues. This technique highlighted residues that were preferentially enriched or depleted at specific positions within peptide sequences, offering deeper insights into sequences associated with neurotoxicity.

2.3 | Correlation between toxicities

To explore the relationships between different toxicity types, we conducted a correlation analysis based on amino acid composition patterns. This analysis aimed to identify potential overlaps or distinctions in compositional features among various toxicity classes, including neurotoxicity, hemotoxicity, and cytotoxicity. By systematically comparing the amino acid composition profiles of these toxicity categories, we sought to determine whether certain residues were consistently associated with toxicity across multiple classes or if unique compositional signatures distinguished one toxicity type from another. This investigation provides valuable insights into the underlying molecular characteristics that drive toxicity and helps in understanding the extent to which different toxic properties share common sequence determinants.

2.4 | Composition and physicochemical property-based prediction

In this analysis, we classified neurotoxic and non-toxic sequences based on the composition of each of the 20 standard amino acids and different types of physicochemical properties (PCP) like polarity, charge, etc. For each feature, we first calculated the mean composition for both the neurotoxic and non-toxic groups. A threshold was then determined as the midpoint between the two means. Sequences were classified according to whether their composition exceeded this threshold. Specifically, if the neurotoxic group had a higher mean composition for a particular amino acid, sequences with values above the threshold were classified as neurotoxic, and vice versa. This classification was performed across all three datasets. To assess the classification performance, we calculated the accuracy and area under the curve (AUC) for each amino acid. The results, including the thresholds, mean compositions, accuracy, and AUC, were compiled into a comprehensive table for further analysis.

2.5 | Feature extraction

In developing a sequence-based predictor for peptide and protein properties, it is crucial to effectively represent sequences to capture their biological functions. We extracted a variety of informative features, including binary profiling, position-specific scoring matrix (PSSM) profiling, composition-based features, physicochemical features, and word embeddings using large language models (LLMs). These features were computed across all three datasets to improve the predictive accuracy and generalizability of the model.

2.5.1 | Binary profiling

To assess the importance of specific residues at the N- and C-terminal positions, we incorporated both frequency and positional information of residues. We developed a method using binary profiles of peptides, generating these profiles for each sequence across all three datasets. Since the minimum sequence length was seven, we extracted the first seven amino acids (NT7) and the last seven amino acids (CT7) from each sequence. Each amino acid in these regions was then one-hot encoded, creating a 14-dimensional binary vector that indicated the presence or absence of the 20 standard amino acids at each position. As a result, each sequence was represented by a 280-dimensional feature vector, capturing the amino acid composition of its terminal regions. These binary profiles were integrated into the dataset, providing valuable features for subsequent ML analyses. This approach has been successfully applied in several existing methods (Lata et al., 2007; Sharma et al., 2013).

2.5.2 | Evolutionary information-based features

Evolutionary information-based features capture sequence conservation and variability, providing insights into functional importance (Kumar et al., 2007; Tang et al., 2021). In this study, we utilized the PSSM, obtained through Position-Specific Iterated- Basic Local Alignment Search Tool (PSI-BLAST) (Altschul et al., 1997), to extract evolutionary information. The PSSM matrix quantifies the evolutionary conservation of each amino acid position in a sequence by comparing it to a set of homologous sequences. We employed multiple modules from the *POSSUM* package (Wang et al., 2017):

1. *aac_pssm*: This module calculates the average activation values for each amino acid position across the sequence, providing a 20-dimensional (20D) feature vector. It captures how strongly each amino acid type is represented at each position, reflecting sequence-specific characteristics.
2. *pssm_composition*: This method treats the entire PSSM matrix as a single vector and computes its average, encoding global properties of the matrix. It results in a 400-dimensional (400D) feature vector that reflects the overall evolutionary profile of the sequence.
3. *medp_pssm* (Mean Evolutionary Difference Profile): This combines features from two approaches-Evolutionary Difference Profile (EDP) and Evolutionary Difference PSSM (EEDP). EDP computes differences between PSSM values for paired residues across rows and columns, highlighting evolutionary variation, while EEDP focuses on evolutionary differences between PSSM values for paired positions, distinguishing conserved from variable regions. The *mepd_pssm* method yields a 420-dimensional (420D) feature vector.

For this study, we generated PSSM matrices for each sequence using Swiss-Prot as the reference database. However, two sequences were excluded from the analysis due to the absence of homologous sequences, preventing PSSM generation. Additionally, we applied these features only to the protein dataset, as PSSM-based methods are generally unreliable for short sequences. Specifically, sequences shorter than 50 residues lack sufficient evolutionary context, leading to poorly constructed PSSM matrices with high variability and low statistical confidence (Liu et al., 2014; Marks et al., 2012; Refahi et al., 2020).

2.5.3 | Composition-based features

For composition-based feature extraction, we used the standalone tool of *Pfeature*, which provided a wide array of descriptors. In total, we collected 9190 features for each peptide, encompassing 18 distinct feature types. These types include Amino Acid Composition (AAC), Dipeptide Composition (DPC), Atom Type Composition, Bond Type Composition, and various PCP, among others. Detailed descriptions of each feature and the corresponding vector lengths are presented in Table S2.

In addition to Pfeature, we also utilized the modlAMP library. We calculated 56 PCPs for each sequence based on their primary sequences, utilizing the *modlAMP* 4.3.0 package in Python (Müller et al., 2017). This included 47 sequence-specific descriptors and nine global descriptors. A comprehensive list of these properties can be found in Table S3. The resulting data was organized into datasets with peptide sequences as rows and their corresponding properties as columns. This approach has been employed in various studies focused on toxicity prediction (Chu et al., 2022).

2.5.4 | Word embeddings

Recent advancements in natural language processing (NLP) have led to the development of PLMs, which utilize amino acids and their combinations (doublets or triplets) as tokens. These models generate fixed-size vectors, known as embeddings, which encapsulate specific peptide sequences. These protein embeddings are crucial for various applications, including structure prediction, novel sequence generation, and protein classification (Bepler & Berger, 2021; Madani et al., 2023). In our research, we utilized two prominent LLMs: ESM-2 (Lin et al., 2022) and ProtBERT. These PLMs are

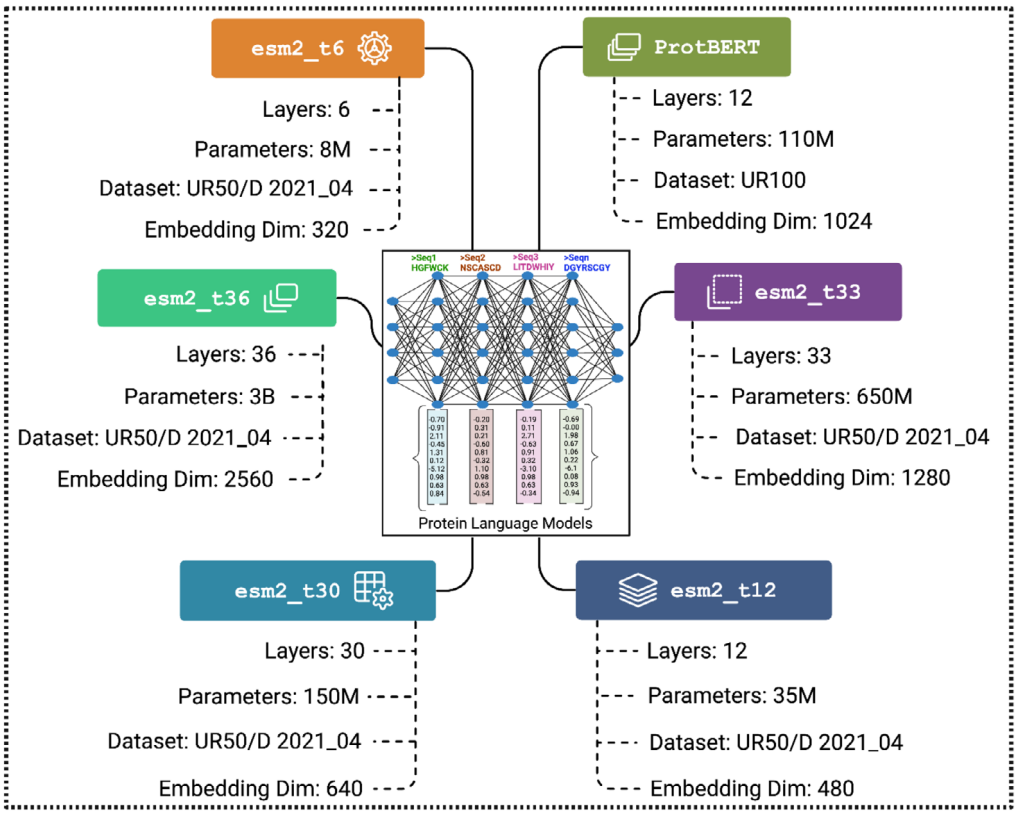

FIGURE 2 Shows different protein language models implemented in this study, including the parameters of the models.

recognized as a SOTA model for predicting various protein properties directly from individual sequences (Brandes et al., 2022; Mall et al., 2024). ProtBERT is an extension of the bidirectional encoder representations from transformers (BERT) model, pre-trained on extensive protein sequence datasets using a self-supervised approach (Elnaggar et al., 2022). ESM-2, a transformer-based PLM, was trained on sequences from the UniRef protein sequence database using a masked language modeling objective. Figure 2 showcases different PLMs used to extract the embeddings, each varying in size and complexity. Several checkpoints of the ESM-2 model, which vary in size, can be found on Hugging Face (Wolf et al., 2019). Generally, larger models tend to achieve marginally improved accuracy; however, they also demand considerably more memory and longer training durations. Key characteristics such as the number of layers, the total number of parameters in which they are trained, and the embedding dimension are displayed for each model. This visualization provides a concise overview of these PLMs, highlighting their key specifications.

2.6 | Machine-learning-based classifiers

ML algorithms have been extensively utilized to distinguish neurotoxic peptides and proteins from non-toxic ones. In this study, we explored a range of classifiers, including Logistic Regression (LR), Ridge, Lasso, ElasticNet, Random Forest (RF), Extra Trees (ET), Gaussian Naïve Bayes (GNB), Decision Trees (DT), Multi-Layer Perceptron (MLP), eXtreme Gradient Boosting (XGB), AdaBoost (AdB), and Support Vector Classifier (SVC) with various kernel functions (linear, radial basis function, polynomial, and sigmoid). Each model was systematically optimized by fine-tuning multiple hyperparameters to achieve optimal classification performance. The complete workflow of the NTxPred2 system is depicted in Figure 3.

2.7 | Feature selection

Feature selection is a crucial step in ML that enhances model performance by identifying and retaining only the most informative features while eliminating redundant or irrelevant ones. This process not only improves prediction accuracy but also reduces computational complexity and enhances interpretability. In this study, we employed multiple feature selection techniques to optimize ML models. These methods include Variance Threshold (VTh), which removes low-variance features; SelectKBest, which ranks features based on statistical tests; and SVC with L1-based feature selection (SVC-L1), which leverages sparsity to identify important features. Additionally, we utilized tree-based feature selection (ET) to assess feature

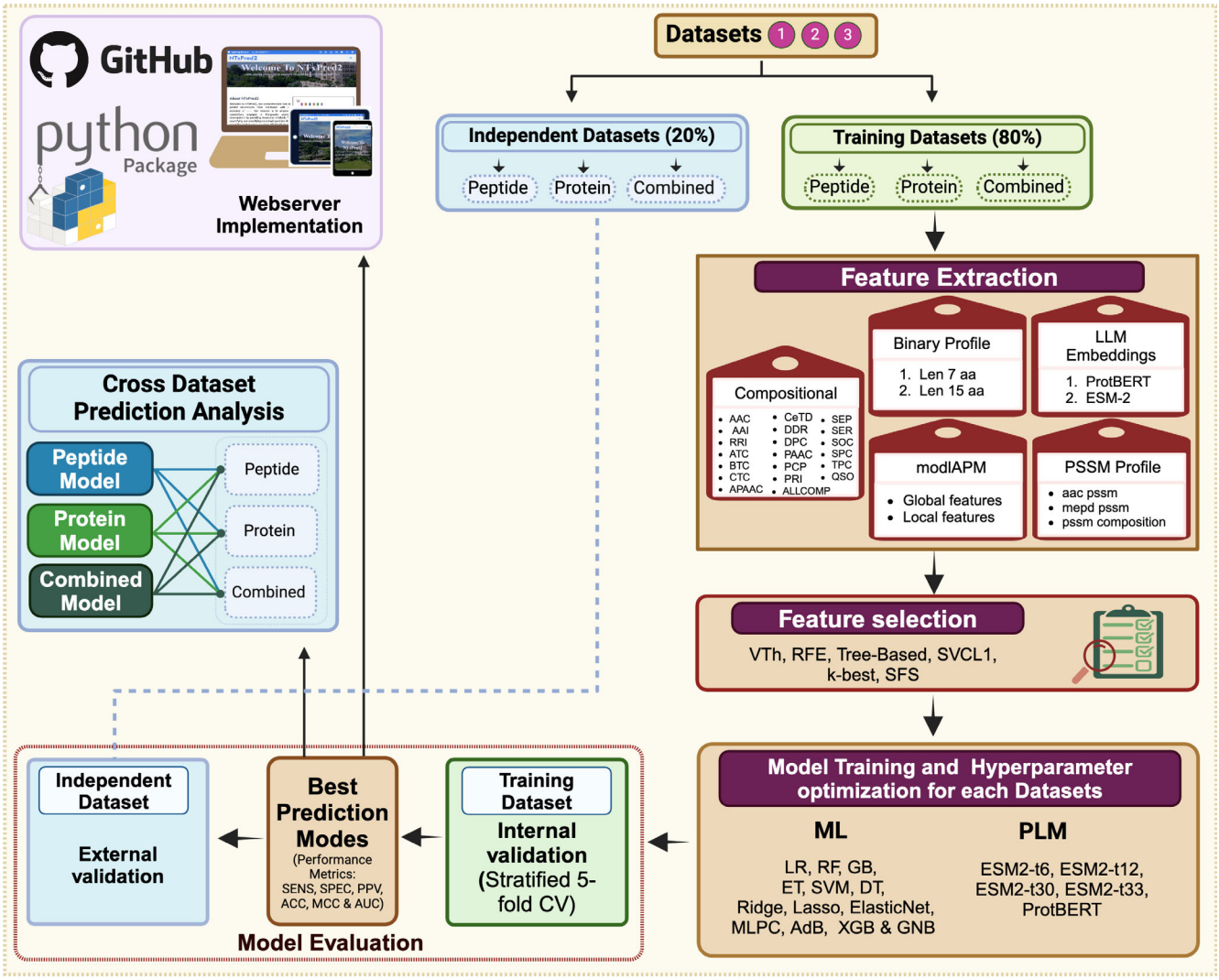

FIGURE 3 Flowchart illustrating the complete architecture of NTxPred2, including cross-validation techniques, feature extraction methods, and prediction models. AAC, Amino Acid Composition; ACC, accuracy; AdB, AdaBoost; ALLCOMP, All Composition features extracted using *Pfeature* tool; AUC, area under the curve; DPC, Dipeptide Composition; DT, Decision Trees; ESM, Evolutionary Scale Modeling; ET, Extra Trees; GNB, Gaussian Naïve Bayes; LLM, large language models; LR, Logistic Regression; MCC, Matthews correlation coefficient; ML, machine-learning; PCP, physicochemical properties; PLM, protein language models; PPV, positive predictive value; PSSM, position-specific scoring matrix; SENS, sensitivity; SFS, sequential forward selection; SPEC, specificity; SVCL1, Support Vector Classifier with L1-based feature selection; SVM, support vector machine; RF, Random Forest.

importance using ensemble learning, recursive feature elimination (RFE) to iteratively refine the most relevant subset, and sequential forward selection (SFS) to progressively add features that contribute most to the model's performance. By integrating these diverse approaches, we ensured a robust and comprehensive feature selection process, ultimately enhancing the efficiency and reliability of our ML models.

2.8 | Protein language model as classifiers

Utilizing classical ML models laid the groundwork for our research, but we advanced our approach by incorporating pre-trained language models (PLMs). These computational frameworks leverage sophisticated NLP techniques to analyze the intricate details of protein structures, functions, and interactions. In particular, we employed several models from the ESM series, specifically esm2-t33, esm2-t30, esm2-t12, and esm2-t6. These models are pre-trained on extensive protein sequence datasets and excel in various tasks such as structure prediction and variant effect analysis. The ESM series has demonstrated remarkable capabilities in capturing the evolutionary relationships and functional aspects of proteins due to their architecture and training methodology (Lin et al., 2022). In addition to the ESM models, we also integrated ProtBERT into our framework. ProtBERT is a protein-specific model

derived from the BERT architecture and is pre-trained on a vast corpus of protein sequences. This model is particularly adept at understanding the contextual relationships within protein sequences, which enhances its performance in tasks related to protein classification and function prediction. The combination of these models significantly improved the accuracy and reliability of our classification system for neurotoxic and non-toxic peptides (Liu & Tian, 2023; Zhang et al., 2024).

To fine-tune these models for our task, we optimized the hyperparameters of these models. For ProtBERT, sequences were tokenized with maximum length sequences and a batch size of 32. The model was fine-tuned across multiple epochs (6, 12, and 32) using the AdamW optimizer with a learning rate of 1×10^{-5}. Similarly, for ESM, sequences were tokenized, and the *esmForSequenceClassification* model was trained for the same range of epochs using Adam with an identical learning rate. Cross-entropy loss was utilized, and both models employed truncation, padding, and gradient updates to enhance classification performance for neurotoxic and non-toxic peptides.

Overall, the integration of PLMs like ESM2 and ProtBERT into our research not only enhanced our classification model's performance but also underscored the importance of leveraging SOTA computational techniques in understanding complex biological phenomena. These advancements pave the way for more accurate predictions and deeper insights into protein functionality and interactions, ultimately contributing to fields such as drug discovery and synthetic biology.

2.9 | Cross-validation and performance metrics

As outlined in Section 2.1, for model training and evaluation, datasets were split into 80% training and 20% independent test sets. The test set remained unused during training or tuning. Peptide and protein training sets were merged into a combined training dataset, and their test sets were merged into a combined independent dataset. To prevent data leakage and ensure robust model performance estimates, a stratified five-fold cross-validation procedure was implemented exclusively within the training set. Stratified cross-validation maintains the original class proportions in each fold, ensuring that the characteristics of the original dataset are preserved during the evaluation process. This approach provides a more reliable estimate of model performance compared to simple *k*-fold cross-validation, especially when dealing with imbalanced datasets. The effectiveness of various ML models was evaluated using both threshold-dependent and threshold-independent metrics. The area AUC serves as a threshold-independent measure, while sensitivity, specificity, accuracy, and the Matthews correlation coefficient (MCC) are classified as threshold-dependent metrics. These evaluation criteria have been thoroughly documented in earlier research studies (Rathore, Choudhury, et al., 2024; Saha & Raghava, 2007b).

2.10 | Benchmarking and statistical comparison of tools

To rigorously assess the performance of NTxPred2, an enhanced version of NTxPred incorporating fine-tuned PLMs and novel features, we conducted a comprehensive comparative analysis against existing predictive tools. We utilized an independent, unseen dataset and employed a 10-fold stratified cross-validation strategy to ensure robust evaluation and maintain original class distributions across folds. We evaluated different predictive tools by applying the Friedman test, a non-parametric statistical method suitable for repeated-measures comparisons of multiple methods. Each tool generated prediction scores for identical peptide sequences, with ground-truth binary labels (0 or 1) provided. To ensure robustness, we performed 10-fold stratified cross-validation using scikit-learn's *StratifiedKFold* (v1.3.0), preserving class distribution in each fold. For each fold, we calculated the area under the receiver operating characteristic curve (AUROC) Basic Local Alignment Search Tool for all tools using sklearn.metrics' *roc_auc_score*. The Friedman test, implemented via *scipy.stats*' *friedmanchisquare* (v1.11.0), was applied to this matrix, with a *p-value* <0.05 indicating statistical significance. Significant results prompted a post-hoc Nemenyi test using *scikit-posthocs* (v0.7.0) to identify pairwise differences in tool performance. This approach ensures a fair, robust, and statistically valid comparison across identical data splits.

3 | RESULTS

3.1 | Compositional analysis of different toxicities

Firstly, we compare the average amino acid composition of neurotoxic and non-toxic sequences. The amino acid profiles are analyzed across all three datasets, and a comparative assessment of their average amino acid compositions was conducted (Figure S1). Notably, the occurrence of cysteine (C), a polar and uncharged amino acid, was significantly greater in the neurotoxin sequences compared to those of non-toxins. This observation aligns with trends identified in previous research (Saha & Raghava, 2007b). Statistical analysis identified significant differences in amino acid composition between neurotoxic and non-toxic peptides. Neurotoxic peptides exhibited a significantly higher abundance of C,

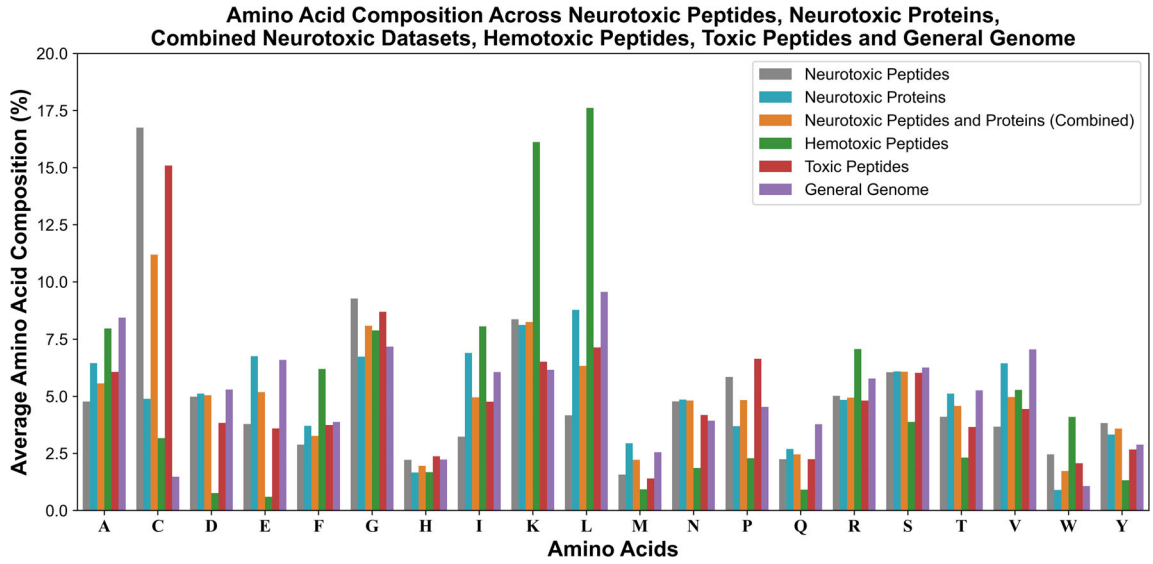

FIGURE 4 Amino acid composition of different types of toxicity-associated proteins and peptides that includes neurotoxic peptides/proteins, hemotoxic, cytotoxic peptides, and the human genome.

glycine, and lysine (K), with a corrected *p*-value of 0.001. Conversely, non-toxic peptides demonstrated a significantly higher composition of alanine (A), isoleucine (I), leucine (L), and glutamine (Q) across all three datasets.

In the past, number of methods has been developed for predicting different types of toxicities. We compare the composition profiles of different types of toxicities to understand the preference of different amino acids. In Figure 4, we show the composition profile of six distinct datasets: neurotoxic peptides, neurotoxic proteins, combined neurotoxic datasets (peptides + proteins), hemotoxic peptides, toxic peptides (general toxicity), and the general genome. This aims to identify amino acid compositional biases that distinguish functionally specialized toxic biomolecules from the broader genomic background. Neurotoxic peptides and proteins are analyzed separately and collectively to evaluate whether their amino acid profiles converge or diverge, while hemotoxic and general toxic peptides are included to assess if neurotoxicity correlates with unique compositional signatures. The inclusion of the general genome provides a baseline for distinguishing natural amino acid distributions from those enriched in toxic entities. Key observations likely include elevated frequencies of hydrophobic or charged residues (e.g., C and tryptophan [W]) in neurotoxic and hemotoxic peptides, which are often critical for membrane interactions or receptor targeting. Conversely, the general genome may show higher proportions of smaller, residues (e.g., A and glycine). This comparative visualization highlights how evolutionary and functional pressures shape amino acid utilization in toxic molecules, offering insights into potential structural or mechanistic hallmarks of toxicity.

3.2 | Correlation between toxicities

The amino acid composition correlation matrix reveals distinct relationships among various toxic components and the general genome (Table 1). It was observed that neurotoxic peptides and neurotoxins have a poor correlation that further supports the development of separate models for neurotoxic peptides and neurotoxins. Neurotoxic peptides show a strong positive correlation with the overall toxin profile ($r \approx 0.94$), indicating their significant role in neurotoxicity. Neurotoxic peptides have a negligible correlation with the general genome ($r \approx -0.03$), suggesting a high degree of specialization, diverging from the general genome. Neurotoxic proteins exhibit moderate correlations with the general genome ($r \approx 0.86$), reflecting a balance between maintaining essential physiological functions and contributing to neurotoxic effects. Hemotoxins display a moderate correlation with neurotoxic proteins ($r \approx 0.68$) and the general genome ($r \approx 0.61$), but a low correlation with neurotoxic peptides ($r \approx 0.18$), suggesting that hemotoxins and neurotoxins may have distinct evolutionary pathways and functional roles. These findings align with research indicating that snake venoms are complex mixtures of proteins and peptides, each evolving to target specific physiological systems, such as the nervous or circulatory systems, in their prey (Osipov & Utkin, 2023).

3.3 | Positional analysis

We generated TSLs for three datasets to analyze amino acid preferences in neurotoxic and non-toxic sequences, as shown in Figure 5. The TSLs highlight

TABLE 1 Confusion matrix shows a correlation between different types of toxicities and the general genome. Amino acid residues significantly higher (corrected *p*-value <0.05) in each dataset are highlighted using their single-letter codes, with color coding applied to distinguish between datasets.

	Neurotoxic peptides	Neurotoxic proteins	Neurotoxic combined	Hemotoxins	Toxins	General
Neurotoxic peptides	1	0.325 C, G, H, P, W, Y	0.916 C, G, H, P, W	0.178 C, D, E, G, H, M, N, P, Q, S, T, Y	0.942 C, D, G, K, N, T, W, Y	−0.027 C, G, K, N, P, W, Y
Neurotoxic proteins	0.325 A, E, F, I, L, M, Q, T, V	1	0.678 A, E, F, I, L, M, Q, T, V	0.677 C, D, E, M, N, P, Q, S, T, V, Y	0.454 A, D, E, I, K, L, M, N, Q, T, V, Y	0.862 C, I, K, M, N, Y
Neurotoxic combined	0.916 A, E, F, I, L, M, T, V	0.678 C, G, H, P, W, Y	1	0.426 C, D, E, M, N, P, Q, S, T, Y	0.925 D, E, K, M, N, Q, T, V, Y	0.345 C, G, K, N, P, W, Y
Hemotoxins	0.178 A, F, I, K, L, R, V, W	0.677 A, F, G, I, K, L, R, W	0.426 A, F, I, K, L, R, W	1	0.337 A, F, I, K, L, R, V, W	0.610 C, F, G, I, K, L, R, W
Toxins	0.942 A, F, I, L, P, V	0.454 C, G, H, P, W	0.925 A, C, F, G, H, L, P, W	0.337 C, D, E, G, H, M, N, P, Q, S, T, Y	1	0.167 C, G, H, K, N, P, W
General	-0.027 A, D, E, F, I, L, M, Q, R, T, V	0.862 A, D, F, G, H, L, P, Q, R, V, W	0.345 A, D, E, F, H, I, L, M, Q, R, S, T, V	0.610 D, E, H, M, N, P, Q, S, T, V, Y	0.167 A, D, E, F, I, L, M, Q, R, S, T, V, Y	1

the positional significance of amino acid residues at a *p*-value threshold of 0.05. In neurotoxic peptides, C is predominantly favored across all positions except the first, indicating its critical role at both N- and C-termini (Figure 5a). Additionally, K and W are more prevalent in the C-terminal region, suggesting a preference for these residues in this segment. Non-toxic peptides predominantly feature methionine (M) at the first position and L at the C-terminus. In contrast, neurotoxic proteins show a preference for C residues at positions 8–11 of the C-terminal region (Figure 5b). Residues like K, L, and I are more favored at the N-terminal. The combined dataset exhibits a pattern similar to the peptide dataset (Figure 5c). These findings indicate distinct amino acid preferences between protein and peptide datasets, highlighting the specialized roles of specific residues in neurotoxic functions.

3.4 | Amino acid composition-based prediction

In this analysis, we evaluated the predictive power of AAC for distinguishing between neurotoxic and non-toxic sequences by examining three datasets: peptide, protein, and a combined set. In Table S4, results reveal significant differences in the amino acid profiles; for instance, C consistently emerged as a key differentiator. C shows markedly different mean values between neurotoxic and non-toxic peptides compared to proteins, suggesting that peptides have a unique compositional profile that may confer specific functional attributes related to neurotoxicity. In the peptide dataset, the dramatic contrast in the composition of C with neurotoxic sequences exhibiting much higher values than non-toxic ones results in high accuracy (0.867) and AUC (0.901), whereas the protein dataset shows a less pronounced difference, with lower accuracy and AUC values. Similar trends are observed for other amino acids like L, where the separation between neurotoxic and non-toxic groups differs significantly between peptides and proteins. These compositional disparities underline the biological and structural differences inherent to peptides versus full-length proteins. Consequently, this suggests the development of separate predictive models for peptides and proteins, as a unified model may fail to capture the nuanced variations in amino acid distributions that are critical for accurately discriminating neurotoxins from non-toxins. Overall, this compositional analysis highlights the utility of amino acid profiles as discriminative features for the prediction of neurotoxicity, offering valuable insights that may aid in understanding the molecular determinants of neurotoxic activity. A detailed result is present in Table S4.

3.5 | Physicochemical property-based prediction

Physicochemical property-based prediction analysis reveals that the predictive utility of PCPs differs

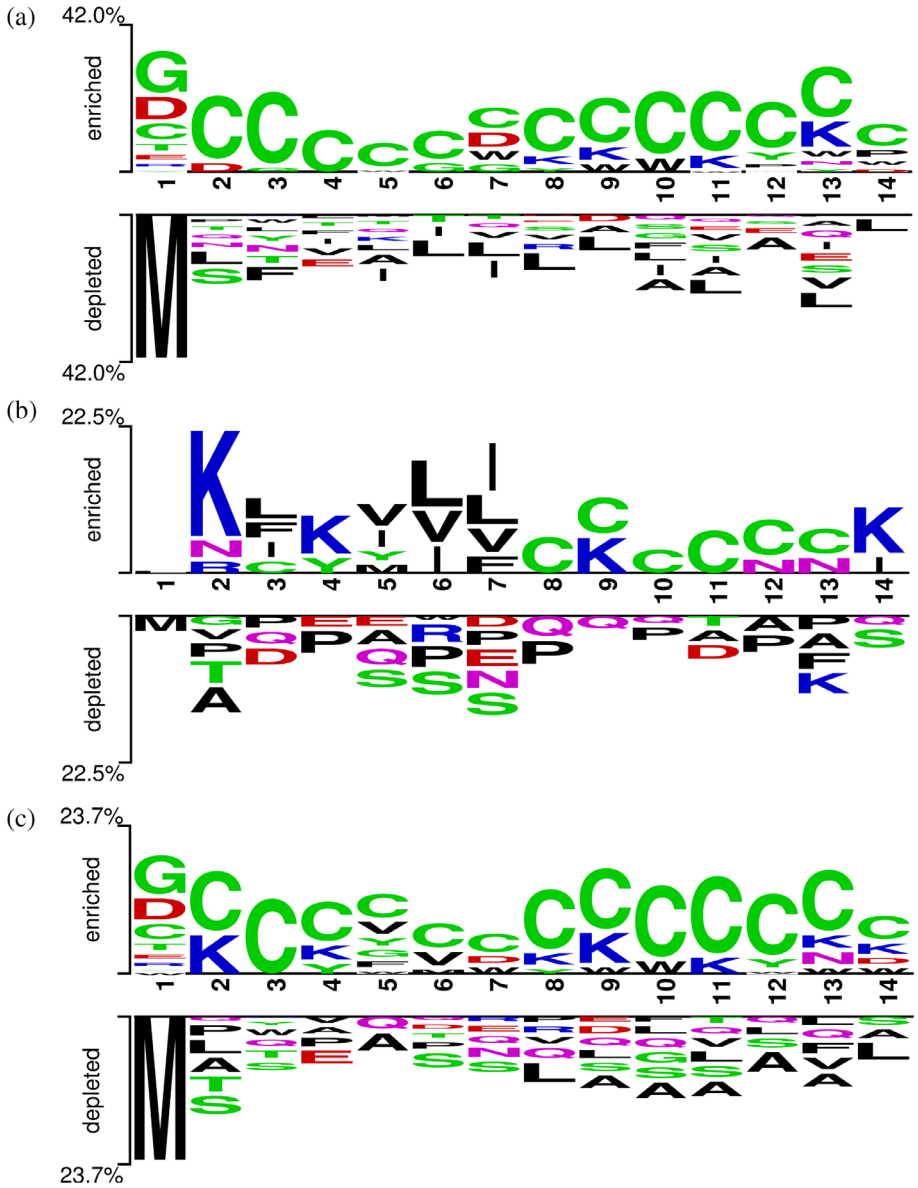

FIGURE 5 Residue preferences at different positions in neurotoxic and non-toxic sequences are illustrated using Two Sample Logo for the peptide dataset (a), protein dataset (b), and combined dataset (c). The first seven positions represent the N-terminal region, while the last seven positions represent the C-terminal region of the peptides.

markedly across the peptide, protein, and combined datasets. In the peptide dataset, descriptors such as the composition of sulfur-containing residues (PCP_SC) and advanced indices like PCP_Z3 (polarity/charge ratio) and PCP_Z5 achieved high accuracy (0.830) and AUC (0.879), indicating these properties can strongly discriminate between neurotoxic and non-toxic peptides. In contrast, for the protein dataset, most properties showed substantially lower performance with accuracies generally around 50%–60% and AUCs not exceeding 0.70–0.75, suggesting that the inherent differences in physicochemical profiles between neurotoxic and non-toxic proteins are less pronounced. The combined dataset produced intermediate metrics, reflecting a dilution of discriminative power when peptides and proteins are analyzed together. Overall, these findings critically support the development of separate, predictive models, as the features that robustly distinguish neurotoxicity in peptides do not translate equivalently to proteins. Comprehensive results are available in Table S5.

3.6 | Analysis of cysteine count

The C distribution is a key determinant of peptide and protein stability, particularly in neurotoxic peptides, which predominantly feature even C counts. This facilitates disulfide bond formation, enhancing structural rigidity, protease resistance, and bioactivity, as seen in conotoxins and scorpion toxins (Bin et al., 2017; Fahey, 1977). In peptide dataset analysis, it shows that among 877 neurotoxic peptides, 738 have an even C count, while only 78 have an odd count and 61 lack C (Table 2). In contrast, non-toxic peptides exhibit a more balanced distribution, with 195 having even

TABLE 2 The distribution of cysteine (even and odd numbers) in neurotoxic peptides, neurotoxins, non-neurotoxic peptides, and non-neurotoxins.

Datasets		Sequences without cysteine	Sequences with even cysteine count	Sequences with odd cysteine count
Peptides	Neurotoxin (877)	61	738	78
	Non-toxin (877)	515	195	167
Protein	Neurotoxin (775)	66	385	324
	Non-toxin (775)	186	247	342

counts and 515 lacking C. This suggests that even-numbered Cs can be a defining feature of neurotoxins, stabilizing their three-dimensional conformation and enhancing toxicity. In neurotoxic proteins, the difference in even and odd C counts is less pronounced; similarly, non-toxic proteins show no strong preference. The higher odd-C occurrence in proteins indicates additional stabilizing factors, such as hydrogen bonding or salt bridges (Fry, 2005).

3.7 | Machine-learning-based models

Five distinct feature sets were employed to develop ML models for classifying neurotoxic peptides: binary profiles, PSSM profiles, *Pfeature* compositional features, 56D *modlAMP* descriptors, and protein embeddings generated from LLM. These feature sets served as input for subsequent ML model training and evaluation using the Scikit-learn package (Pedregosa et al., 2012).

3.7.1 | Binary profiles

Using NT7 and CT7 binary profile features, a range of ML models were developed for all three datasets. Table 3 presents the performance metrics best models utilizing binary composition-based features. Among the evaluated various ML models, tree-based algorithms, specifically ET and RF demonstrated superior performance on all three datasets. The ET models achieved AUC values of 0.949 and 0.846 on peptide and combined independent datasets, respectively. In the protein-independent dataset, the RF model exhibited AUC values of 0.976. Similarly, using the same strategy, we also generated NT15 and CT15 binary profile features. However, the performance remained comparable. Table S6 provides performance metrics of all the models on different datasets.

3.7.2 | PSSM profiles

This study investigates the role of evolutionary information in predicting neurotoxins. Due to the short length of peptides, their PSSM profiles are unreliable, so we extracted PSSM features only for the protein dataset. Among various PSSM-derived features, PSSM composition performed best, achieving an AUC of 0.889 with the ET model on an independent protein dataset. When combined with AAC composition, the performance varied *aac_pssm* + AAC composition, decreasing the AUC to 0.861, whereas combining *pssm_composition* or *medp_pssm* with AAC composition led to a decline in performance (Table 4). These findings highlight the selective impact of evolutionary features on neurotoxin prediction. Table S7 provides detailed performance metrics of all the models on different datasets.

3.7.3 | Compositional features

Using *Pfeature*, 9190 features were generated per sequence and used to train and evaluate ML models. Among the evaluated models, tree-based models consistently exhibited superior predictive performance. Table 5 summarizes the performance of the best-performing models on all three datasets across different feature combinations. The RF model achieved a peak AUC of 0.969 and 0.873 on peptide and protein-independent datasets, respectively, using All Composition features extracted using *Pfeature* tool (ALLCOMP) features. While ALLCOMP features provided the highest accuracy, their utilization resulted in a significant computational burden. The prediction time for the full ALLCOMP set was approximately 515.23 s. In contrast, employing AAC and DPC features reduced the computational time to 48.51 s with only a minor decline in AUC. All computational experiments were conducted on a high-performance computing system featuring an *Intel(R) Xeon(R) Gold 6338 CPU (2.00GHz, 16 cores, 64-bit)*, 100.58 GB of RAM, and 768 MB L3 cache, operating within a VMware virtualization environment. In the case of a combined independent dataset, the ET model performs best with an AUC of 0.919 on AAC features.

Similarly, the *modlAMP* library is used to compute sequence-based features, such as physicochemical and structural properties, and facilitates the development of ML models for neurotoxic peptide classification and prediction. On the peptide dataset, the ET model

TABLE 3 The performance of the best machine-learning-based models developed using binary profile features on different datasets.

		Cross-validation dataset						Independent dataset					
Dataset	**Classifier**	**SENS**	**SPEC**	**PPV**	**ACC**	**MCC**	**AUC**	**SENS**	**SPEC**	**PPV**	**ACC**	**MCC**	**AUC**
Peptides	RF	0.909	0.864	0.870	0.887	0.774	0.945	0.931	0.841	0.853	0.886	0.775	0.946
	ET	0.905	0.860	0.866	0.882	0.766	0.942	0.914	0.847	0.856	0.880	0.763	0.949
Protein	RF	0.677	0.706	0.698	0.692	0.384	0.752	0.665	0.716	0.701	0.690	0.381	0.776
	Lasso	0.661	0.703	0.690	0.682	0.365	0.743	0.658	0.755	0.729	0.706	0.415	0.756
Combined	ET	0.804	0.727	0.747	0.766	0.533	0.836	0.797	0.731	0.747	0.764	0.529	0.846
	RF	0.830	0.692	0.729	0.761	0.527	0.830	0.833	0.683	0.724	0.758	0.522	0.841

Abbreviations: ACC, accuracy; AUC, area under receiver operating characteristic; ET, Extra Trees; MCC, Matthews correlation coefficient; PPV, precision; RF, Random Forest; SENS, sensitivity; SPEC, specificity.

exhibited superior performance among other models, achieving an AUC of 0.934 on independent data, as detailed in Table S8. On a protein and combined dataset, the RF model demonstrated best performance, attaining an AUC of 0.845 and 0.907 on the protein and combined independent datasets, respectively.

3.7.4 | Embeddings from fine-tuned PLMs

Embedding-based ML models have become increasingly important in various domains due to their ability to effectively represent complex data types in a lower-dimensional space (Erckert & Rost, 2024; Ko et al., 2024; Pokharel et al., 2022). In this study, we utilized embeddings from various fine-tuned ESM checkpoints and the ProtBERT model. These embeddings, derived from fine-tuned PLMs such as ESM and ProtBERT, capture both evolutionary and structural information. On the peptide dataset, ProtBERT embeddings using the LR model achieved the best performance with an AUC of 0.963 on independent data. For the protein and combined datasets, esm2-t30 embeddings with the ET model performed best, attaining an AUC of 0.934 and 0.949 on protein and combined independent data, respectively. Detailed results are presented in Table 6.

3.7.5 | Feature selection

The evaluation of various feature selection techniques across peptide, protein, and combined datasets (Table 7) highlights the potential of dimensionality reduction in maintaining or enhancing predictive performance. In the peptide dataset, we applied various feature selection methods to the ALLCOMP features, reducing them to subsets such as SVC with L1-based feature selection (SVCL1) (467 features) and tree-based selection (100 features), which achieved comparable AUCs of 0.955 and 0.952, respectively. Notably, the ET model based on simple AAC + DPC features outperformed the feature selection-based approaches. In contrast, feature selection proved highly effective for the protein and combined datasets. In the protein dataset, RFE using only 48 features achieved an independent AUC of 0.937, demonstrating that a compact feature subset is sufficient for robust neurotoxin prediction. Similarly, in the combined dataset, the SVCL1 method, which selected just 84 features, attained the highest independent AUC of 0.954. These findings underscore the importance of feature selection in reducing model complexity while preserving or even enhancing predictive accuracy, particularly in protein and combined datasets. Techniques such as RFE and SVCL1 effectively reduced the number of features while maintaining or improving key performance indicators, including AUC, sensitivity, and MCC, making them valuable strategies for optimizing neurotoxicity prediction models.

3.8 | Protein language models

In this work, we employed various PLMs to predict neurotoxic proteins and peptides. Since these models are not explicitly trained or fine-tuned for classifying neurotoxic peptides, they primarily rely on the general representations learned from large, diverse datasets. To tailor these models for the specific task, we optimized their hyperparameters using a training dataset composed of neurotoxic and non-toxic proteins and peptides, enabling them to perform effectively in this specialized classification task. The performance of fine-tuned PLMs on the different datasets is shown in Table 8. In the peptide dataset among all models, the esm2-t30 achieved the best performance on the independent data with an AUC of 0.984, while the esm2-t6 excelled on the protein and combined independent data with an AUC of 0.912 and 0.943, respectively.

3.9 | Final models

In this study, we developed a range of ML models and PLMs to predict neurotoxic and non-toxic peptides

TABLE 4 The performance of best machine-learning-based models on the protein dataset developed using position-specific scoring matrix (PSSM) profile features and in combination with Amino Acid Composition (AAC) features.

		Cross-validation dataset						Independent dataset					
Features	**Classifier**	**SENS**	**SPEC**	**PPV**	**ACC**	**MCC**	**AUC**	**SENS**	**SPEC**	**PPV**	**ACC**	**MCC**	**AUC**
aac_pssm (20)	SVC (rbf)	0.785	0.757	0.765	0.771	0.543	0.835	0.781	0.742	0.752	0.761	0.523	0.841
aac_pssm (20) + AAC composition (20)	SVC (rbf)	0.814	0.802	0.804	0.808	0.616	0.871	0.800	0.787	0.790	0.794	0.587	0.861
pssm_composition (400)	ET	**0.837**	**0.832**	**0.833**	**0.834**	**0.669**	**0.885**	**0.813**	**0.806**	**0.808**	**0.810**	**0.619**	**0.889**
pssm_composition (400) + AAC composition (20)	ET	0.672	0.716	0.702	0.694	0.388	0.760	0.697	0.710	0.706	0.703	0.406	0.747
medp_pssm (420)	RF	0.768	0.772	0.771	0.770	0.540	0.841	0.806	0.735	0.753	0.771	0.543	0.851
medp_pssm (420) + AAC composition (20)	EF	0.735	0.727	0.729	0.731	0.462	0.797	0.748	0.639	0.674	0.694	0.389	0.766

Note: The bold values indicate the best-performing models for each dataset.
Abbreviations: aac_pssm, amino acid composition from position-specific scoring matrix; ACC, accuracy; AUC, area under the receiver operating characteristic curve; ET, Extra Trees; pssm_composition, PSSM Composition; medp_pssm, combines features from two approaches, EDP and EEDP; MCC, Matthews Correlation Coefficient; PPV, positive predictive value; RF, Random Forest; SENS, sensitivity; SPEC, specificity; SVC, Support Vector Classifier.

TABLE 5 The performance of best machine-learning-based models developed using different compositional features on all datasets.

			Cross-validation dataset						Independent dataset					
Dataset	**Features**	**Classifier**	**SENS**	**SPEC**	**PPV**	**ACC**	**MCC**	**AUC**	**SENS**	**SPEC**	**PPV**	**ACC**	**MCC**	**AUC**
Peptide	AAC	RF	0.874	0.877	0.877	0.876	0.752	0.940	0.926	0.891	0.896	0.909	0.818	0.954
	DPC	RF	0.904	0.852	0.859	0.878	0.757	0.948	0.915	0.869	0.875	0.892	0.784	0.947
	AAC + DPC	ET	0.890	0.877	0.875	0.883	0.766	0.951	0.911	0.880	0.874	0.895	0.790	0.959
	ALLCOMP	RF	**0.890**	**0.870**	**0.873**	**0.880**	**0.761**	**0.944**	**0.937**	**0.903**	**0.906**	**0.920**	**0.841**	**0.969**
Protein	AAC	ET	**0.805**	**0.815**	**0.813**	**0.810**	**0.619**	**0.876**	**0.794**	**0.813**	**0.809**	**0.803**	**0.607**	**0.873**
	DPC	ET	0.795	0.805	0.803	0.800	0.600	0.864	0.794	0.774	0.778	0.784	0.568	0.852
	AAC + DPC	ET	0.794	0.813	0.809	0.803	0.607	0.865	0.794	0.742	0.755	0.768	0.536	0.850
	ALLCOMP	RF	0.773	0.852	0.839	0.812	0.626	0.877	0.748	0.858	0.841	0.803	0.610	0.873
Combined	AAC	ET	**0.833**	**0.846**	**0.844**	**0.840**	**0.679**	**0.905**	**0.855**	**0.848**	**0.850**	**0.852**	**0.703**	**0.919**
	DPC	ET	0.854	0.828	0.832	0.841	0.682	0.907	0.846	0.833	0.836	0.840	0.679	0.907
	AAC + DPC	ET	0.845	0.843	0.843	0.844	0.687	0.909	0.840	0.845	0.845	0.843	0.685	0.914
	ALLCOMP	ET	0.832	0.865	0.860	0.848	0.697	0.916	0.849	0.855	0.854	0.852	0.703	0.917

Note: The bold values indicate the best-performing models for each dataset.
Abbreviations: AAC, Amino Acid Composition; ACC, accuracy; ALLCOMP, All Composition features extracted using *Pfeature* tool; AUC, area under the receiver operating characteristic curve; DPC, Dipeptide Composition; ET, Extra Trees; MCC, Matthews Correlation Coefficient; PPV, positive predictive value; RF, Random Forest; SENS, sensitivity; SPEC, specificity.

TABLE 6 The performance of machine-learning models developed using protein language model embeddings on all three datasets.

			Cross-validation data						Independent data					
Dataset	**Embedding source**	**Classifier**	**SENS**	**SPEC**	**PPV**	**ACC**	**MCC**	**AUC**	**SENS**	**SPEC**	**PPV**	**ACC**	**MCC**	**AUC**
Peptide	esm2 – t6	ET	0.882	0.883	0.883	0.882	0.765	0.944	0.897	0.858	0.863	0.877	0.756	0.947
	esm2 – t12	ET	0.869	0.902	0.898	0.885	0.771	0.949	0.880	0.881	0.880	0.880	0.761	0.941
	esm2 – t30	ET	0.875	0.914	0.911	0.895	0.790	0.955	0.880	0.886	0.885	0.883	0.766	0.954
	esm2 – t33	ET	0.870	0.920	0.916	0.895	0.791	0.951	0.891	0.903	0.902	0.897	0.795	0.950
	esm2 – t36	RF	0.859	0.913	0.908	0.886	0.773	0.952	0.863	0.915	0.910	0.889	0.779	0.953
	ProtBERT	LR	**0.932**	**0.919**	**0.920**	**0.925**	**0.850**	**0.964**	**0.920**	**0.891**	**0.895**	**0.906**	**0.812**	**0.963**
Protein	esm2 – t6	ET	0.803	0.860	0.851	0.831	0.664	0.902	0.800	0.865	0.855	0.832	0.666	0.896
	esm2 – t12	ET	0.790	0.874	0.863	0.832	0.667	0.908	0.787	0.910	0.897	0.848	0.702	0.918
	esm2 – t30	ET	**0.832**	**0.894**	**0.887**	**0.863**	**0.727**	**0.923**	**0.819**	**0.903**	**0.894**	**0.861**	**0.725**	**0.934**
	esm2 – t33	ET	0.817	0.905	0.906	0.851	0.725	0.925	0.806	0.931	0.924	0.844	0.737	0.930
	esm2 – t36	ET	0.806	0.923	0.912	0.865	0.734	0.922	0.800	0.918	0.919	0.854	0.717	0.924
	ProtBERT	ET	0.727	0.815	0.797	0.771	0.544	0.853	0.729	0.774	0.764	0.752	0.504	0.861
Combined	esm2 – t6	ET	0.832	0.878	0.872	0.855	0.711	0.920	0.831	0.888	0.881	0.859	0.720	0.931
	esm2 – t12	ET	0.831	0.888	0.881	0.860	0.720	0.927	0.840	0.906	0.900	0.873	0.748	0.934
	esm2 – t30	ET	**0.846**	**0.899**	**0.893**	**0.872**	**0.746**	**0.936**	**0.855**	**0.930**	**0.925**	**0.893**	**0.787**	**0.949**
	esm2 – t33	SVC (linear)	0.839	0.864	0.864	0.866	0.733	0.931	0.850	0.912	0.909	0.891	0.773	0.949
	esm2 – t36	ElasticNet	0.871	0.886	0.884	0.879	0.757	0.945	0.870	0.876	0.875	0.873	0.746	0.943
	ProtBERT	ET	0.795	0.803	0.801	0.799	0.597	0.886	0.831	0.803	0.809	0.817	0.634	0.903

Note: The bold values indicate the best-performing models for each dataset.

Abbreviations: ACC, accuracy; AUC, area under the receiver operating characteristic curve; ET, Extra Trees; LR, Logistic Regression; MCC, Matthews Correlation Coefficient; PPV, positive predictive value; RF, Random Forest; SENS, sensitivity; SPEC, specificity; SVC, Support Vector Classifier.

TABLE 7 The performance of extra tree-based models developed using a selected set of features obtained using different feature selection methods.

			Cross-validation data						Independent data					
Dataset	**Featutre selection method**	**No of features**	**SENS**	**SPEC**	**PREC**	**ACC**	**MCC**	**AUC**	**SENS**	**SPEC**	**PREC**	**ACC**	**MCC**	**AUC**
Peptides	VTh	184	0.892	0.883	0.884	0.887	0.775	0.950	0.926	0.886	0.891	0.906	0.813	0.956
	k-best	55	0.890	0.872	0.874	0.881	0.762	0.947	0.920	0.880	0.885	0.900	0.801	0.954
	RFE	148	0.876	0.866	0.867	0.871	0.742	0.945	0.920	0.891	0.895	0.906	0.812	0.956
	SFS	128	0.874	0.870	0.871	0.872	0.745	0.941	0.920	0.886	0.890	0.903	0.807	0.944
	SVCL1	467	0.906	0.885	0.887	0.895	0.791	0.959	0.909	0.897	0.899	0.903	0.806	0.955
	Tree based	100	0.894	0.872	0.874	0.883	0.766	0.946	0.915	0.880	0.885	0.897	0.795	0.952
	No feature selection	**420 (AAC + DPC)**	**0.890**	**0.877**	**0.875**	**0.883**	**0.766**	**0.951**	**0.911**	**0.880**	**0.874**	**0.895**	**0.790**	**0.959**
Proteins	VTh	67	0.821	0.895	0.887	0.858	0.718	0.914	0.813	0.903	0.894	0.858	0.719	0.935
	k-best	50	0.824	0.861	0.856	0.843	0.686	0.917	0.826	0.877	0.871	0.852	0.704	0.929
	RFE	**48**	**0.850**	**0.889**	**0.884**	**0.869**	**0.739**	**0.929**	**0.826**	**0.903**	**0.895**	**0.865**	**0.731**	**0.937**
	SFS	50	0.794	0.877	0.866	0.835	0.673	0.905	0.794	0.884	0.872	0.839	0.680	0.924
	SVCL1	66	0.840	0.881	0.876	0.860	0.722	0.927	0.852	0.877	0.874	0.865	0.729	0.934
	Tree based	65	0.831	0.871	0.866	0.851	0.702	0.919	0.826	0.884	0.877	0.855	0.711	0.930
	No feature selection (640 esm2-t30 embeddings)	640	0.832	0.894	0.887	0.863	0.727	0.923	0.819	0.903	0.894	0.861	0.725	0.934
Combined	VTh	66	0.833	0.893	0.886	0.863	0.728	0.926	0.849	0.915	0.909	0.882	0.766	0.941
	k-best	50	0.830	0.886	0.879	0.858	0.717	0.926	0.825	0.894	0.886	0.859	0.720	0.943
	RFE	70	0.853	0.896	0.892	0.875	0.750	0.937	0.855	0.927	0.922	0.891	0.784	0.949
	SFS	50	0.839	0.890	0.884	0.864	0.729	0.924	0.843	0.903	0.897	0.873	0.747	0.930
	SVCL1	**84**	**0.852**	**0.898**	**0.893**	**0.875**	**0.751**	**0.940**	**0.870**	**0.909**	**0.906**	**0.890**	**0.780**	**0.954**
	Tree based	50	0.835	0.891	0.885	0.863	0.727	0.927	0.846	0.903	0.897	0.874	0.750	0.940
	No feature selection (640 esm2-t30 embeddings)	640	0.846	0.899	0.893	0.872	0.746	0.936	0.855	0.93	0.925	0.893	0.787	0.949

Note: The bold values indicate the best-performing models for each dataset.

Abbreviations: AAC, Amino Acid Composition; ACC, accuracy; AUC, area under the receiver operating characteristic curve; DPC, Dipeptide Composition; MCC, Matthews Correlation Coefficient; PPV, positive predictive value; PREC, precision; RFE, recursive feature elimination; SENS, sensitivity; SFS, sequential forward selection; SPEC, specificity; SVCL1, Support Vector Classifier with L1-based feature selection; VTh, Variance Threshold.

TABLE 8 The performance evaluation of protein language models (PLMs) on all the datasets.

		Cross-validation data						Independent data					
Dataset	**PLM classifier**	**SENS**	**SPEC**	**PPV**	**ACC**	**MCC**	**AUC**	**SENS**	**SPEC**	**PPV**	**ACC**	**MCC**	**AUC**
Peptides	esm2-t6	0.892	0.881	0.882	0.923	0.770	0.952	0.960	0.847	0.862	0.903	0.817	0.950
	esm2-t12	0.879	0.901	0.882	0.883	0.776	0.951	0.942	0.887	0.862	0.913	0.820	0.961
	esm2-t30	**0.937**	**0.941**	**0.944**	**0.943**	**0.877**	**0.981**	**0.926**	**0.970**	**0.970**	**0.951**	**0.898**	**0.984**
	esm2-t33	0.648	0.972	0.835	0.937	0.712	0.952	0.952	0.880	0.883	0.911	0.830	0.961
	ProtBERT	0.976	0.862	0.876	0.919	0.843	0.970	0.949	0.800	0.827	0.875	0.758	0.961
Proteins	esm2-t6	**0.89**	**0.903**	**0.917**	**0.891**	**0.800**	**0.930**	**0.804**	**0.890**	**0.903**	**0.841**	**0.693**	**0.912**
	esm2-t12	0.907	0.898	0.911	0.894	0.843	0.940	0.839	0.787	0.797	0.810	0.630	0.900
	esm2-t30	0.888	0.912	0.910	0.900	0.801	0.953	0.804	0.873	0.864	0.838	0.678	0.912
	esm2-t33	0.834	0.870	0.865	0.852	0.704	0.915	0.833	0.859	0.855	0.846	0.692	0.905
	ProtBERT	0.827	0.800	0.805	0.813	0.627	0.877	0.813	0.791	0.796	0.802	0.604	0.875
Combined	esm2-t6	**0.921**	**0.886**	**0.889**	**0.900**	**0.803**	**0.952**	**0.867**	**0.875**	**0.875**	**0.873**	**0.742**	**0.943**
	esm2-t12	0.872	0.869	0.854	0.897	0.793	0.941	0.778	0.894	0.893	0.84	0.68	0.922
	esm2-t30	0.853	0.856	0.856	0.855	0.710	0.917	0.865	0.865	0.865	0.865	0.729	0.931
	esm2-t33	0.848	0.860	0.858	0.854	0.708	0.907	0.865	0.871	0.870	0.868	0.735	0.920
	ProtBERT	0.777	0.815	0.807	0.796	0.592	0.868	0.755	0.787	0.780	0.771	0.542	0.846

Note: The bold values indicate the best-performing models for each dataset.
Abbreviations: ACC, accuracy; AUC, area under the receiver operating characteristic curve; MCC, Matthews Correlation Coefficient; PPV, positive predictive value; SENS, sensitivity; SPEC, specificity.

and proteins using three distinct datasets. Among these, the esm2-t30 model demonstrated the highest performance on the peptide dataset, achieving an AUC of 0.984 and an MCC of 0.898. For both the protein and combined datasets, the ET model, utilizing esm2-t30 embeddings as features, exhibited superior predictive capability. In the protein dataset, the ET model achieved the best performance with an AUC of 0.937 and an MCC of 0.731, having been trained on 48 embeddings selected through RFE. Similarly, in the combined dataset, the ET model attained optimal performance with an AUC of 0.954 and an MCC of 0.780, leveraging 84 embeddings selected via the SVCL1 method. These findings highlight the effectiveness of feature selection techniques in optimizing model performance across different datasets.

3.10 | Cross-dataset prediction of final models

In this section, we evaluated the performance of the models across all three datasets to determine their generalizability. Specifically, we assessed whether a model trained on peptides could effectively predict neurotoxic proteins, whether a model trained on proteins could accurately predict neurotoxic peptides, and whether a combined model exhibited superior predictive performance compared to models trained exclusively on peptide or protein datasets. This analysis provides insights into whether separate models are necessary for peptides and proteins or if a unified model can effectively generalize across both categories.

In Table 9 cross-dataset analysis reveals significant variations in model performance depending on the dataset used for training and validation. The ESM2-t30 model, developed on the neurotoxic peptide dataset, exhibits exceptional performance on the independent peptide dataset (AUC 0.984), but it performed very poorly on the protein dataset (AUC 0.752). This discrepancy suggests that the peptide-specific model is not well-suited for identifying neurotoxic proteins, likely due to intrinsic differences in sequence composition and structural features between peptides and proteins. The ET model developed on the neurotoxic protein dataset shows high performance on the independent protein dataset (AUC 0.937) but performed poorly on the peptide dataset (AUC 0.883). Notably, the ET model developed on the combined dataset delivers consistently high performance across all independent datasets, achieving AUCs of 0.954, 0.939, and 0.967 for the combined, protein, and peptide datasets, respectively. However, the method developed on a combined dataset performs poorer than the model trained on peptides for predicting peptides. Overall, it is clear from the above analysis that there is a need to develop separate models for proteins and peptides.

TABLE 9 The performance of models on cross datasets, performance of each model predicted on different datasets.

Final model	Dataset	SENS	SPEC	PPV	ACC	MCC	AUC
esm2-t30 (peptide dataset)	Independent peptide dataset	**0.926**	**0.970**	**0.970**	**0.951**	**0.898**	**0.984**
	Independent protein dataset	0.386	0.981	0.951	0.674	0.441	0.752
	Independent combined dataset	0.665	0.976	0.963	0.821	0.674	0.879
Extra Tree (protein dataset)	Independent protein dataset	**0.826**	**0.903**	**0.895**	**0.865**	**0.731**	**0.937**
	Independent peptide dataset	0.653	0.931	0.906	0.792	0.608	0.883
	Independent combined dataset	0.734	0.918	0.9	0.826	0.663	0.904
Extra Tree (combined dataset)	Independent combined dataset	**0.870**	**0.909**	**0.906**	**0.890**	**0.780**	**0.954**
	Independent protein dataset	0.838	0.91	0.903	0.874	0.75	0.939
	Independent peptide dataset	0.892	0.914	0.912	0.903	0.807	0.967

Note: The bold values indicate the best-performing models in particular dataset.
Abbreviations: ACC, accuracy; AUC, area under the receiver operating characteristic curve; MCC, Matthews Correlation Coefficient; PPV, positive predictive value; SENS, sensitivity; SPEC, specificity.

3.11 | Benchmarking with existing methods

Thorough comparison against existing methods is crucial for evaluating the advancements of newly developed approaches. NTxPred2, an updated version of NTxPred, incorporates several key innovations, including the integration of fine-tuned PLMs, novel composition and physicochemical features. Table 10 presents a comparative analysis of the proposed NTxPred2 framework against its predecessor, NTxPred, using independent datasets. To ensure a comprehensive evaluation, NTxPred2 was benchmarked against various neurotoxicity prediction models previously developed by Saha et al. Among these, the AAC-based SVM model exhibited the highest performance on the independent peptide dataset, achieving an AUC of 0.922. For protein and combined independent datasets, the SVM (DPC) model demonstrated superior performance, with AUC values of 0.715 and 0.783, respectively. Additionally, the convolutional neural network (CNN) model developed by Lee et al. (2021) achieved an AUC of 0.689 on the independent protein dataset. However, a direct comparative evaluation with tools developed by Yang and Li (2009), Guang et al. (2010), Song Chaohong et al. (2012), Tang et al. (2017), Huo et al. (2017), Koua and Kuhn-Nentwig (2017), Mei and Zhao (2018), and Wan et al. (2023) was not feasible due to the unavailability of their underlying algorithms. Pippin, a RF-based method (Li et al., 2020), could not be compared due to the web service being nonfunctional. The results clearly demonstrate that NTxPred2 outperforms existing methods. Its consistently superior performance highlights its potential as a valuable tool for therapeutic peptide development, particularly in the classification of neurotoxic proteins and peptides.

To assess whether the differences in predictive performance among these tools were statistically significant, we applied the Friedman test separately on the peptide, protein, and combined datasets. The results indicated significant variation in AUROC scores across the tools for all three datasets: peptide dataset ($\chi^2 = 37.37$, $p = 1.51 \times 10^{-7}$), protein dataset ($\chi^2 = 32.94$, $p = 1.23 \times 10^{-6}$), and the combined dataset ($\chi^2 = 38.72$, $p = 7.96 \times 10^{-8}$). These findings confirm that at least one tool performs significantly differently from the others in each dataset. Among the evaluated tools, our proposed model, NTxPred2, consistently achieved the highest performance with notably low variability: AUROC = 0.984 ± 0.02 on the peptide dataset, 0.937 ± 0.05 on the protein dataset, and 0.954 ± 0.02 on the combined dataset. The lower standard deviation compared to other tools indicates that NTxPred2 provides not only high but also stable and reliable performance across different data splits. To further investigate tool-wise differences, we performed a post hoc Nemenyi test, and the detailed pairwise comparison results are provided in Table S9.

3.12 | Case study

To test the robustness and flexibility of our predictive program, NTxPred2, we tested its performance on a separate data set of therapeutic peptides curated from THPdb2, a database of approved therapeutic peptides and proteins. Although it is generally considered that these peptides are not toxic because they are used therapeutically, it is noteworthy that some peptides, like vasopressin (Ito et al, 1997), have been documented to be neurotoxic under certain conditions. Therefore, it becomes increasingly significant to reassess therapeutic peptides based on predictive methods like NTxPred2 for the estimation of possible neurotoxicity. A total of 30 peptides were collected and tested using the prediction model esm2-t30 of our tool (Table S10). It was observed that almost all the therapeutic peptides were non-toxic, with some exceptions, such as vasopressin, insulin regular, and nesiritide. Our validation further establishes NTxPred2's predictive accuracy on

TABLE 10 Benchmarking of existing methods on independent datasets of NTxPred2.

Tool	Model	Data	SENS	SPEC	PPV	ACC	MCC	AUC
NTxPred	SVM (DPC + Length)	Independent peptide dataset	0.790 ± 0.11	0.794 ± 0.09	0.794 ± 0.09	0.792 ± 0.10	0.584 ± 0.16	0.827 ± 0.13
		Independent protein dataset	0.412 ± 0.09	0.761 ± 0.03	0.634 ± 0.06	0.587 ± 0.04	0.185 ± 0.05	0.589 ± 0.11
		Independent combined dataset	0.589 ± 0.10	0.603 ± 0.07	0.598 ± 0.06	0.596 ± 0.08	0.192 ± 0.08	0.656 ± 011
	SVM (DPC)	Independent peptide dataset	0.898 ± 0.14	0.811 ± 0.14	0.827 ± 0.06	0.855 ± 0.06	0.712 ± 0.12	0.916 ± 0.08
		Independent protein dataset	0.639 ± 0.18	0.619 ± 0.19	0.627 ± 0.33	0.629 ± 0.11	0.258 ± 0.18	0.715 ± 0.11
		Independent combined dataset	0.652 ± 0.13	0.772 ± 0.15	0.742 ± 0.05	0.712 ± 0.07	0.428 ± 0.10	0.783 ± 0.06
	SVM (AAC + Length)	Independent peptide dataset	0.813 ± 0.09	0.709 ± 0.11	0.737 ± 0.16	0.761 ± 0.12	0.524 ± 0.13	0.832 ± 0.08
		Independent protein dataset	0.529 ± 0.11	0.658 ± 0.03	0.607 ± 0.03	0.593 ± 0.05	0.188 ± 0.08	0.648 ± 0.09
		Independent combined dataset	0.607 ± 0.10	0.755 ± 0.06	0.712 ± 0.12	0.680 ± 0.07	0.365 ± 0.10	0.692 ± 0.09
	SVM (AAC)	Independent peptide dataset	0.932 ± 0.08	0.777 ± 0.09	0.808 ± 0.11	0.855 ± 0.06	0.718 ± 0.11	0.922 ± 0.13
		Independent protein dataset	0.574 ± 0.12	0.651 ± 0.11	0.622 ± 0.09	0.612 ± 0.09	0.226 ± 0.18	0.669 ± 0.12
		Independent combined dataset	0.661 ± 0.09	0.766 ± 0.09	0.640 ± 0.10	0.714 ± 0.07	0.430 ± 0.14	0.770 ± 0.11
Lee et al. (2021)	CNN	Independent peptide dataset	0.00 ± 0	1.00 ± 0	0.00 ± 0	0.498 ± 0.01	0.00 ± 0	0.500 ± 0
		Independent protein dataset	0.277 ± 0.09	0.993 ± 0.08	0.977 ± 0.14	0.635 ± 0.07	0.388 ± 0.10	0.63 ± 0.20
		Independent combined dataset	0.130 ± 0.03	0.996 ± 0.00	0.977 ± 0.52	0.563 ± 0.02	0.245 ± 0.09	0.563 ± 0.02
NTxPred2	esm2-t30 (peptide dataset)	Independent peptide dataset	**0.926 ± 0.03**	**0.970 ± 0.04**	**0.970 ± 0.04**	**0.951 ± 0.01**	**0.898 ± 0.03**	**0.984 ± 0.02**
	Extra Tree (protein dataset)	Independent protein dataset	**0.827 ± 0.11**	**0.903 ± 0.05**	**0.895 ± 0.05**	**0.865 ± 0.07**	**0.731 ± 0.14**	**0.937 ± 0.05**
	Extra Tree (combined dataset)	Independent combined dataset	**0.870 ± 0.05**	**0.909 ± 0.05**	**0.906 ± 0.05**	**0.890 ± 0.04**	**0.780 ± 0.08**	**0.954 ± 0.02**

Note: The bold values indicate the best-performing models in particular dataset.

Abbreviations: AAC, Amino Acid Composition; ACC, accuracy; AUC, area under the receiver operating characteristic curve; CNN, convolutional neural network; DPC, Dipeptide Composition; SENS, sensitivity; SPEC, specificity; SVM, support vector machine; PPV, positive predictive value; MCC, Matthews Correlation Coefficient.

real-world biologically relevant data and opens up possibilities for drug repurposing, as peptides predicted to be non-toxic can be researched for alternative use in neurological or systemic applications where safety is a primary concern.

3.13 | Design and deployment of a web server

To enhance the accessibility and utility of our research findings, we developed NTxPred2, a comprehensive platform consisting of both a user-friendly web server, standalone software, and *pip* package. This platform is freely accessible to the scientific community. The NTxPred2 web server offers a suite of functionalities, including peptide and protein neurotoxicity prediction, protein scanning for potential neurotoxic regions, and design of non-toxic peptides/proteins. For users requiring high-throughput analysis, a standalone software and a *pip* package are available for download, enabling efficient and large-scale neurotoxic peptide prediction.

4 | DISCUSSION

The genomes of numerous organisms have been sequenced over the past three decades due to advancements in sequencing technology. This has resulted in a rapid expansion of protein databases,

which store information on protein sequences, structures, functions, and therapeutic applications. A major challenge in the post-genomic era is the annotation of these rapidly growing databases of proteins. Highlighting the significance of protein annotation, the 2024 Nobel Prize was awarded to researchers for their groundbreaking contributions to the structural annotation of proteins. Therapeutic annotation of a protein is another crucial aspect of annotation, where researchers are exploiting therapeutic applications of proteins. Over the past few decades, the US FDA has approved a significant number of proteins for therapeutic applications, recognizing their potential in treating various diseases (Arora et al., 2024; Gahlot et al., 2024; Kumar et al., 2025). In addition to sequencing, GM organisms, GM crops, and recombinant technology have played a key role in the large-scale production of therapeutic proteins like monoclonal antibodies and hormones (Plotkin, 2010; Walsh, 2018). The technological advancement has also played a crucial role in shifting vaccine development from traditional whole-organism vaccines to protein-based vaccines (Díaz-Dinamarca et al., 2022; Draper et al., 2015).

In summary, the use of proteins in therapeutics is rapidly increasing, presenting significant challenges in assessing their safety for humans, animals, and the environment. To address these concerns, the WHO has established guidelines for evaluating the safety of GM foods and protein-based therapies. Toxicity assessment is a crucial component of the safety evaluation of GM foods and therapeutic proteins. Ensuring that these proteins do not cause adverse effects is essential for their regulatory approval and clinical use. Previously, a number of computational tools have been developed to predict and assess cytotoxicity, hemotoxicity, immunotoxicity, and neurotoxicity. The motivation for this work stems from the growing demand for accurate, scalable, and ethical tools to assess protein toxicity, particularly in the context of therapeutic development and GM foods (Oh et al., 2009). While earlier tools like NTxPred (Saha & Raghava, 2007a, 2007b) used a single model for neurotoxins, recent advances in both biological data availability and AI allow us to build more specialized and accurate systems.

Our primary analysis revealed significant differences in the amino acid composition of neurotoxic sequences compared to non-toxic sequences, particularly a higher frequency of C. Enrichment of C in neurotoxins enables disulfide-bonded scaffolds, conferring protease resistance and structural stability, a hallmark of venom peptides evolved under predator–prey arms races (Jimenez & Cruz, 2016). We also compared the amino acid composition of neurotoxic peptides and neurotoxins and observed a significant difference. Neurotoxic peptides exhibit a very high association with overall toxin content, reflecting rapid divergence driven by strong positive selection for optimized PCP such as charge and hydrophobicity that enhance ion channel interactions (Casewell et al., 2013; Fry et al., 2009). In contrast, neurotoxins typically are larger and preserve conserved residues like inhibitory cystine knots, displaying a strong correlation to the general genome, suggesting that their evolutionary trajectory is tempered by the need to maintain structural integrity while permitting functional diversification (Pineda et al., 2020; Sunagar et al., 2016). This composition difference justifies our hypothesis to develop a separate method for predicting neurotoxic peptides and for predicting neurotoxins. We also compare the amino acid composition of proteins/peptides having different types of toxicities. Our primary analysis justifies the development model for predicting neurotoxic peptides/proteins.

Composition and physicochemical property-based prediction analysis highlight significant differences in discriminatory power between neurotoxins and non-toxins. In peptides, cystine and L composition alone achieved high predictive performance, with AUC values of 0.90 and 0.75 and accuracies of 0.87 and 0.70, respectively. However, in proteins, cystine-based composition performed less effectively, with an AUC of 0.69 and an accuracy of 0.65. Physicochemical descriptors such as polarity (PCP_PO), non-polar residues (PCP_NP), aliphatic side chains (PCP_AL), and sulfur-containing residues (PCP_SC) demonstrated strong predictive power in peptides but were less effective in proteins. Notably, advanced indices (PCP_Z3, PCP_Z5) remained reliable across both datasets, indicating their robustness in toxicity prediction. Finally, the C count analysis revealed that non-toxic peptides tend to favor odd C counts, whereas neurotoxic peptides are more likely to exhibit even C counts. This finding suggests that for the development of therapeutic peptides, particularly those intended to be non-toxic, careful consideration should be given to C distribution, as odd C counts may contribute to reduced toxicity.

This study explored a diverse range of feature-generation and feature-selection strategies to optimize predictive performance. Among the traditional ML models, tree-based models (ET and RF) exhibited superior performance. In peptide datasets, the esm2-t30 model performed best (AUC 0.98) on peptide-independent datasets. For both the protein and combined datasets, the ET model, utilizing esm2-t30 embeddings as features, exhibited superior predictive capability. In the protein dataset, the ET model achieved the best performance with an AUC of 0.94 and an MCC of 0.73, having been trained on 48 embeddings selected through RFE. Similarly, in the combined dataset, the ET model attained optimal performance with an AUC of 0.95 and an MCC of 0.78, using 84 embeddings selected via the SVCL1 method. Cross-dataset predictions demonstrated superior performance for models trained on the peptide dataset

compared to those trained on combined or protein datasets. While the protein model (ET) achieved an AUC of 0.937, comparable to the combined model (AUC 0.939), these results underscore inherent compositional and functional differences between peptides and proteins, supporting the development of specialized models. Although the combined model integrates diverse sequence characteristics, the observed performance differences suggest that domain-specific models remain preferable for maximizing predictive accuracy within their respective classes. Efficient computation is a critical aspect of developing practical in silico tools, especially for large-scale applications in therapeutic screening and food safety assessment.

Although our ESM-based model demonstrates high predictive accuracy for neurotoxin proteins, several important limitations should be considered. The dataset is composed exclusively of reviewed (Swiss-Prot) UniProt entries, ensuring high annotation quality, but it may introduce curation and research focus biases, particularly an overrepresentation of neurotoxins from well-studied species such as Clostridium botulinum, spiders, and scorpions. Also, the dataset may miss out on lesser-known or novel neurotoxins due to their poor annotation. This taxonomic skew may limit the model's ability to generalize to neurotoxins from less-studied or non-model organisms. Additionally, due to the complexity of functional annotations and keyword-based searches in UniProt, it is possible that some proteins included in our dataset are not neurotoxins themselves but are proteins affected by neurotoxins, potentially introducing noise and affecting model specificity. To assess the model's generalizability beyond the training and independent data, we evaluated its predictions on a set of therapeutic peptides from THPdb2, which are not expected to be toxic. The model correctly identified 27 out of 30 of these peptides as non-toxic, suggesting a promising ability to distinguish neurotoxins from unrelated therapeutic peptides. This is important, as even therapeutically approved peptides (e.g., vasopressin and nesiritide) have the potential to cause neurotoxic hazards under specific conditions. It is important to note that even therapeutically approved peptides have the potential to cause neurotoxic hazards under specific conditions (Ito et al., 1997; Nampoothiri et al., 2014). These results highlight not only the model's robustness but also its potential utility in screening candidate therapeutics for unintended neurotoxic liabilities. Furthermore, such therapeutic peptides could be explored in future drug repurposing studies, particularly where neuroactivity or central nervous system targeting is relevant. In this context, the integration of DL techniques, such as autoencoders for multi-omics data fusion, as demonstrated in DeepDRA (Mohammadzadeh-Vardin et al., 2024), and comprehensive multivariate pattern recognition frameworks (Razzaghi et al., 2023), could further enhance the interpretability and precision of such predictions. Nonetheless, broader validation on more diverse and experimentally confirmed datasets will be essential to fully establish the robustness and applicability of our approach.

High computational cost can hinder real-time analysis, limit accessibility on standard hardware, and slow down model deployment in resource-constrained settings. To address this challenge, we implemented tailored strategies that balance performance with efficiency. For peptides, we selected the smaller esm2-t30 variant, which offers strong predictive power with reduced memory and processing requirements. For proteins and hybrid datasets, we employed ET classifiers known for their speed, scalability, and suitability for high-dimensional biological data. Additionally, we used parallel processing and stratified cross-validation to optimize training time while preserving model robustness. To ensure broad usability, we also provide a standalone version of NTxPred2, enabling accurate predictions on local machines with limited computational resources.

5 | LIMITATIONS AND FUTURE DIRECTIONS

There is a growing need for unified in silico platforms to predict multiple forms of toxicity in therapeutic proteins and GM foods. These tools should follow WHO guidelines, reduce reliance on animal testing, and support comprehensive, scalable safety assessments in line with the 3Rs principle (Replacement, Reduction, and Refinement). One limitation of the present study lies in the strategy used for data curation: neurotoxin sequences were extracted from Swiss-Prot using keyword-based filtering. While this approach ensures specificity, it may also exclude relevant sequences or include ambiguous entries based solely on annotations. A more robust future direction would be to incorporate sequence-level validation or functional screening prior to inclusion, such as domain/motif-based filtering and manual curation. Such methods could significantly enhance the biological reliability of the dataset and improve downstream model performance. Looking forward, one important avenue is to incorporate biological mechanisms of neurotoxicity, particularly the disruption of ion channels and impairment of neurotransmitter activity, which are key drivers of neural dysfunction. Future models can integrate such mechanistic features, such as known ion-channel binding motifs or neurotransmitter interaction sites, to enhance both the biological interpretability and functional relevance of predictions. Developing and refining such computational models will accelerate drug discovery, improve safety assessments, and support regulatory decision-making, ultimately leading to safer and more effective therapeutic interventions.

6 | CONCLUSION

This study introduces NTxPred2, a comprehensive platform for classifying neurotoxic and non-toxic sequences. NTxPred2 integrates a high-accuracy classification model and is freely accessible via a user-friendly web server. All three best-performing models are deployed on our server. To cater to diverse research needs, a pip package and a standalone software package are also provided. While NTxPred2 offers significant advancements in neurotoxicity prediction, it is important to acknowledge certain limitations. Currently, the platform relies solely on peptide sequence information for classification and does not incorporate information regarding the peptide's source of origin and structure. Furthermore, NTxPred2 is primarily designed for natural peptides, excluding non-canonical amino acids, modified peptides, and peptides shorter than seven residues. This focused approach, while enhancing model efficiency, may limit the applicability of NTxPred2 to specific peptide classes.

AUTHOR CONTRIBUTIONS

Anand Singh Rathore: Software; conceptualization; methodology; data curation; writing – original draft; writing – review and editing; formal analysis; investigation. **Saloni Jain:** Software; validation; visualization; formal analysis. **Shubham Choudhury:** Software; writing – review and editing. **Gajendra P. S. Raghava:** Project administration; resources; supervision; writing – review and editing; funding acquisition; investigation.

ACKNOWLEDGMENTS

The authors express their gratitude to the University Grants Commission (UGC), Council of Scientific and Industrial Research (CSIR), Department of Science and Technology (DST-INSPIRE), and Department of Biotechnology (DBT) for their generous fellowships and financial support. They also thank the Department of Computational Biology, IIITD, New Delhi, for the excellent infrastructure and facilities. The authors would like to acknowledge the Department of Biotechnology (DBT) for the infrastructure grant awarded to the institute. Furthermore, they would like to acknowledge BioRender.com for creating the figures utilized in this work. The authors acknowledge the use of ChatGPT in improving the language of the manuscript.

CONFLICT OF INTEREST STATEMENT

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

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be accessed on the "NTxPred2" web server at https://webs.iiitd.edu.in/raghava/ntxpred2/download.php, publicly available on GitHub https://github.com/raghavagps/ntxpred2, and from Zenodo (https://zenodo.org/records/15030415) (Rathore, Jain, et al., 2025). The source code for this study is publicly available on GitHub and can be found at https://github.com/raghavagps/ntxpred2, the "NTxPred2" web server at https://webs.iiitd.edu.in/raghava/ntxpred2/download.php, Zenodo (https://zenodo.org/records/15030415) (Rathore, Kumar, et al., 2025). BioRxiv doi: https://doi.org/10.1101/2025.03.01.640936, and on Hugging Face Model Hub https://huggingface.co/raghavagps-group/NTxPred2.

ORCID

Anand Singh Rathore https://orcid.org/0009-0004-8907-7174
Saloni Jain https://orcid.org/0009-0009-3181-1037
Shubham Choudhury https://orcid.org/0000-0002-4509-4683
Gajendra P. S. Raghava https://orcid.org/0000-0002-8902-2876

REFERENCES

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389–402. https://doi.org/10.1093/nar/25.17.3389

Ansari M, White AD. Serverless prediction of peptide properties with recurrent neural networks. J Chem Inf Model. 2023;63(8):2546–53. https://doi.org/10.1021/acs.jcim.2c01317

Arora A, Patiyal S, Sharma N, Devi NL, Kaur D, Raghava GPS. A random forest model for predicting exosomal proteins using evolutionary information and motifs. Proteomics. 2024;24(6):e2300231. https://doi.org/10.1002/pmic.202300231

Beltrán JF, Herrera-Belén L, Parraguez-Contreras F, Farías JG, Machuca-Sepúlveda J, Short S. MultiToxPred 1.0: a novel comprehensive tool for predicting 27 classes of protein toxins using an ensemble machine learning approach. BMC Bioinform. 2024;25(1):148. https://doi.org/10.1186/s12859-024-05748-z

Bepler T, Berger B. Learning the protein language: evolution, structure, and function. Cell Syst. 2021;12(6):654–669.e3. https://doi.org/10.1016/j.cels.2021.05.017

Bin P, Huang R, Zhou X. Oxidation resistance of the sulfur amino acids: methionine and cysteine. Biomed Res Int. 2017;2017:9584932. https://doi.org/10.1155/2017/9584932

Brandes N, Ofer D, Peleg Y, Rappoport N, Linial M. ProteinBERT: a universal deep-learning model of protein sequence and function. Bioinformatics. 2022;38(8):2102–10. https://doi.org/10.1093/bioinformatics/btac020

Casewell NR, Wüster W, Vonk FJ, Harrison RA, Fry BG. Complex cocktails: the evolutionary novelty of venoms. Trends Ecol Evol. 2013;28(4):219–29. https://doi.org/10.1016/j.tree.2012.10.020

Chaohong S. Prediction of presynaptic and postsynaptic neurotoxins by bi-layer support vector machine with multi-features. Afr J Microbiol Res. 2012;6(6):1354–8. https://doi.org/10.5897/ajmr11.1536

Chaudhary K, Kumar R, Singh S, Tuknait A, Gautam A, Mathur D, et al. A web server and mobile app for computing hemolytic potency of peptides. Sci Rep. 2016;6(1):22843. https://doi.org/10.1038/srep22843

Chu Y, Zhang H, Zhang L. Function prediction of peptide toxins with sequence-based multi-tasking PU learning method. Toxins. 2022;14(11):811. https://doi.org/10.3390/toxins14110811

Cole TJ, Brewer MS. TOXIFY: a deep learning approach to classify animal venom proteins. PeerJ. 2019;7(e7200):e7200. https://doi.org/10.7717/peerj.7200

Devlin J, Chang M-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional Transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis: Association for Computational Linguistics; 2018. p. 4171–86. https://doi.org/10.48550/ARXIV.1810.04805

Dhall A, Patiyal S, Sharma N, Usmani SS, Raghava GPS. Computer-aided prediction and design of IL-6 inducing peptides: IL-6 plays a crucial role in COVID-19. Brief Bioinform. 2021;22(2):936–45. https://doi.org/10.1093/bib/bbaa259

Díaz-Dinamarca DA, Salazar ML, Castillo BN, Manubens A, Vasquez AE, Salazar F, et al. Protein-based adjuvants for vaccines as immunomodulators of the innate and adaptive immune response: current knowledge, challenges, and future opportunities. Pharmaceutics. 2022;14(8):1671. https://doi.org/10.3390/pharmaceutics14081671

Draper SJ, Angov E, Horii T, Miller LH, Srinivasan P, Theisen M, et al. Recent advances in recombinant protein-based malaria vaccines. Vaccine. 2015;33(52):7433–43. https://doi.org/10.1016/j.vaccine.2015.09.093

Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, et al. ProtTrans: toward understanding the language of life through self-supervised learning. IEEE Trans Pattern Anal Mach Intell. 2022;44(10):7112–27. https://doi.org/10.1109/TPAMI.2021.3095381

Erckert K, Rost B. Assessing the role of evolutionary information for enhancing protein language model embeddings. Sci Rep. 2024;14(1):20692. https://doi.org/10.1038/s41598-024-71783-8

Escoubas P, Diochot S, Corzo G. Structure and pharmacology of spider venom neurotoxins. Biochimie. 2000;82(9–10):893–907. https://doi.org/10.1016/s0300-9084(00)01166-4

Fahey RC. Biologically important thiol-disulfide reactions and the role of cyst(E)Ine in proteins: an evolutionary perspective. Advances in experimental medicine and biology. Boston: Springer US; 1977. p. 1–30. https://doi.org/10.1007/978-1-4684-3282-4_1

Fry BG. From genome to "venome": molecular origin and evolution of the snake venom proteome inferred from phylogenetic analysis of toxin sequences and related body proteins. Genome Res. 2005;15(3):403–20. https://doi.org/10.1101/gr.3228405

Fry BG, Roelants K, Champagne DE, Scheib H, Tyndall JDA, King GF, et al. The toxicogenomic multiverse: convergent recruitment of proteins into animal venoms. Annu Rev Genomics Hum Genet. 2009;10(1):483–511. https://doi.org/10.1146/annurev.genom.9.081307.164356

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. https://doi.org/10.7717/peerj-cs.90

Gahlot PS, Choudhury S, Bajiya N, Kumar N, Raghava GPS. Prediction of plant resistance proteins using alignment-based and alignment-free approaches. Proteomics. 2024;25:e202400261. https://doi.org/10.1002/pmic.202400261

Grishin E. Polypeptide neurotoxins from spider venoms. Eur J Biochem. 1999;264(2):276–80. https://doi.org/10.1046/j.1432-1327.1999.00622.x

Guang X-M, Guo Y-Z, Wang X, Li M-L. Prediction of neurotoxins by support vector machine based on multiple feature vectors. Interdiscip Sci Comput Life Sci. 2010;2(3):241–6. https://doi.org/10.1007/s12539-010-0044-7

Guntuboina C, Das A, Mollaei P, Kim S, Barati Farimani A. PeptideBERT: a language model based on transformers for peptide property prediction. J Phys Chem Lett. 2023;14(46):10427–34. https://doi.org/10.1021/acs.jpclett.3c02398

Gupta S, Kapoor P, Chaudhary K, Gautam A, Kumar R, Raghava GPS. Peptide toxicity prediction. Methods Mol Biol. 2015;1268:143–57. https://doi.org/10.1007/978-1-4939-2285-7_7

Huo H, Li T, Wang S, Lv Y, Zuo Y, Yang L. Prediction of presynaptic and postsynaptic neurotoxins by combining various Chou's pseudo components. Sci Rep. 2017;7(1):5827. https://doi.org/10.1038/s41598-017-06195-y

Ito M, Jameson JL, Ito M. Molecular basis of autosomal dominant neurohypophyseal diabetes insipidus. Cellular toxicity caused by the accumulation of mutant vasopressin precursors within the endoplasmic reticulum. J Clin Invest. 1997;99(8):1897–905. https://doi.org/10.1172/JCI119357

Jain S, Gupta S, Patiyal S, Raghava GPS. THPdb2: compilation of FDA approved therapeutic peptides and proteins. Drug Discov Today. 2024;29(7):104047. https://doi.org/10.1016/j.drudis.2024.104047

Jimenez EC, Cruz LJ. Conotoxins as tools in research on nicotinic receptors. Toxins and drug discovery. Dordrecht: Springer Netherlands; 2016. p. 1–17. https://doi.org/10.1007/978-94-007-6726-3_17-2

Johnson IS. Human insulin from recombinant DNA technology. Science. 1983;219(4585):632–7. https://doi.org/10.1126/science.6337396

Kaufmann SHE. Emil von Behring: translational medicine at the dawn of immunology. Nat Rev Immunol. 2017;17(6):341–3. https://doi.org/10.1038/nri.2017.37

Ko Y. S., Parkinson J., & Wang W. 2024. Benchmarking text-integrated protein language model embeddings and embedding fusion on diverse downstream tasks. https://doi.org/10.1101/2024.08.24.609531.

Koua D, Kuhn-Nentwig L. Spider neurotoxins, short linear cationic peptides and venom protein classification improved by an automated competition between exhaustive profile HMM classifiers. Toxins. 2017;9(8):245. https://doi.org/10.3390/toxins9080245

Kumar N, Choudhury S, Bajiya N, Patiyal S, Raghava GPS. Prediction of anti-freezing proteins from their evolutionary profile. Proteomics. 2025;25(3):e202400157. https://doi.org/10.1002/pmic.202400157

Kumar M, Gromiha MM, Raghava GPS. Identification of DNA-binding proteins using support vector machines and evolutionary profiles. BMC Bioinform. 2007;8(1):463. https://doi.org/10.1186/1471-2105-8-463

Kumar V, Kumar R, Agrawal P, Patiyal S, Raghava GPS. A method for predicting hemolytic potency of chemically modified peptides from its structure. Front Pharmacol. 2020;11:54. https://doi.org/10.3389/fphar.2020.00054

Lata S, Sharma BK, Raghava GPS. Analysis and prediction of antibacterial peptides. BMC Bioinform. 2007;8(1):263. https://doi.org/10.1186/1471-2105-8-263

Latham AS, Moreno JA, Geer CE. Biological agents and the aging brain: glial inflammation and neurotoxic signaling. Front Aging. 2023;4:1244149. https://doi.org/10.3389/fragi.2023.1244149

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(22):12291. https://doi.org/10.3390/ijms222212291

Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22(13):1658–9. https://doi.org/10.1093/bioinformatics/btl158

Li P, Zhang H, Zhao X, Jia C, Li F, Song J. Pippin: a random forest-based method for identifying presynaptic and postsynaptic neurotoxins. J Bioinform Comput Biol. 2020;18:2050008. https://doi.org/10.1142/S0219720020500080

Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic level protein structure with a language

model. Science. 2022;379:1123–30. https://doi.org/10.1101/2022.07.20.500902

Liu Y, Tian B. Protein-DNA binding sites prediction based on pre-trained protein language model and contrastive learning. Brief Bioinform. 2023;25(1):bbad488. https://doi.org/10.1093/bib/bbad488

Liu B, Zhang D, Xu R, Xu J, Wang X, Chen Q, et al. Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. Bioinformatics. 2014;30(4):472–9. https://doi.org/10.1093/bioinformatics/btt709

Madani A, Krause B, Greene ER, Subramanian S, Mohr BP, Holton JM, et al. Large language models generate functional protein sequences across diverse families. Nat Biotechnol. 2023;41(8):1099–106. https://doi.org/10.1038/s41587-022-01618-2

Mall R, Singh A, Patel CN, Guirimand G, Castiglione F. VISH-Pred: an ensemble of fine-tuned ESM models for protein toxicity prediction. Brief Bioinform. 2024;25(4):bbae270. https://doi.org/10.1093/bib/bbae270

Marks DS, Hopf TA, Sander C. Protein structure prediction from sequence variation. Nat Biotechnol. 2012;30(11):1072–80. https://doi.org/10.1038/nbt.2419

Mei J, Zhao J. Analysis and prediction of presynaptic and postsynaptic neurotoxins by Chou's general pseudo amino acid composition and motif features. J Theor Biol. 2018;447:147–53. https://doi.org/10.1016/j.jtbi.2018.03.034

Mohammadzadeh-Vardin T, Ghareyazi A, Gharizadeh A, Abbasi K, Rabiee HR. DeepDRA: drug repurposing using multi-omics data integration with autoencoders. PLoS One. 2024;19(7):e0307649. https://doi.org/10.1371/journal.pone.0307649

Morozov V, Rodrigues CHM, Ascher DB. CSM-toxin: a web-server for predicting protein toxicity. Pharmaceutics. 2023;15(2):431. https://doi.org/10.3390/pharmaceutics15020431

Müller AT, Gabernet G, Hiss JA, Schneider G. modlAMP: python for antimicrobial peptides. Bioinformatics. 2017;33(17):2753–5. https://doi.org/10.1093/bioinformatics/btx285

Nagpal G, Chaudhary K, Agrawal P, Raghava GPS. Computer-aided prediction of antigen presenting cell modulators for designing peptide-based vaccine adjuvants. J Transl Med. 2018;16(1):181. https://doi.org/10.1186/s12967-018-1560-1

Nampoothiri M, Reddy ND, John J, Kumar N, Kutty Nampurath G, Rao Chamallamudi M. Insulin blocks glutamate-induced neurotoxicity in differentiated SH-SY5Y neuronal cells. Behav Neurol. 2014;2014:674164. https://doi.org/10.1155/2014/674164

Oh J, Ko M, Lee H. Evaluation for allergenicity for genetically modified organic foods. J Allergy Clin Immunol. 2009;123(2):S244. https://doi.org/10.1016/j.jaci.2008.12.942

Osipov A, Utkin Y. What are the neurotoxins in hemotoxic snake venoms? Int J Mol Sci. 2023;24(3):2919. https://doi.org/10.3390/ijms24032919

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(21):5159–68. https://doi.org/10.1093/bioinformatics/btaa656

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al.2012. Scikit-learn: machine learning in Python. ArXiv [Cs.LG]. https://doi.org/10.48550/ARXIV.1201.0490

Pineda SS, Chin YK-Y, Undheim EAB, Senff S, Mobli M, Dauly C, et al. Structural venomics reveals evolution of a complex venom by duplication and diversification of an ancient peptide-encoding gene. Proc Natl Acad Sci U S A. 2020;117(21):11399–408. https://doi.org/10.1073/pnas.1914536117

Pirazzini M, Rossetto O, Eleopra R, Montecucco C. Botulinum neurotoxins: biology, pharmacology, and toxicology. Pharmacol Rev. 2017;69(2):200–35. https://doi.org/10.1124/pr.116.012658

Plotkin SA. Correlates of protection induced by vaccination. Clin Vaccine Immunol. 2010;17(7):1055–65. https://doi.org/10.1128/CVI.00131-10

Pokharel S, Pratyush P, Heinzinger M, Newman RH, Kc DB. Improving protein succinylation sites prediction using embeddings from protein language model. Sci Rep. 2022;12(1):16933. https://doi.org/10.1038/s41598-022-21366-2

Prasad PL, Rai PL. Prospective study of diphtheria for neurological complications. J Pediatr Neurosci. 2018;13(3):313–6. https://doi.org/10.4103/JPN.JPN_130_17

Rathore AS, Choudhury S, Arora A, Tijare P, Raghava GPS. ToxinPred 3.0: an improved method for predicting the toxicity of peptides. Comput Biol Med. 2024;179(108926):108926. https://doi.org/10.1016/j.compbiomed.2024.108926

Rathore AS, Jain S, Choudhury S, Raghava GPS. NTxPred2: a large language model for predicting neurotoxic peptides and neurotoxins. Zenodo. 2025 https://doi.org/10.5281/ZENODO.15030415

Rathore AS, Kumar N, Choudhury S, Mehta NK, Raghava GPS. Prediction of hemolytic peptides and their hemolytic concentration (HC50). Commun Biol. 2024;8:176. https://doi.org/10.1101/2024.07.23.604887

Rathore AS, Kumar N, Choudhury S, Mehta NK, Raghava GPS. Prediction of hemolytic peptides and their hemolytic concentration. Commun Biol. 2025;8(1):176. https://doi.org/10.1038/s42003-025-07615-w

Razzaghi P, Abbasi K, Ghasemi JB. Multivariate pattern recognition by machine learning methods. Machine learning and pattern recognition methods in chemistry from multivariate and data driven modeling. Amsterdam: Elsevier; 2023. p. 47–72. https://doi.org/10.1016/b978-0-323-90408-7.00002-2

Refahi MS, Mir A, Nasiri JA. A novel fusion based on the evolutionary features for protein fold recognition using support vector machines. Sci Rep. 2020;10(1):14368. https://doi.org/10.1038/s41598-020-71172-x

Romey G, Abita JP, Chicheportiche R, Rochat H, Lazdunski M. Scorpion neurotoxin. Mode of action on neuromuscular junctions and synaptosomes. Biochim Biophys Acta. 1976;448(4):607–19. https://doi.org/10.1016/0005-2736(76)90114-0

Saha S, Raghava GPS. BTXpred: prediction of bacterial toxins. In Silico Biol. 2007a;7(4–5):405–12.

Saha S, Raghava GPS. Prediction of neurotoxins based on their function and source. In Silico Biol. 2007b;7(4–5):369–87.

Salem M, Keshavarzi Arshadi A, Yuan JS. AMPDeep: hemolytic activity prediction of antimicrobial peptides using transfer learning. BMC Bioinform. 2022;23(1):389. https://doi.org/10.1186/s12859-022-04952-z

Sharma A, Kapoor P, Gautam A, Chaudhary K, Kumar R, Chauhan JS, et al. Computational approach for designing tumor homing peptides. Sci Rep. 2013;3(1):1607. https://doi.org/10.1038/srep01607

Sharma N, Naorem LD, Jain S, Raghava GPS. ToxinPred2: an improved method for predicting toxicity of proteins. Brief Bioinform. 2022;23(5):bbac174. https://doi.org/10.1093/bib/bbac174

Sharma N, Patiyal S, Dhall A, Pande A, Arora C, Raghava GPS. AlgPred 2.0: an improved method for predicting allergenic proteins and mapping of IgE epitopes. Brief Bioinform. 2021;22(4):bbaa294. https://doi.org/10.1093/bib/bbaa294

Shi H, Li Y, Chen Y, Qin Y, Tang Y, Zhou X, et al. ToxMVA: an end-to-end multi-view deep autoencoder method for protein toxicity prediction. Comput Biol Med. 2022;151(Pt B):106322. https://doi.org/10.1016/j.compbiomed.2022.106322

Su M-G, Huang C-H, Lee T-Y, Chen Y-J, Wu H-Y. Incorporating amino acids composition and functional domains for identifying bacterial toxin proteins. Biomed Res Int. 2014;2014:972692. https://doi.org/10.1155/2014/972692

Sunagar K, Casewell NR, Varma S, Kolla R, Antunes A, Moran Y. Deadly innovations: unraveling the molecular evolution of animal

venoms. Venom genomics and proteomics. Dordrecht: Springer Netherlands; 2016. p. 1–27. https://doi.org/10.1007/978-94-007-6416-3_27
Tang M, Wu L, Yu X, Chu Z, Jin S, Liu J. Prediction of protein-protein interaction sites based on stratified attentional mechanisms. Front Genet. 2021;12:784863. https://doi.org/10.3389/fgene.2021.784863
Tang H, Yang Y, Zhang C, Chen R, Huang P, Duan C, et al. Predicting presynaptic and postsynaptic neurotoxins by developing feature selection technique. Biomed Res Int. 2017;2017:3267325. https://doi.org/10.1155/2017/3267325
Timmons PB, Hewage CM. HAPPENN is a novel tool for hemolytic activity prediction for therapeutic peptides which employs neural networks. Sci Rep. 2020;10(1):10869. https://doi.org/10.1038/s41598-020-67701-3
Usmani SS, Bedi G, Samuel JS, Singh S, Kalra S, Kumar P, et al. THPdb: database of FDA-approved peptide and protein therapeutics. PLoS One. 2017;12(7):e0181748. https://doi.org/10.1371/journal.pone.0181748
Walsh G. Biopharmaceutical benchmarks 2018. Nat Biotechnol. 2018;36(12):1136–45. https://doi.org/10.1038/nbt.4305
Wan H, Liu Q, Ju Y. Utilize a few features to classify presynaptic and postsynaptic neurotoxins. Comput Biol Med. 2023;152:106380. https://doi.org/10.1016/j.compbiomed.2022.106380
Wang J, Yang B, Revote J, Leier A, Marquez-Lago TT, Webb G, et al. POSSUM: a bioinformatics toolkit for generating numerical sequence feature descriptors based on PSSM profiles. Bioinformatics. 2017;33(17):2756–8. https://doi.org/10.1093/bioinformatics/btx302
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(6):1514–24. https://doi.org/10.1093/bioinformatics/btac006
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(5):bbab041. https://doi.org/10.1093/bib/bbab041
Win TS, Malik AA, Prachayasittikul V, Wikberg S, J. E, Nantasenamat C, et al. HemoPred: a web server for predicting the hemolytic activity of peptides. Future Med Chem. 2017;9(3):275–91. https://doi.org/10.4155/fmc-2016-0188
Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., et al. HuggingFace's transformers: state-of-the-art natural language processing. 2019. https://doi.org/10.48550/ARXIV.1910.03771.
Xu C, Ge L, Zhang Y, Dehmer M, Gutman I. Computational prediction of therapeutic peptides based on graph index. J Biomed Inform. 2017;75:63–9. https://doi.org/10.1016/j.jbi.2017.09.011
Yang L, Li Q. Prediction of presynaptic and postsynaptic neurotoxins by the increment of diversity. Toxicol In Vitro. 2009;23(2):346–8. https://doi.org/10.1016/j.tiv.2008.12.015
Zhang F, Li J, Wen Z, Fang C. FusPB-ESM2: fusion model of ProtBERT and ESM-2 for cell-penetrating peptide prediction. Comput Biol Chem. 2024;111:108098. https://doi.org/10.1016/j.compbiolchem.2024.108098

AUTHOR BIOGRAPHIES

Anand Singh Rathore is currently pursuing a Ph.D. in Computational Biology at the Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India.

Saloni Jain is currently pursuing a Ph.D. in Computational Biology at the Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India.

Shubham Choudhury is currently pursuing a Ph.D. in Computational Biology at the Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India.

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

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Rathore AS, Jain S, Choudhury S, Raghava GPS. A large language model for predicting neurotoxic peptides and neurotoxins. Protein Science. 2025;34(8):e70200. https://doi.org/10.1002/pro.70200