



ToxGIN: an *In silico* prediction model for peptide toxicity via graph isomorphism networks integrating peptide sequence and structure information

Qiule Yu, Zhixing Zhang, Guixia Liu , Weihua Li, Yun Tang *

Shanghai Frontiers Science Center of Optogenetic Techniques for Cell Metabolism, Shanghai Key Laboratory of New Drug Design, School of Pharmacy, East China University of Science and Technology, 130 Meilong Road, Shanghai 200237, China

*Corresponding author. Shanghai Frontiers Science Center of Optogenetic Techniques for Cell Metabolism, Shanghai Key Laboratory of New Drug Design, School of Pharmacy, East China University of Science and Technology, 130 Meilong Road, Shanghai 200237, China. E-mail: ytang234@ecust.edu.cn

Abstract

Peptide drugs have demonstrated enormous potential in treating a variety of diseases, yet toxicity prediction remains a significant challenge in drug development. Existing models for prediction of peptide toxicity largely rely on sequence information and often neglect the three-dimensional (3D) structures of peptides. This study introduced a novel model for short peptide toxicity prediction, named ToxGIN. The model utilizes Graph Isomorphism Network (GIN), integrating the underlying amino acid sequence composition and the 3D structures of peptides. ToxGIN comprises three primary modules: (i) Sequence processing module, converting peptide 3D structures and sequences into information of nodes and edges; (ii) Feature extraction module, utilizing GIN to learn discriminative features from nodes and edges; (iii) Classification module, employing a fully connected classifier for toxicity prediction. ToxGIN performed well on the independent test set with F1 score = 0.83, AUROC = 0.91, and Matthews correlation coefficient = 0.68, better than existing models for prediction of peptide toxicity. These results validated the effectiveness of integrating 3D structural information with sequence data using GIN for peptide toxicity prediction. The proposed ToxGIN and data can be freely accessible at <https://github.com/cihebiyql/ToxGIN>.

Keywords: graph isomorphism network (GIN); peptide toxicity prediction; deep learning; computational toxicology; protein language models; 3D structure of peptides

Introduction

Peptides are short chains formed by amino acids linked through peptide bonds, playing pivotal roles in numerous biological processes and demonstrating substantial therapeutic potentials [1, 2]. As integral components of drug development, peptides exhibit considerable promise in treating refractory diseases, owing to their specificity and selectivity as optimal therapeutic targets [3, 4]. Compared to small molecular drugs, therapeutic peptides offer heightened specificity, efficacy, safety, and reduced immunogenicity. In contrast to biologics, peptides present lower immunogenicity, enhanced membrane permeability, and lower therapy costs [5–7]. However, certain peptides exhibit toxicity, presenting challenges for their uses in drug development [8, 9]. Therefore, accurate prediction of peptide toxicity is crucial for designing safe and effective peptide-based drugs. While peptides share fundamental similarities with proteins, they differ significantly in terms of length and structural rigidity. Peptides are typically composed of fewer than 50 amino acids, making them relatively short, flexible, and variable. This contrasts with proteins, which consist of longer amino acid chains that form more stable and complex three-dimensional (3D) structures. The shorter length and increased flexibility of peptides contribute to their unique

biological functions and therapeutic applications, necessitating specialized approaches for their analysis and toxicity prediction.

Toxicity prediction is a crucial aspect of drug development. Traditional methods rely heavily on experimental validation, which is not only time-consuming and expensive but also plagued by issues of experimental conditions and reproducibility [10]. This challenge becomes particularly evident as the number of potential therapeutic peptides rapidly increases. With the advancements in computational biology and machine learning, computational approaches have emerged as powerful tools. These methods can generally be categorized into two types: similarity-based methods and machine learning-based methods. Similarity-based methods use alignment tools to measure local and global sequence similarities, such as BLAST [11] and BLAST-score [11]. However, these methods have several drawbacks: they require the target peptide to have homologous toxic ones; and their performance degrades significantly when handling large datasets. Hence, they necessitate setting an e-value cutoff and an arbitrary sequence similarity threshold, which can affect prediction accuracy. In contrast, machine learning-based methods focus on using manually extracted protein sequence features and

Received: August 14, 2024. Revised: October 22, 2024. Accepted: October 29, 2024

© The Author(s) 2024. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

For commercial re-use, please contact journals.permissions@oup.com

positive–negative samples to predict peptide toxicity. For instance, in 2009, Naamati et al. developed ClanTox [12], a machine learning predictor for animal toxins. It distinguishes toxic peptides from non-toxic ones using features extracted from protein sequences, trained with boosted stump classifiers. In 2013, Gupta et al. introduced ToxinPred [13], an *in silico* method using machine learning to predict peptide and protein toxicity. It employs sequence-based feature descriptors, such as amino acid composition (AAC), dipeptide composition, and sequential motifs.

Recently, deep learning methods have been increasingly applied in protein toxicity prediction. In 2021, ToxDL [14] encoded protein sequences using convolutional neural networks and employed Skip-gram [15] to generate domain embeddings for toxicity prediction. In 2022, ToxMVA [16] was introduced as a novel end-to-end deep learning architecture that integrates sequence-based features through an autoencoder network, combining sequence, physicochemical, and contextual semantic information into discriminative latent representations. In 2023, Morozov et al. developed CSM-Toxin [17], a tool that relies entirely on the primary protein sequence by utilizing a model adapted from ProteinBERT [18], a deep learning language model originally developed to understand protein sequences in a manner analogous to natural language processing. In 2024, VISH-Pred [19] was introduced as an ensemble model that combines a fine-tuned Transformer model with LightGBM [20] and XGBoost [21], effectively addressing class imbalance issues and enhancing prediction accuracy.

Deep learning methods have made significant advances in predicting toxicity of therapeutic peptides. For example, ATSE [22], the first deep learning model dedicated to peptide toxicity, utilizes molecular graph representations extracted by RDKit [23] and evolutionary information represented by position-specific scoring matrixes (PSSMs) [24] for peptide toxicity prediction. However, this model tends to perform poorly for peptides lacking homologous sequences. ToxIBTL [25] employs the BLOSUM62 (BLOcks SUbstitution Matrix) [26] and the FEFS (Feature Extraction based on Graphical and Statistical features) [27] to extract sequence features of peptides. CAPTP [28] introduces Transformer, leveraging convolution and self-attention mechanisms to enhance peptide toxicity prediction from amino acid sequences. Nevertheless, these models only extract features from protein sequences and do not consider the 3D structures of proteins, which is a crucial determinant of protein properties [29].

In 2018, the emergence of AlphaFold [30] revolutionized protein design research by enabling the prediction of 3D structures solely from protein sequences. The transformative advancements in Natural Language Processing have been applied to protein sequence analysis, with the advent of protein language models offering a novel approach to extracting sequence features [31]. Hence, in this study, we proposed a novel model based on Graph Isomorphism Networks (GIN) [32] to predict short peptide toxicity. Initially, we represented the 3D structures of peptides predicted by ColabFold [33] as graphs, with amino acid residues and their interactions serving as nodes and edges, respectively. Next, to leverage the capabilities of the ESM2 protein language model [34], we extracted deep biological features from peptide sequences and further enriched the feature representation of each amino acid node with physicochemical properties. Subsequently, GIN aggregated information from neighboring nodes to extract local and global features, followed by nonlinear transformation to output toxicity prediction probabilities. This approach effectively captured complex amino acid interactions, and

Table 1. Overview of the benchmark datasets.

| Dataset | Number of positives | Number of negatives |
|--------------|---------------------|---------------------|
| Training set | 1932 | 1932 |
| Testing set | 282 | 282 |

significantly enhanced the accuracy and robustness of toxicity predictions. The proposed ToxGIN and data could be freely accessible at <https://github.com/cihebiyql/ToxGIN>.

Methods and materials

Data collection and preparation

To ensure consistent model comparisons, we adopted the data collection methodology used in ATSE [22] and ToxIBTL [25]. Our dataset comprises both training and test sets, each containing toxic and non-toxic peptide sequences ranging from 10 to 50 amino acids in length. For the training set, we used the same set of 1932 toxic peptide sequences and 1932 non-toxic peptide sequences as employed by ATSE and ToxIBTL. For the test set, we collected additional toxic peptide sequences from three public databases.

UniProt [35]. Retrieved 1777 toxic peptide sequences using the keywords ‘KW-0800 AND (reviewed) AND (length: [10 TO 50])’.

ConoServer [36]. Collected 706 toxic peptide sequences related to conopeptides.

ArachnoServer [37]. Obtained 271 toxic peptide sequences from spider venom toxins.

After merging and removing duplicates, we obtained a total of 2400 unique toxic peptide sequences. Excluding the 1932 sequences used in the training set, we had 468 toxic sequences remaining for test. We then applied CD-HIT [38] with a 90% sequence identity threshold to remove highly similar sequences, resulting in 282 toxic peptide sequences for the positive test set. For non-toxic peptides, we retrieved 10,484 sequences from UniProt using the keywords ‘NOT KW-0800 AND NOT KW-0020 AND (reviewed) AND (length: [10 TO 50])’. After excluding the 1932 non-toxic sequences used in the training set and applying CD-HIT at a 90% similarity threshold, we obtained 2886 non-toxic sequences. From this set, we randomly selected 282 sequences to serve as the negative test set.

By balancing the test set with an equal number of positive and negative samples (282 each), we ensured that the evaluation metrics would not be biased due to class imbalance. Applying CD-HIT with a 90% sequence identity threshold reduced redundancy and minimized potential prediction bias due to highly similar sequences. A summary of the datasets used in this study is presented in Table 1. A comprehensive table was provided in the Supplementary Materials (Supplementary Tables S1).

Overview of the proposed ToxGIN

Our ToxGIN model architecture, as illustrated in Fig. 1, comprises three modules: (i) sequence processing module, (ii) feature extraction module, and (iii) classification module. In the first module, we depicted the 3D structures of peptides predicted by ColabFold [33] as nodes and edges. Leveraging ESM2 [34], we extracted profound biological features from peptide sequences, which enhanced the feature representation of each amino acid node with their respective physicochemical properties. The second module employed GIN [32] to aggregate information

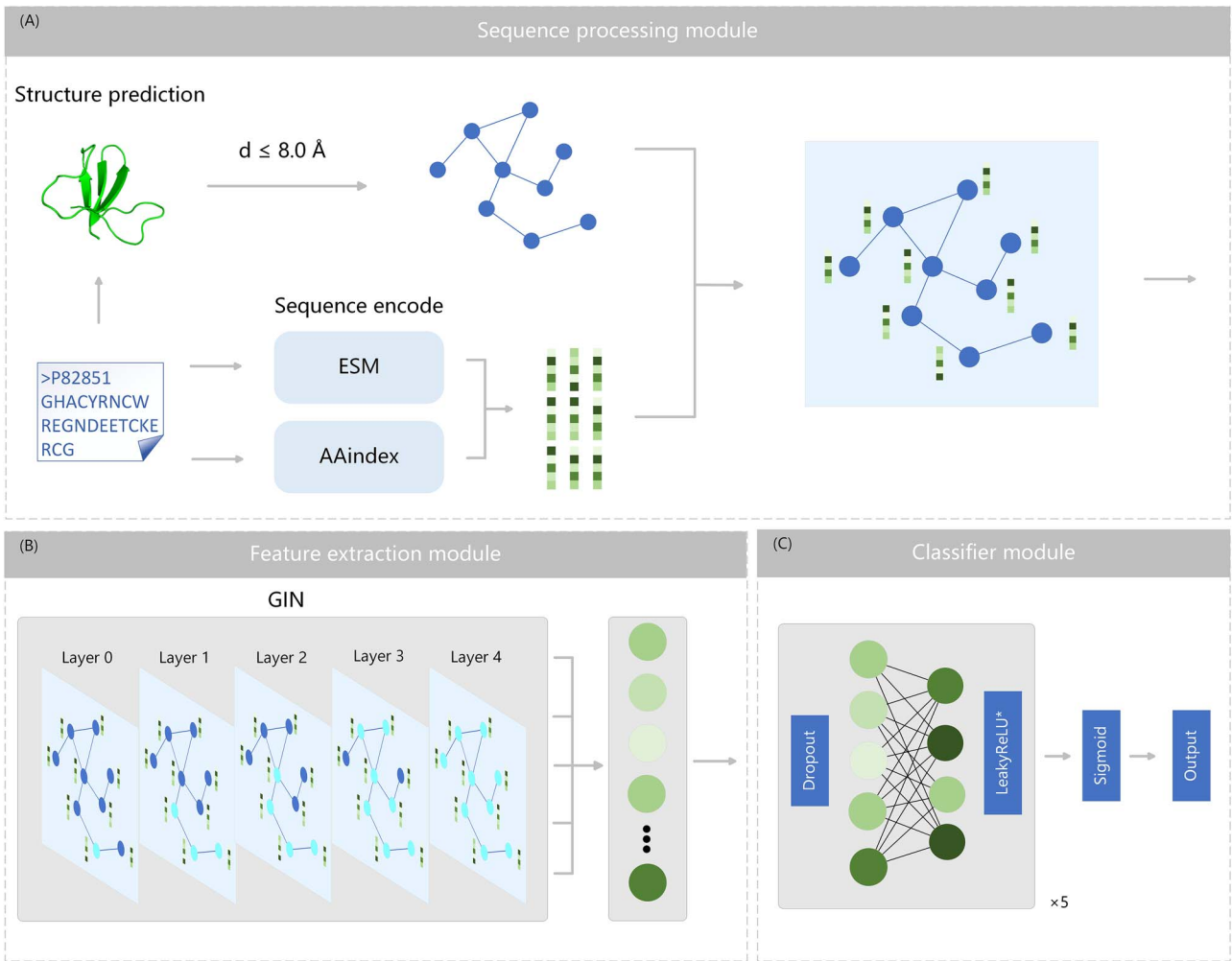


Figure 1. Flowchart of the proposed ToxGIN model, comprising three main modules: (A) sequence processing module, which extracts information about nodes and edges from the peptide’s 3D structure and sequence; (B) feature extraction module, which employs GIN to aggregate information from neighboring nodes and edges; and (C) classification module, which uses these features to generate the toxicity probability.

from neighboring nodes, which could effectively extract both local and global features. These features were subsequently transformed from node-level to graph-level representations, capturing richer, higher-dimensional information. The third module, the classification module, integrated four fully connected layers and a sigmoid layer to predict the toxicity probability of a given peptide. Detailed explanation of each module was provided below.

Sequence processing module

Peptide 3D structure feature extraction

In this study, we utilized version 1.5 of local ColabFold [33] to predict the 3D structures of peptide sequences. To enhance our understanding of amino acid interactions, we transformed structural data into a graph representation. Central to this approach is the construction of a graph’s adjacency matrix based on a distance threshold, which delineates nodes and edges. Initially, we computed the coordinate differences for each pair of amino acids in terms of their 3D positions. Let $(r_i = (x_i, y_i, z_i))$ denote the coordinates of amino acid i , and $(r_j = (x_j, y_j, z_j))$ for amino acid j , the difference (Δr_{ij}) is defined as:

$$\Delta r_{ij} = r_i - r_j = (x_i - x_j, y_i - y_j, z_i - z_j) \quad (1)$$

Next, we computed the Euclidean distance d_{ij} between every pair of amino acids in the peptide:

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \quad (2)$$

By setting a distance threshold θ of 8.0 Å, we generate an adjacency matrix (A). If the distance between two amino acids is less than the threshold, an edge is established between them. The elements of the adjacency matrix (A_{ij}) are defined as follows:

$$A_{ij} = \begin{cases} 1 & \text{if } d_{ij} < \theta \\ 0 & \text{if } d_{ij} \geq \theta \end{cases} \quad (3)$$

Each element in the adjacency matrix indicates whether there is an edge between the corresponding pair of amino acids. By applying this threshold to the distances, we determine which pairs of amino acids interact. This graph representation method effectively captures the complex spatial relationships between amino acids in the 3D structure of the peptide.

Feature extraction using ESM2

Protein sequences carry essential information about biological structure and function across evolutionary scales. This is because

protein's biological properties limit sequence mutations, thereby preserving patterns of structural and functional insights that artificial intelligence can reveal [31]. ESM2 [34] serves as an advanced tool for extracting features from extensive protein sequence datasets using deep learning techniques. ESM2 uses an advanced transformer architecture and Masked Language Modeling to randomly mask amino acid segments in sequences. It predicts these masked positions using contextual clues, capturing intricate sequence dependencies. In our study, we applied the encoder component of ESM2 to extract features from amino acids in short peptide sequences. By inputting peptide sequences into pretrained ESM2, we generated high-dimensional feature vectors containing profound characteristics of peptide sequences, which are effective inputs for subsequent toxicity prediction models.

Derivation of physicochemical properties from the AAindex database

AAindex [39] is a numerical database that catalogs various physicochemical and biochemical properties of amino acids and their pairs. From AAindex version 9.2, we extracted 566 physicochemical properties, each represented by a set of 20 numerical values corresponding to the 20 standard amino acids. It's noteworthy that we focused on 553 properties devoid of NaN values. In contrast to other models using this database, such as ToxMVA [16], our study did not average the 553 AAindex descriptors across each amino acid in the sequence to derive a feature representation for each sequence. Instead, we enhance the feature set of each amino acid node by incorporating the physicochemical properties from AAindex.

Feature extraction module

In this study, we utilized the GIN [32] as the foundational model component for representing the graph structure of short peptide sequences. The process comprised three main steps: node feature extraction, node feature assignment, and graph-level feature aggregation.

Node feature extraction

Assuming the graph \mathbf{G} has \mathbf{N} nodes, and each node v has a feature vector h_v , the node feature representation $h_v^{(k)}$ computed through the GIN at the k -th layer can be expressed as:

$$h_v^{(k)} = \text{MLP}^{(k)} \left((1 + \epsilon^{(k)}) \cdot h_v^{(k-1)} + \sum_{u \in N(v)} h_u^{(k-1)} \right) \quad (4)$$

where $\text{MLP}^{(k)}$ is a multi-layer perceptron, $\epsilon^{(k)}$ is a learnable or fixed scalar parameter, and $u \in N(v)$ represents the set of neighbors of node v . This can be divided into two parts: $(1 + \epsilon^{(k)}) \cdot h_v^{(k-1)}$ is the self-loop term, representing the features of node v itself, adjusted by multiplying $1 + \epsilon^{(k)}$ to modify its influence; $\sum_{u \in N(v)} h_u^{(k-1)}$ is the neighbor aggregation term, representing the sum of features of all neighboring nodes of v . The aggregated features undergo a nonlinear transformation through MLP.

The updated node features obtained through MLP are accumulated with each layer of GIN convolution, allowing the node's feature representation to progressively gather more neighbor and graph structure information. The first layer mainly focuses on the node's direct neighbors (1-hop). The second layer aggregates the neighbors' neighbors (2-hop). Similarly, the k -th layer aggregates the k -hop neighbors' information. By concatenating the features extracted from each layer, different levels of structural information could be retained, forming a richer and more discriminative

feature representation:

$$h_v = \text{Concat} \left(h_v^{(0)}, h_v^{(1)}, \dots, h_v^{(k)} \right) \quad (5)$$

where $h_v^{(k)}$ represents the feature representation of node v at the k -th layer, and Concat denotes concatenation along the feature dimension. The final feature representation h_v of node v contains all the information from layer 0 to layer k .

Node feature assignment

Each layer of GIN convolution extracts different levels of graph structural information, and the node features h_v computed by GIN are assigned to the original graph's node data h_N :

$$H = \{h_1, h_2, h_3, \dots, h_N\} \quad (6)$$

Graph-level feature aggregation

We use a global aggregation operation (such as summation) to aggregate all node features into graph-level features h_G :

$$h_G = \sum_{i=1}^N h_i \quad (7)$$

Classification module

The classification module aims to translate the graph-level feature representation into the final toxicity prediction outcome. This crucial step employs primarily fully connected layers and a *Sigmoid* activation function. Initially, the aggregated graph-level feature h_G is inputted into multiple fully connected layers with *LeakyRelu* activation. Dropout techniques are then applied after each fully connected layer to prevent overfitting. Following this, classification is performed using a *Sigmoid* layer, defined as:

$$z = \text{LeakyRelu}(wh_G + b) \quad (8)$$

$$\hat{y} = \text{sigmoid}(z) = \frac{1}{1 + e^{-z}} \quad (9)$$

where w represents the weights of the fully connected layer, and b represents the corresponding biases. The output value is a probability between 0 and 1. If the probability value is >0.5 , the sequence is classified as the toxic peptide class, and vice versa.

Evaluation metrics

To comprehensively evaluate the performance of our proposed model, we employed several key metrics, including Sensitivity (SE), AUROC, AUPRC, F1 Score, and Matthews Correlation Coefficient (MCC). The definitions of these metrics are as follows:

$$SE = \frac{TP}{TP + FN} \quad (10)$$

$$F1 = 2 \times \frac{PPV \times SE}{PPV + SE} = \frac{2 \times TP}{2 \times TP + FP + FN} \quad (11)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \quad (12)$$

Among these metrics, TP (True Positive) and TN (True Negative) represent the number of correctly predicted positive and negative cases, respectively. FP (False Positive) and FN (False Negative) represent the number of incorrectly predicted positive and negative

cases, respectively. SE measure the classifier's ability to predict positive cases, while MCC and F1 evaluate the overall predictive performance of the classifier. AUROC provides an aggregate measure of performance by calculating the area under the curve. AUPRC focuses on the trade-off between precision and recall, providing insights into the model's performance with respect to positive class predictions.

Model effectiveness evaluation

To evaluate the effectiveness of ESM2 and AAindex for extracting sequence features, we compared them with four typical hand-crafted features: Amino Acid Composition (AAC), Adaptive Skip Dipeptide Composition (ASDC), Pseudo-amino Acid Composition (Pse-AAC) [40], and Amphiphilic Pseudo Amino Acid Composition (APAAC) [41]. We employed six classifiers: Random Forest (RF), Support Vector Machine with Radial Basis Function kernel (SVM-RBF), Gaussian Naive Bayes (GNB), LightGBM [20], Logistic Regression (LR), and K-Nearest Neighbors (KNN). We conducted experiments using ten-fold cross-validation on the training dataset and evaluated them on the test set. To visually demonstrate the discriminative power of features extracted by ToxGIN compared to hand-crafted features, we used t-Distributed Stochastic Neighbor Embedding (t-SNE) [42]. Additionally, we performed ablation studies to understand the contribution of each component of the ToxGIN model to its overall performance. This involved systematically removing one component at a time and observing its impact on the model's performance. Furthermore, we investigated the impact of replacing the GIN component with other Graph Neural Network (GNN) [43] architectures, including Graph Convolutional Network (GCN) [44], Graph Attention Network (GAT) [45], Topology Adaptive Graph Convolutional Network (TAG) [46], Approximate Personalized Propagation of Neural Predictions (APPNP) [47], and GraphSAGE [48], under the same experimental conditions.

Comparison with other models

To evaluate the effectiveness of our proposed model, we compared it with other peptide toxicity prediction models, including ATSE [22], ToxIBTL [25], CAPTP [28], tAMPer [49], and an optimized version of ToxIBTL, referred to as Toxicity-vib [50].

- 1) ATSE. ATSE converts peptide sequences into molecular and evolutionary graphs, learning distinguishing features from both graph structural information and evolutionary information. It employs an optimized attention mechanism to determine whether a peptide is toxic or non-toxic.
- 2) ToxIBTL. ToxIBTL learns features from evolutionary, graphical and statistical information. It combines the information bottleneck principle with transfer learning technique, initially pre-training the model on a protein dataset and then fine-tuning it on a short peptide dataset, transferring the knowledge acquired from proteins to peptides.
- 3) CAPTP. CAPTP is an end-to-end model that integrates a novel encoder combining convolutional modulation and self-attention. This design allows it to automatically learn representations of peptide sequences using only the amino acid sequence as input, facilitating the prediction of peptide toxicity.
- 4) tAMPer. tAMPer utilizes Bi-directional Gated Recurrent Units (Bi-GRUs) to capture sequence features and Geometric Vector Perceptron (GVP) [51] to handle geometric and vector features. It enhances its expressive power and predictive performance by integrating these features using a multi-head attention mechanism.

- 5) Toxicity-vib. The model fine-tunes ToxIBTL by integrating an attention mechanism into its original architecture. This addition aims to enhance feature extraction capabilities while reducing the dimensionality of feature vectors.

Analysis of computational complexity

To comprehensively evaluate ToxGIN, we conducted an experimental analysis of its time and space complexity in comparison with other peptide toxicity prediction models. The models considered include ATSE [22], ToxIBTL [25], CAPTP [28], and tAMPer [49]. We measured metrics such as preprocessing time, training time per epoch, total training time, memory consumption (CPU and GPU), model size, and the total number of parameters. All experiments were conducted using a consistent hardware configuration. For training, we used 3864 sequences with an 80:20 train-validation split, and for testing, we used 564 sequences. Detailed results of the computational complexity analysis are provided in the Supplementary Materials (Supplementary Tables S2 and S3).

Results

Comparison of sequence features

To assess the effectiveness of ESM2 and AAindex for extracting sequence features, we compared the features generated by them with four typical hand-crafted features.

Figure 2 presents the comparison results of the six classifiers on four metrics (SE, AUROC, AUPRC, F1, MCC). The results indicate that the features generated by ESM2 outperform both the AAindex and the four typical hand-crafted features across most metrics for all classifiers. The hand-crafted features generally showed lower performance compared to the ESM2 and AAindex.

To visually illustrate the discriminative power of ESM2 and AAindex, we employed t-SNE [42] on the training dataset to reduce all six feature sets to two dimensions for visualization. Figure 3 shows that compared to the four handcrafted features, features extracted by ESM2 and AAindex more effectively cluster positive and negative samples, thereby reducing sample overlap.

Ablation study of ToxGIN

The exceptional performance of ToxGIN hinges on extracting comprehensive features from peptide sequences and their 3D structures. Specifically, the model leverages protein language models (ESM2) [34] for sequence feature extraction and incorporates physicochemical properties of amino acids from AAindex [39]. To assess the impact of each component on overall performance, we conducted ablation studies by systematically removing or modifying components of the model. Furthermore, we investigated how variations in parameters of the protein language model influence feature extraction and performance. Table 2 shows the ESM2 with different parameters. In this study, we used `esm2_t36_3B_UR50D`. The variants of ToxGIN evaluated in the ablation study are as follows.

- ToxGIN (ESM2_t36). Incorporates all components, including 3D structural information, ESM2-derived sequence features, and AAindex physicochemical properties.
- ToxGIN without 3D structures (w/o Structures). Excludes 3D structural information, utilizing only sequence features.
- ToxGIN without ESM2 (w/o ESM2). Omits features extracted by the ESM2 language model.

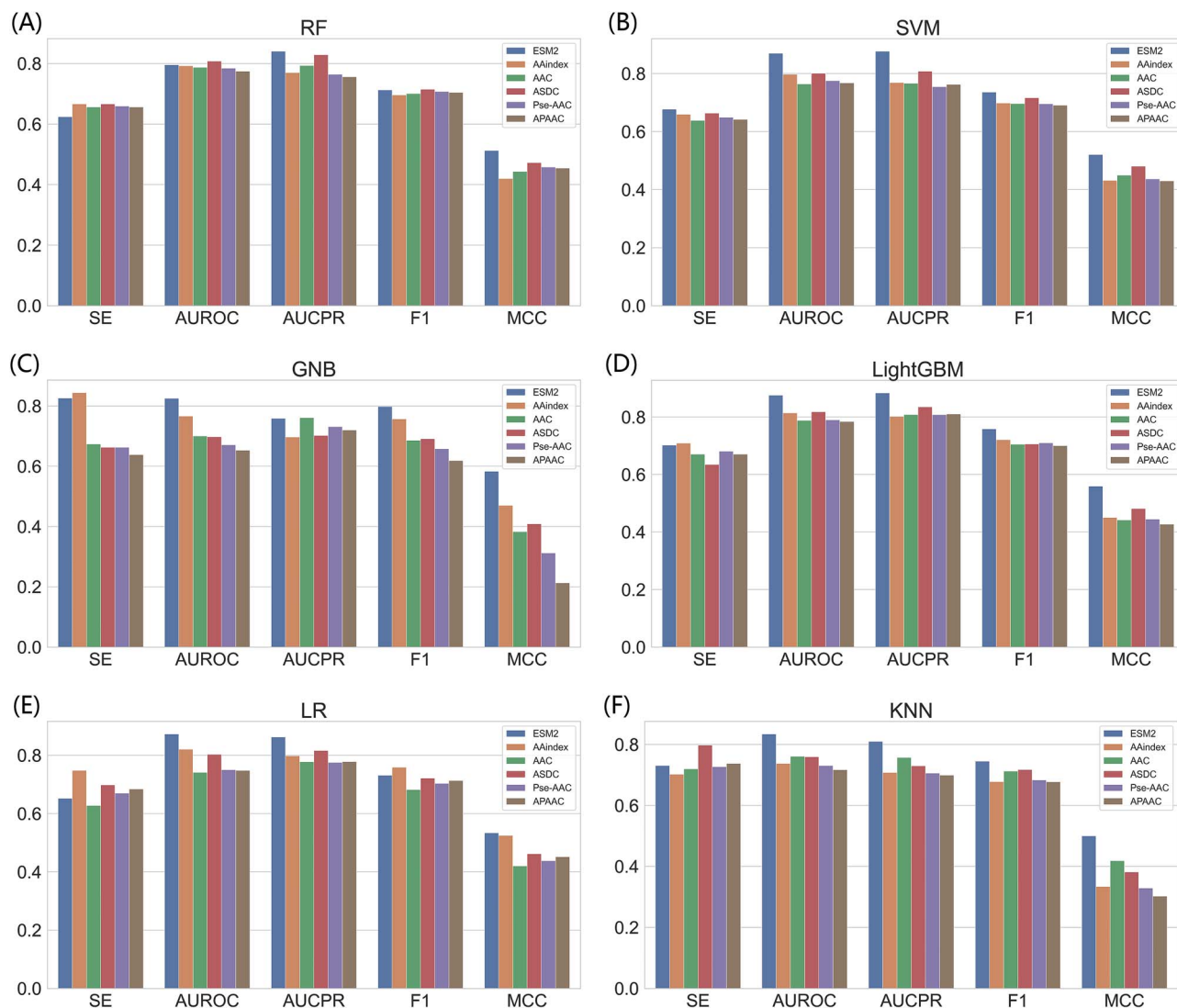


Figure 2. The ten-fold cross validation results of ESM2, AAindex, AAC, ASDC, Pse-AAC and APAAC are based on the six basic classifiers. (A) Results based on RF. (B) Results based on SVM. (C) Results based on GNB. (D) Results based on LightGBM. (E) Results based on LR. (F) Results based on KNN.

Table 2. The different parameters of protein language models.

| Checkpoint name | Num layers | Num parameters |
|---------------------|------------|----------------|
| ESM2_t36_3B_UR50D | 36 | 3B |
| ESM2_t33_650M_UR50D | 33 | 650 M |
| ESM2_t30_150M_UR50D | 30 | 150 M |
| ESM2_t12_35M_UR50D | 12 | 35 M |

- ToxGIN without AAindex (w/o AAindex). Excludes physico-chemical properties from AAindex.
- ToxGIN with different ESM2 models. Employs ESM2 models of varying sizes (ESM2_t33, ESM2_t30, ESM2_t12) to assess the effect of model complexity.

As illustrated in Fig. 4, ToxGIN consistently outperforms its variants across all evaluation metrics. Notably, the removal of 3D structural information results in a significant decrease in performance compared to the full model. Similarly, using smaller ESM2 models leads to reduced effectiveness, indicating the importance of model capacity in capturing sequence features.

Evaluation of GIN component variants

To evaluate the significance of the GIN [32] module in the ToxGIN model, we conducted experiments where we replaced GIN with other GNN [43] architectures, specifically GCN [44], GAT [45], TAG [46], APPNP [47], and GraphSAGE [48]. All other components and settings of the model were kept unchanged to isolate the effect of the GNN architecture on prediction performance. The performance metrics of these model variants are summarized in Table 3.

As shown in Table 3, ToxGIN, utilizing the GIN architecture, achieves the highest SE of 0.8014, AUROC of 0.9172, F1 Score of 0.8354, and MCC of 0.6866 among all the models tested. When GIN is replaced with other GNN architectures, there is a noticeable decline in performance. For instance, replacing GIN with GCN results in a decrease in SE to 0.7577, AUROC to 0.9082, F1 Score to 0.8095, and MCC to 0.6487. Similarly, using GAT leads to further reduction in SE to 0.7189 and MCC to 0.6132, although AUPRC slightly increases to 0.8983. TAG, APPNP, and GraphSAGE also exhibit lower performance compared to ToxGIN, with SE ranging from 0.7538 to 0.7790 and MCC ranging from 0.6185 to 0.6487. Additionally, we conducted interpretability analysis on the

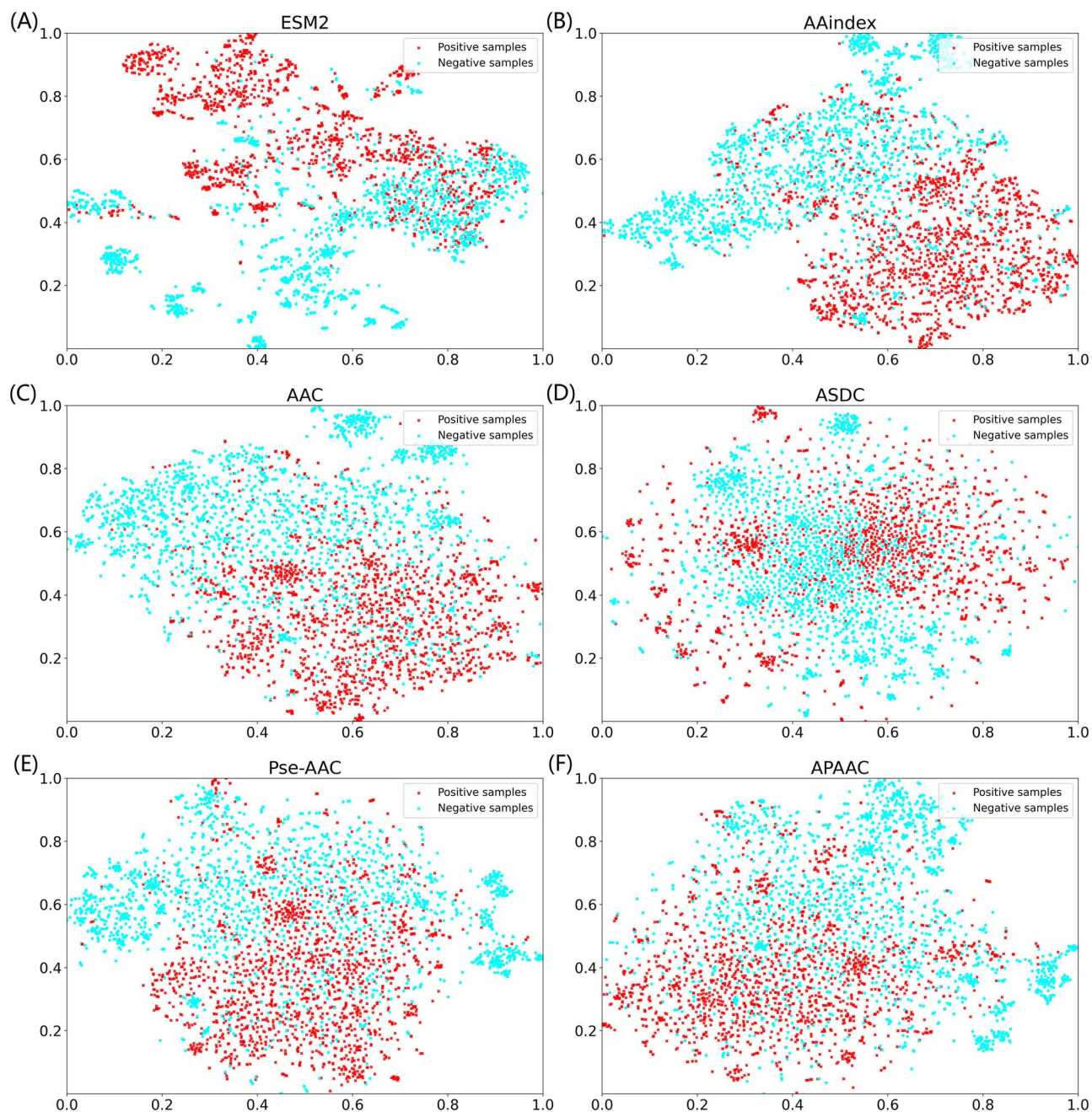


Figure 3. Feature visualization of ESM2, AAindex and other four hand-crafted features. (A) Feature visualization of ESM2. (B) Feature visualization of AAindex. (C) Feature visualization of AAC. (D) Feature visualization of ASDC. (E) Feature visualization of Pse-AAC. (F) Feature visualization of APAAC.

Table 3. Performance comparison of ToxGIN with different GNN architectures.

| Model | SE | AUROC | AUPRC | F1 | MCC |
|-----------|--------|--------|--------|--------|--------|
| ToxGIN | 0.8014 | 0.9172 | 0.89 | 0.8354 | 0.6866 |
| GCN | 0.7577 | 0.9082 | 0.8807 | 0.8095 | 0.6487 |
| GAT | 0.7189 | 0.9058 | 0.8983 | 0.7843 | 0.6132 |
| TAG | 0.7655 | 0.9011 | 0.8714 | 0.7995 | 0.6185 |
| APNP | 0.779 | 0.8952 | 0.8731 | 0.8048 | 0.6233 |
| GraphSAGE | 0.7538 | 0.9079 | 0.8788 | 0.802 | 0.6324 |

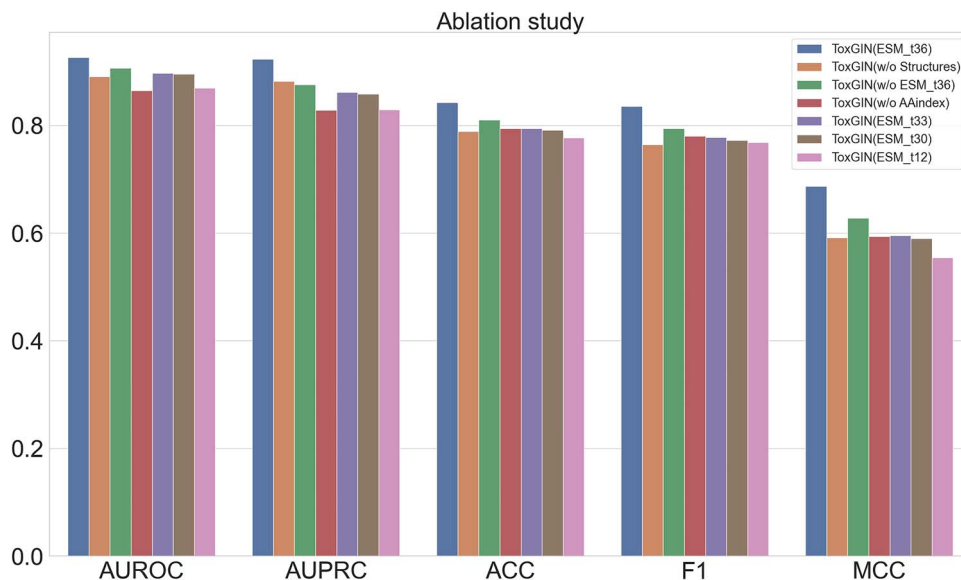


Figure 4. Performance comparison of ToxGIN and its variants across evaluation metrics. The figure displays the effectiveness of the full model (ToxGIN with ESM2_t36) alongside its variants: Excluding 3D structures (w/o structures), ESM2 features (w/o ESM2), and AAindex properties (w/o AAindex), as well as results from various ESM2 model sizes (ESM2_t33, ESM2_t30, ESM2_t12).

GIN model's predictions using GNNExplainer [52], as detailed in [Supplementary Section S9](#).

Comparison with existing models

To ensure a fair comparison, we utilized the same peptide sequence training data and test sets. Due to the unavailability of the online platforms for ATSE [22] and ToxIBTL [25], we replicated these models using their provided data and code, selecting the best-performing models to compare with our own. Although CAPTP [28] and tAMPer [49] offered pre-trained models trained on different datasets, we reproduced and evaluated these models using identical training and test datasets for consistency. It is noteworthy that all models were optimized using the authors' recommended parameters for peptides of identical lengths and significant dataset overlap. For Toxicity-vib, which provided a pre-trained model trained on identical data, we directly utilized this model. Further details are available in the Supplementary Materials ([Supplementary Tables S4, S5, S6, S7, and S8](#)).

According to the data in [Table 4](#), ToxGIN demonstrates outstanding performance across multiple evaluation metrics. Specifically, ToxGIN achieves an AUROC of 0.9172 and an AUPRC of 0.8900, indicating superior model performance over the entire range of classification thresholds. Compared to ToxIBTL, ToxGIN shows improvements of 5.56% in AUROC and 3.50% in AUPRC. When compared to tAMPer, ToxGIN improves AUROC by 4.30% and achieves comparable AUPRC. In threshold-dependent metrics, ToxGIN outperforms other models in SE, F1 Score, and MCC. Specifically, ToxGIN achieves an SE of 0.8014, F1 Score of 0.8354, and MCC of 0.6866, showing improvements of 10.23%, 4.61%, and 13.36%, respectively, compared to ToxIBTL. Compared to tAMPer, ToxGIN shows improvements of 12.04% in SE, 6.41% in F1 Score, and 10.94% in MCC.

Discussion

In this study, we proposed ToxGIN, a prediction model for peptide toxicity. Unlike existing models that solely extracted features from peptide sequences, ToxGIN advanced peptide toxicity prediction

by integrating 3D structural information. This approach transformed peptide structures into graph representations, incorporating GIN to integrate structural insights with sequence data, thereby improving predictive accuracy and robustness. Additionally, the model employed ESM2 [34] to extract deep biological features from peptide sequences and incorporated the physicochemical properties of amino acids, enhancing the understanding of peptide sequences and their toxic properties.

To validate the innovation of ToxGIN, we conducted both horizontal (against similar deep learning models) and vertical (against traditional methods) comparisons. In horizontal comparisons with models like ATSE [22], ToxIBTL [25], and CAPTP [28], which only extracted features from protein sequences, ToxGIN demonstrated superior performance. Using GIN, the model aggregated information from both local (node-level) and global (graph-level) perspectives, ensuring comprehensive feature extraction. This multi-layered approach enhanced the model's ability to distinguish subtle differences between toxic and non-toxic peptides, surpassing the predictive accuracy of previous benchmarks.

In comparisons against traditional methods, ESM2 outperformed better than the AAindex and the four hand-crafted features. Despite the AAindex exhibiting only moderate performance when used alone, we still included them as node features in our model. Because AAindex provides a comprehensive set of physicochemical properties that are essential for understanding peptide behavior and interactions. These properties offer complementary information to the sequence features captured by ESM2, enabling the model to consider both the structural and biochemical aspects of peptides.

Our ablation studies further demonstrated the impact of 3D structural information on model performance. The results revealed that the inclusion of 3D structural data substantially improved predictive accuracy, highlighting the importance of integrating multi-view information for accurate peptide toxicity prediction. Moreover, we investigated the importance of the GIN module within ToxGIN by conducting experiments where GIN was replaced with GCN, GAT, TAG, APPNP, and GraphSAGE. To further elucidate why GIN outperforms other GNN algorithms in handling

Table 4. Comparison of the proposed ToxGIN with existing models.

| Models | SE | AUROC | AUPRC | F1 | MCC |
|--------------|--------|--------|--------|--------|--------|
| ATSE | 0.4716 | 0.4734 | 0.4898 | 0.4787 | 0.0536 |
| ToxIBTL | 0.727 | 0.8617 | 0.86 | 0.7986 | 0.6057 |
| CAPTP | 0.6348 | 0.7592 | 0.7721 | 0.7178 | 0.4462 |
| tAMPer | 0.7153 | 0.8742 | 0.8919 | 0.7851 | 0.6189 |
| Toxicity-vib | 0.6809 | 0.8125 | 0.8025 | 0.7505 | 0.5087 |
| ToxGIN | 0.8014 | 0.9172 | 0.89 | 0.8354 | 0.6866 |

Note: To account for variability during deep learning model training, reported results represent the best outcomes from 20 independent runs.

peptide structural information, we employed GNNExplainer [52] for interpretability analysis, as detailed in [Supplementary Section S9](#). Our analysis, illustrated through [Figs. S1 to S5](#), revealed that GIN identified key amino acid residues critical for peptide bioactivity with higher importance weights compared to GAT and GCN. This demonstrates GIN not only achieves higher predictive accuracy but also provides a more interpretable framework for understanding the molecular determinants of peptide toxicity. Finally, the computational complexity analysis demonstrated that while ToxGIN requires more preprocessing time and has a larger model size than some existing methods, including tAMPer, the increased computational resources are justified by its enhanced predictive accuracy.

We have noted similarities between tAMPer [49] and our approach in integrating peptide amino acid sequences and 3D structures for toxicity prediction. tAMPer utilizes GRUs (Gated Recurrent Units) to process sequence information extracted by ESM2 and employs the GVP [51] to handle the 3D structures of peptides, capturing geometric and vector features. By integrating sequence and structural features through a multi-head attention mechanism, tAMPer generates predictions of peptide toxicity. Despite tAMPer's simpler model architecture and lower computational complexity, ToxGIN outperforms it on the same training and testing datasets. This suggests that ToxGIN's approach of utilizing GIN for graph representations more effectively captures structural information, leading to improved predictive performance.

Conclusions

Accurate prediction of peptide toxicity is crucial for discovery and development of peptide-based drugs. Traditional toxicity prediction methods often rely on time-consuming and expensive experimental validation. With the advancements in computational biology and machine learning, computational methods have emerged as a powerful tool. However, existing peptide toxicity prediction models are primarily based on sequence information and do not consider the 3D structure of peptides. In this study, we introduced a novel peptide toxicity prediction model, ToxGIN, based on GIN. We represented the 3D structures of peptides as graphs, utilizing protein language models to extract deep biological features from peptide sequences, and integrated the physicochemical properties of amino acids to further enrich the feature representation of each amino acid node. ToxGIN uses GIN to aggregate information from node neighbors to extract local and global features and performs nonlinear transformation through MLPs, outputting toxicity prediction probabilities. Experimental results on the same training and test sets show that ToxGIN performs exceptionally well on multiple key performance indicators, surpassing existing advanced peptide toxicity prediction models. The performance of our model validates the effectiveness of combining protein

structural predict model with language model and highlights the importance of integrating 3D structural information in toxicity prediction. We hope that ToxGIN can offer an alternative method for addressing other biological challenges and contribute to advancements in computational peptide toxicity prediction.

Key Points

- In this study, we proposed a deep learning-based method called ToxGIN to improve the prediction of peptide toxicity.
- Unlike existing peptide toxicity prediction models, ToxGIN, as proposed, extracts information from the 3D structures of peptides, not solely from their sequences.
- Comparative studies on the same datasets demonstrated that the proposed ToxGIN outperformed existing models in prediction of peptide toxicity.

Abbreviations

AAC, amino acid composition; DPC, dipeptide composition; NLP, Natural Language Processing; GIN, Graph Isomorphism Networks; TP, True Positive; TN, True Negative; FP, False Positive; FN, False Negative; SE, Sensitivity; MCC, Matthews Correlation Coefficient; ASDC, Pseudo-amino Acid Composition; Pse-AAC, Pseudo-amino Acid Composition; APAAC, Amphiphilic Pseudo Amino Acid Composition; RF, Random Forest; SVM-RBF, Support Vector Machine with Radial Basis Function kernel; GNB, Gaussian Naive Bayes; LR, Logistic Regression; KNN, K-Nearest Neighbors; t-SNE, t-Distributed Stochastic Neighbor Embedding; Bi-GRU, Bi-directional Gated Recurrent Unit; GVP, Geometric Vector Perceptron; GNN, Graph Neural Network; GAT, Graph Attention Network; TAG, Topology Adaptive Graph Convolutional Network; APPNP, Approximate Personalized Propagation of Neural Predictions; GraphSAGE, Graph Sample and Aggregation; AUROC, Area Under the Receiver Operating Characteristic Curve; AUPRC, Area Under the Precision-Recall Curve

Supplementary Data

Supplementary data are available at *Briefings in Bioinformatics* online.

Author contributions

Yun Tang supervised the research project. Qiule Yu designed and implemented the ToxGIN method. Qiule Yu constructed the models and evaluated the performance. Qiule Yu collected, compiled and processed the data sets. Qiule Yu and Zhixing Zhang analyzed

the results. Qiule Yu, Zhixing Zhang, Weihua Li, Guixia Liu, and Yun Tang wrote the manuscript. All authors read the manuscript and approved the final version.

Conflict of interest: The authors declare that they have no conflict of interest.

Funding

National Key Research and Development Program of China (Grant 2023YFF1204904), the National Natural Science Foundation of China (Grants U23A20530 and 82173746) and Shanghai Frontiers Science Center of Optogenetic Techniques for Cell Metabolism (Shanghai Municipal Education Commission).

Code and data availability

All the datasets generated and source code for this study are hosted on GitHub and can be found at <https://github.com/cihebiyql/ToxGIN>.

References

- Albericio F, Kruger HG. Therapeutic peptides. *Future Med Chem* 2012;**4**:1527–31. <https://doi.org/10.4155/fmc.12.94>.
- Guntuboina C, Das A, Mollaei P. et al. PeptideBERT: A language model based on transformers for peptide property prediction. *J Phys Chem Lett* 2023;**14**:10427–34. <https://doi.org/10.1021/acs.jpcclett.3c02398>.
- Chiangjong W, Chutipongtanate S, Hongeng S. Anticancer peptide: Physicochemical property, functional aspect and trend in clinical application (review). *Int J Oncol* 2020;**57**:678–96. <https://doi.org/10.3892/ijo.2020.5099>.
- Lei J, Sun LC, Huang S. et al. The antimicrobial peptides and their potential clinical applications. *Am J Transl Res* 2019;**11**:3919–31.
- Craik DJ, Fairlie DP, Liras S. et al. The future of peptide-based drugs. *Chem Biol Drug Des* 2013;**81**:136–47. <https://doi.org/10.1111/cbdd.12055>.
- Haggag YA, Donia AA, Osman MA. et al. Peptides as drug candidates: Limitations and recent development perspectives. *Biomed J Sci Tech Res* 2018;**8**:6659–62.
- Wang L, Wang N, Zhang W. et al. Therapeutic peptides: Current applications and future directions. *Signal Transduct Target Ther* 2022;**7**:48.
- Blomme EAG, Will Y. Toxicology strategies for drug discovery: Present and future. *Chem Res Toxicol* 2016;**29**:473–504. <https://doi.org/10.1021/acs.chemrestox.5b00407>.
- Khan F, Niaz K, Abdollahi M. Toxicity of biologically active peptides and future safety aspects: An update. *Curr Drug Discov Technol* 2018;**15**:236–42. <https://doi.org/10.2174/1570163815666180219112806>.
- Duracova M, Klimentova J, Fucikova A. et al. Proteomic methods of detection and quantification of protein toxins. *Toxins* 2018;**10**:99.
- Altschul SF, Madden TL, Schäffer AA. et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**:3389–402.
- Naamati G, Askenazi M, Linial M. ClanTox: A classifier of short animal toxins. *Nucleic Acids Res* 2009;**37**:W363–8. <https://doi.org/10.1093/nar/gkp299>.
- Gupta S, Kapoor P, Chaudhary K. et al. In silico approach for predicting toxicity of peptides and proteins. *PLoS One* 2013;**8**:e73957.
- Pan X, Zuallaert J, Wang X. et al. ToxDL: Deep learning using primary structure and domain embeddings for assessing protein toxicity. *Bioinformatics* 2020;**36**:5159–68.
- Mikolov T, Sutskever I, Chen K. et al. Distributed representations of words and phrases and their compositionality. *Adv Neural Inf Process Syst*, Volume 2. Red Hook, NY, United States: Curran Associates Inc., 57 Morehouse Lane, 2013, 3111–9. <https://dl.acm.org/doi/10.5555/2999792.2999959>.
- Shi H, Li Y, Chen Y. et al. ToxMVA: An end-to-end multi-view deep autoencoder method for protein toxicity prediction. *Comput Biol Med* 2022;**151**:106322.
- Morozov V, Rodrigues CHM, Ascher DB. CSM-toxin: A web-server for predicting protein toxicity. *Pharmaceutics* 2023;**15**:431.
- Brandes N, Ofer D, Peleg Y. et al. ProteinBERT: A universal deep-learning model of protein sequence and function. *Bioinformatics* 2022;**38**:2102–10. <https://doi.org/10.1093/bioinformatics/btac020>.
- Mall R, Singh A, Patel CN. et al. VISH-pred: An ensemble of fine-tuned ESM models for protein toxicity prediction. *Brief Bioinform* 2024;**25**. <https://doi.org/10.1093/bib/bbae270>.
- Ke G, Meng Q, Finley T. et al. LightGBM: A highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst* 2017;**30**:52.
- Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY: Association for Computing Machinery, pp. 785–94, 2016. <https://doi.org/10.1145/2939672.2939785>.
- Wei L, Ye X, Xue Y. et al. ATSE: A peptide toxicity predictor by exploiting structural and evolutionary information based on graph neural network and attention mechanism. *Brief Bioinform* 2021;**22**:bbab041.
- Landrum G. RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum* 2013;**8**:5281.
- Zhu XJ, Feng CQ, Lai HY. et al. Predicting protein structural classes for low-similarity sequences by evaluating different features. *Knowl-Based Syst* 2019;**163**:787–93.
- Wei L, Ye X, Sakurai T. et al. ToxIBTL: Prediction of peptide toxicity based on information bottleneck and transfer learning. *Bioinformatics* 2022;**38**:1514–24. <https://doi.org/10.1093/bioinformatics/btac006>.
- Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* 1992;**89**:10915–9.
- Mu Z, Yu T, Liu X. et al. FEFS: A novel feature extraction model for protein sequences and its applications. *BMC Bioinformatics* 2021;**22**:1–15.
- Jiao S, Ye X, Sakurai T. et al. Integrated convolution and self-attention for improving peptide toxicity prediction. *Bioinformatics* 2024;**40**:btac297.
- Prabantu VM, Yazhini A, Srinivasan N. *Manoeuvring Protein Functions and Functional Levels by Structural Excursions*. In: Levine H, Jolly MK, Kulkarni P, Nanjundiah V, editors. *Phenotypic Switching*, p. 77–104. San Diego: Academic Press; 2020. <https://doi.org/10.1016/B978-0-12-817996-3.00006-2>.
- Jumper J, Evans R, Pritzel A. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583–9. <https://doi.org/10.1038/s41586-021-03819-2>.
- Ofer D, Brandes N, Linial M. The language of proteins: NLP, machine learning & protein sequences. *Comput Struct Biotechnol J* 2021;**19**:1750–8. <https://doi.org/10.1016/j.csbj.2021.03.022>.

32. Xu K, Hu W, Leskovec J. et al. How powerful are graph neural networks? arXiv preprint arXiv:1810.00826 2018. <https://doi.org/10.48550/arXiv.1810.00826>.
33. Mirdita M, Schütze K, Moriwaki Y. et al. ColabFold: Making protein folding accessible to all. *Nat Methods* 2022;**19**:679–82. <https://doi.org/10.1038/s41592-022-01488-1>.
34. Lin Z, Akin H, Rao R. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 2023;**379**:1123–30. <https://doi.org/10.1126/science.ade2574>.
35. Bateman A, Martin MJ, Orchard S. et al. UniProt: The universal protein knowledgebase in 2023. *Nucleic Acids Res* 2023;**51**: D523–31.
36. Kaas Q, Yu R, Jin AH. et al. ConoServer: Updated content, knowledge, and discovery tools in the conopeptide database. *Nucleic Acids Res* 2012;**40**:D325–30. <https://doi.org/10.1093/nar/gkr886>.
37. Wood DL, Miljenović T, Cai S. et al. ArachnoServer: A database of protein toxins from spiders. *BMC Genomics* 2009;**10**:1–8.
38. Fu L, Niu B, Zhu Z. et al. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012;**28**:3150–2. <https://doi.org/10.1093/bioinformatics/bts565>.
39. Kawashima S, Pokarowski P, Pokarowska M. et al. AAindex: Amino acid index database, progress report 2008. *Nucleic Acids Res* 2008;**36**:D202–5.
40. Jain S, Ranjan P, Sengupta D. et al. TpPred: A tool for hierarchical prediction of transport proteins using cluster of neural networks and sequence derived features. *IJCB* 2014;**1**: 28–36.
41. Chou KC. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins Struct Funct Genet* 2001;**43**: 246–55.
42. Van Der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res* 2008;**9**:2579–625.
43. Scarselli F, Gori M, Tsoi AC. et al. The graph neural network model. *IEEE Trans Neural Netw* 2008;**20**:61–80.
44. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 2016. <https://doi.org/10.48550/arXiv.1609.02907>.
45. Veličković P, Casanova A, Liò P. et al. Graph attention networks. In: *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, 2018. <https://doi.org/10.48550/arXiv.1710.10903>.
46. Du J, Zhang S, Wu G. et al. Topology adaptive graph convolutional networks. arXiv preprint arXiv:1710.10370 2017. <https://doi.org/10.48550/arXiv.1710.10370>.
47. Gasteiger J, Bojchevski A, Günnemann S. Predict then propagate: Graph neural networks meet personalized pagerank. arXiv preprint arXiv:1810.05997 2018. <https://doi.org/10.48550/arXiv.1810.05997>.
48. Hamilton W, Ying Z, Leskovec J. Inductive representation learning on large graphs. *Adv Neural Inf Process Syst* 2017;**30**:1025–35.
49. Ebrahimikondori H, Sutherland D, Yanai A. et al. Structure-aware deep learning model for peptide toxicity prediction. *Protein Sci* 2024;**33**:e5076.
50. Zhao Z, Gui J, Yao A. et al. Improved prediction model of protein and peptide toxicity by Integrating Channel attention into a convolutional neural network and gated recurrent units. *ACS Omega* 2022;**7**:40569–77. <https://doi.org/10.1021/acsomega.2c05881>.
51. Jing B, Eismann S, Suriana P. et al. Learning from protein structure with geometric vector perceptrons. In: Hofmann, K. et al. (eds.) *International Conference on Learning Representations*. Red Hook, New York, USA: OpenReview.net, 2021. <https://openreview.net/forum?id=1YLJDvSx6J4>.
52. Ying Z, Bourgeois D, You J. et al. Gnnexplainer: Generating explanations for graph neural networks. *Adv Neural Inf Process Syst* 2019;**32**:9240–51.