JSTEM

NeuroSeq: An Integrated Learning Framework for Enhanced Neuropeptide Prediction and Therapeutic Discovery

Rui Wang[1,2,#], Yifan Luo[2,3,#],Lijiang Huang[1], Minfang Zhou[1],Yingzi Feng[1],Yunlong Zhou[2],Yang Wang[2,*], Mengpei Zhang[1,*]

[1] Department of Gastroenterology, the Affiliated Xiangshan Hospital of Wenzhou Medical University. Ningbo 315700, China;
[2] Wenzhou Institute, University of Chinese Academy of Sciences, Wenzhou 325000, China;
[3] The Second Clinical School, Wenzhou Medical University, Wenzhou 325035, China;
* Correspondence: Yang Wang: wangy0727@ucas.ac.cn and Mengpei Zhang: 609957527@qq.com.
# Rui Wang and Yifan Luo contributed equally to this work.

Abstract: Neuropeptides, crucial in neural signaling and present throughout the nervous system, are integral to processes like stress response and neural repair. Traditional prediction methods fall short due to their complexity, necessitating advanced strategies. Our study introduces NeuroSeq, an integrated learning network that improves neuropeptides' prediction accuracy and efficiency. It employs pre-trained convolutional networks, advanced CNNs, and TCNs to learn complex amino acid patterns automatically. Combining machine and deep learning via average probability voting, an ensemble learning method further enhances predictions. We also propose a technique for optimizing neuropeptide sequences and assessing their reliability, generating high-confidence candidates. Experiments show NeuroSeq surpasses existing models, hastening neuropeptide identification and progressing peptide-based neurotherapeutics. This research emphasizes deep learning's potential in neuropeptide studies, advancing biological insights and therapeutic development. NeuroSeq marks a significant advance in computationally unraveling neuropeptide intricacies, benefiting neuroscience and clinical applications.

Keywords: Deep Learning; Neuropeptide; Machine Learning

1. Introduction

Neuropeptides are signaling molecules in the brain involved in numerous physiological functions. As small proteins are produced and released by neurons through regulated secretion pathways, they can modulate other transmitters in the cellular environment and play a hormonal role in the organism[1]. Numerous studies have shown that neuropeptides and neuropeptide receptors in organisms play a significant role in regulating fatty function[2], influencing metabolism, and treating anxiety Disorders[3]. Additionally, research has discovered that NPs such as ghrelin, neuropeptide Y (NPY), neurotensin, and pituitary adenylate cyclase-activating polypeptide (PACAP) exhibit neuroprotective effects on various neurodegenerative diseases, including Parkinson's disease and Alzheimer's disease[4]. Neuropeptides' discovery and functional annotation require tools to deeply understand biosynthesis pathways, activity states, functions, and disease associations[5]. Biochemical and molecular biology methods have made significant contributions in this regard; however, due to the complexity of neuropeptides, it is becoming increasingly necessary to develop more efficient and comprehensive approaches[6]. This involves predicting the sequence of neuropeptides, gaining insight into their functional diversity, and identifying structures corresponding to specific activities[7,8].

In this context, deep learning technology has emerged as a crucial tool at the forefront of artificial intelligence, opening up new horizons for neuropeptide research[9].

Received: May 6, 2025
Revised: May 14, 2025
Accepted: May 15, 2025
Published: May 20, 2025

Citation: Wang R., Luo Y., Huang L., Zhou M., Feng Y., Zhou Y., Wang Y., and Zhang M. NeuroSeq: An Integrated Learning Framework for Enhanced Neuropeptide Prediction and Therapeutic Discovery. Journal of STEM2025, 2(2), 00007. https://doi.org/10.63460/YLUS8074

These algorithms can discover patterns and extract information from vast biological datasets, predicting the potential functions of unknown peptides[10]. The application of deep learning technology significantly accelerates neuropeptides' discovery and functional annotation and holds great significance for understanding biological systems and developing therapies for agarose diseases[11].

Various machine learning-based methods have been developed to predict neuropeptides[12]. NeuroPID[13] is a machine-learning tool that predicts metazoan neuropeptide precursors. It utilizes support vector machines and integrated decision trees, processing over 600 features combining biophysical, chemical, and statistical characteristics. NeuroPP[14] predicts neuropeptides using an optimized predictor based on single amino acid, dipeptide, and tripeptide sequence composition. Both methods use known data to train models for predicting potential new peptides, continuously improving prediction models by introducing new algorithms. The NeuroPred[15] algorithm trains logistic regression models based on cleavage data from different species, providing confidence intervals for cleavage probabilities and assessing the accuracy of these probabilities based on thresholds. NeuroPred can also calculate the mass of predicted peptides and aid in identifying new neuropeptides through mass spectrometry techniques by adding post-translational modifications. In specific application scenarios, such as NeuroPIpred proposed by Agrawal et al.[16] for insect neuropeptides, the advantages of support vector machines (SVM) in particular domains are demonstrated by analyzing positional residue preferences. More complex two-layer stacked models like PredNeuroP[17] combine nine feature descriptors with five machine learning algorithms, introducing 45 models as base learners to predict neuropeptides more accurately based on comprehensive datasets. Jiang et al.'s NeuroPpred-Fuse[18] constructs a refined prediction model by selecting different feature subsets to achieve higher prediction accuracy[19]. Hasan et al.'s NeuroPred-FRL[20] model combines 11 encodings, six classifiers, and feature selection methods through complex steps, building a powerful integrated model to improve recognition accuracy. However, current machine learning models suffer from varying degrees of interpretability issues, and there is still room for improvement in accuracy.

When it comes to encoding schemes for neuropeptide sequences, the challenges lie in capturing sufficient sequence information and ensuring the diversity and reliability of that information. Methods based on deep learning excel in learning the distribution of data and prevent the need for complex feature engineering. Liu et al.[20] devised the NeuroPpred-SVM model, which uses embeddings based on bidirectional encoder representations and employs a Support Vector Machine (SVM) for neuropeptide prediction. They trained a Transformer-based pre-trained language model with over 680 million parameters to forecast protein contact maps. Wang et al.[21] developed the interpretable and stable NeuroPred-PLM neuropeptide prediction model. This model utilizes a protein language model to derive semantic representations of neuropeptides, simplifying feature engineering, and incorporates a multi-scale convolutional neural network to enhance the local feature representation of neuropeptide embeddings. For interpretability, NeuroPred-PLM introduces a multi-head attention network that captures the contribution of positions to neuropeptide prediction through attention scores. Although deep learning has demonstrated remarkable feature extraction and learning capabilities from data, current methods still face challenges in obtaining complete neuropeptide sequences and representing comprehensive features.

There are various tools in the field of neuropeptide prediction, such as BLAST and PeptideCutter. Among them, BLAST identifies homologous neuropeptide sequences by searching for similar sequences in a database and predicting their potential functions. However, its reliance on local alignment makes it sensitive to sequence variations and requires significant computational resources and time when dealing with long sequences. PeptideCutter is primarily used to predict potential cleavage sites in protein sequences to aid in identifying regions where neuropeptides may be generated. However, its prediction effectiveness may not be ideal for enzymes that have not been sufficiently studied.

To address the current challenges in neuropeptide prediction, this paper presents the NeuroSeq integrated learning network architecture designed to improve prediction accuracy and efficiency. Firstly, we propose a neuropeptide prediction method based on a pre-trained convolutional network to overcome the limitations of traditional machine learning methods that rely on manual feature extraction and are susceptible to human bias. This method combines a pre-trained model[22], convolutional neural networks, and temporal convolutional networks to automatically capture and learn complex patterns and biological functions of amino acid sequences. Secondly, we introduce a prediction method based on ensemble learning to address the incomplete feature extraction that can occur with deep learning methods based on sequence features. This approach integrates machine learning with deep learning models, achieving higher prediction validity through averaged probability voting. Finally, to optimize the generation of high-probability neuropeptides, we describe a method for optimal neuropeptide sequence selection and reliability evaluation based on ensemble deep learning.

2. Results

2.1 Model Selection

In evaluating neural peptide prediction methodologies based on pre-trained convolutional networks that integrate ProteinBERT. Convolutional Neural Networks (CNNs) and Temporal Convolutional Networks (TCNs), various combination strategies were analyzed in Figure 1B. Figure 1A shows that initially, the models exhibited lower accuracies, which significantly improved as training progressed, culminating in a training set accuracy of 94.58% and a validation set accuracy of 90.9%. By harnessing the pre-training knowledge of ProteinBERT alongside the potent feature extraction capabilities of CNNs and TCNs, the model effectively extracts informative features from sequence data of neuropeptides.

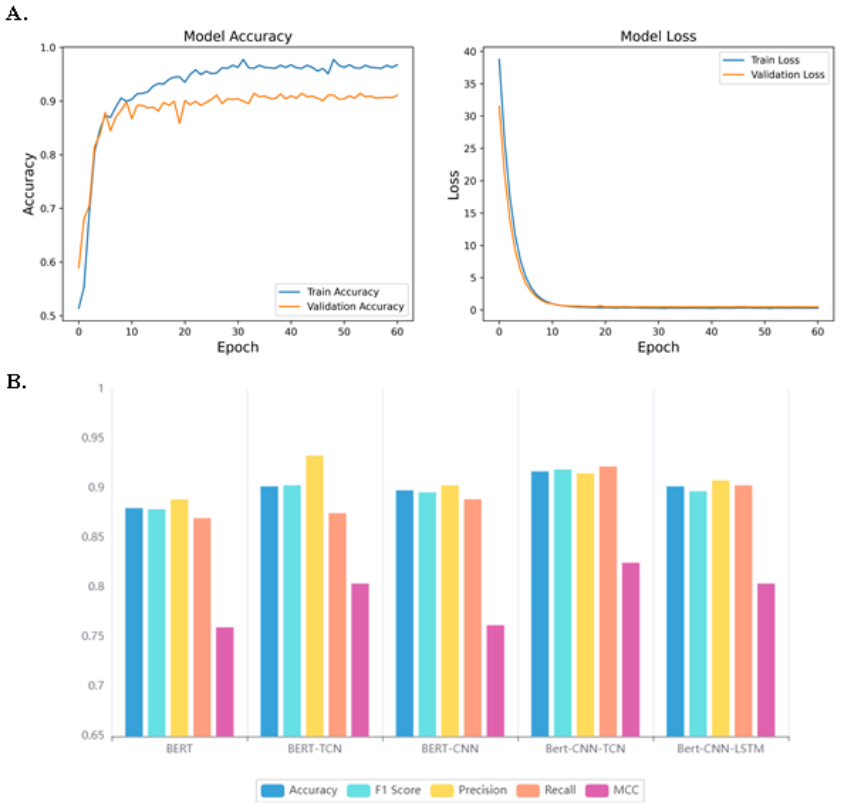

Figure 1. Accuracy and loss graphs of deep learning models and performance evaluation of different deep learning models.

To comprehensively assess the accuracy of different feature extraction techniques combined with classifiers in distinguishing neuropeptides from non-neuropeptides, Figure 2A embodies experiments conducted with eight widely adopted feature extraction methodologies and nine classical classifiers. These feature extraction techniques span from basic compositional descriptors (e.g., Amphiphilic Pseudo-Amino Acid Composition (AAP), Amino Acid Composition (AAC), and Dipeptide Composition (DPC) to more intricate sequence descriptors (such as Grouped Amino Acid Pairs (GGAP), Amphiphilic Symmetry Distributions (ASDC), Composition-Transition-Distribution (CTD), and Pseudo Amino Acid Composition (PSAAC), while classifiers encompass rule-based systems (Random Forest, RF, and Extremely Randomized Trees, ERT), instance-based methods (K-Nearest Neighbors, KNN), linear models (Logistic Regression, LR), and ensemble learning algorithms (Adaptive Boosting, AdaBoost, Gradient Boosting Decision Trees, GBDT, and XGBoost). The results depicted in Figure 1B indicate that combinations involving PSAAC, ASDC, and CTD for feature extraction paired with RF, LR, GBDT, ERT, and XGBoost classifiers significantly outperform other configurations on the test set.

CTD, PSAAC, and ASDC as feature extraction methods were selected and juxtaposed against a merged feature set (Merge), each being trained with five distinct classifiers: RF, ERT, GBDT, LR, and XGBoost. Their performances were evaluated using the standard training set after applying 5-fold cross-validation. Observations from the training set indicate that PSAAC coupled with the GBDT classifier achieves the highest accuracy of 0.913. A similar trend is evident in the results following the 5-fold cross-validation process.

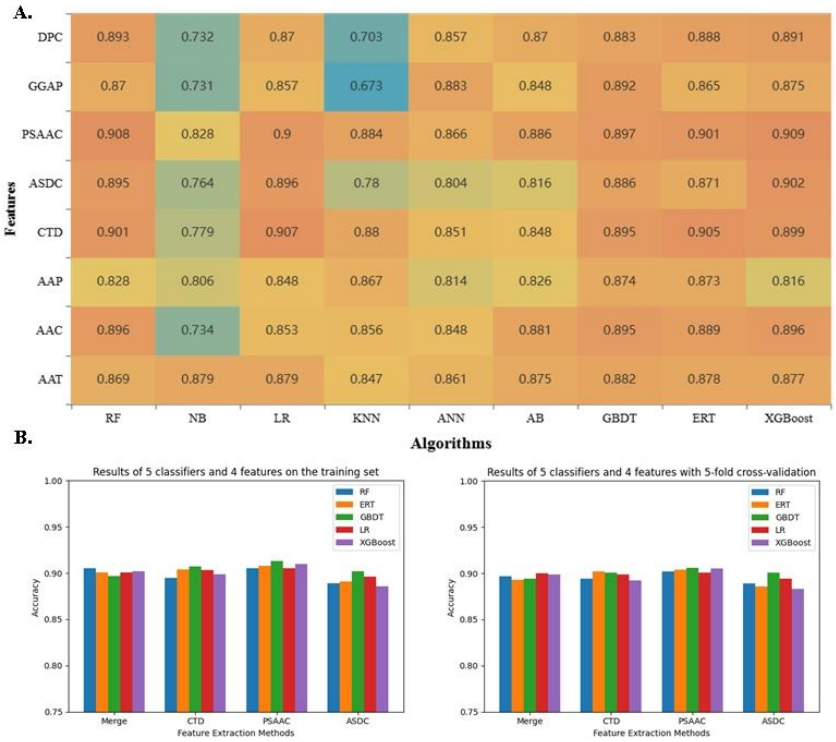

Figure 2. Evaluation of the performance of different machine learning models. (A) Accuracy heat map of 72 machine learning models on a test set. (B) The accuracy of 5 classifiers and four feature extraction methods after training set and 5-fold cross-check.

2.2 Comparison with State-of-the-Art Methods for Neuropeptide Prediction

In pursuit of more precise identification and classification of neural peptide sequences, the NeuroSeq model was developed, excelling across multiple performance metrics. NeuroSeq incorporates multi-feature fusion strategies with advanced machine learning algorithms to enhance differentiation between neuropeptides and non-neuropeptides.

We benchmarked NeuroSeq against several established models, including PredNeuroP, NeuroPred-FRL, NeuroPpred-Fuse, and NeuroPred-PLM. Notably, as the source code for the NeuroPred-FRL model was not publicly available, our replication relied on reconstructing the model architecture based on literature descriptions and adopting parameters reported for its utilization.

As can be seen from the comparison results in Table 1, NeuroSeq excels in several significant performance indicators, demonstrating its efficiency and superiority in the task of neuropeptide identification. Its high accuracy reflects the model's formidable classification capability. In classification tasks, precision and recall are two crucial indicators: the former reflects the actual accuracy of the sequences predicted to be neuropeptides, and the latter signifies the proportion of correctly predicted neuropeptides among all neuropeptide sequences. In both respects, NeuroSeq performs exceptionally well. Specifically, its high recall rate suggests NeuroSeq is highly suitable for high-coverage screening of neuropeptides. Simultaneously, as more comprehensive measurements, the F1 score and MCC have also proven that NeuroSeq displays high quality and accuracy in all classification decisions. Therefore, we can assert that NeuroSeq can precisely predict neuropeptide sequences, potentially unlocking many neuropeptides' latent functions and paving the way for new methods to treat various diseases. Table 1 shows the ACC, Pre, Rec, F1, and MCC of NeuroSeq are 0. 918, 0. 915, 0. 922, 0. 919, and 0. 837, respectively. Compared to other models, NeuroSeq performs excellently across all indicators.

Table 1. Comparison with existing models.

Methods	ACC	Pre	Rec	F1	MCC
PredNeuroP	0.870	0.873	0.865	0.874	0.801
NeuroPred-FRL	0.901	0.913	0.876	0.884	0.814
NeuroPpred-Fuse	0.874	0.896	0.902	0.887	0.809
NeuroPred-PLM	0.911	0.905	0.883	0.920	0.812
NeuroSeq	0.918	0.915	0.922	0.919	0.837

2.3 Effectiveness of Features Extracted by Networks

Figure 3 illustrates the changes in the feature space of samples before and after model training. One can notice in the visualization of feature distribution after reducing the feature dimensions to two using t-distributed Stochastic Neighbor Embedding (t-SNE) that each point represents a peptide sequence, with different colors distinguishing between neuropeptide and non-neuropeptide categories. After model training, these two categories of peptide samples exhibit a more apparent distribution in the reduced dimensional feature space. This indicates that the model, when handling the task of neuropeptide classification, can effectively learn and capture high-quality discriminative features, significantly enhancing the separation degree of sample points in the feature space.

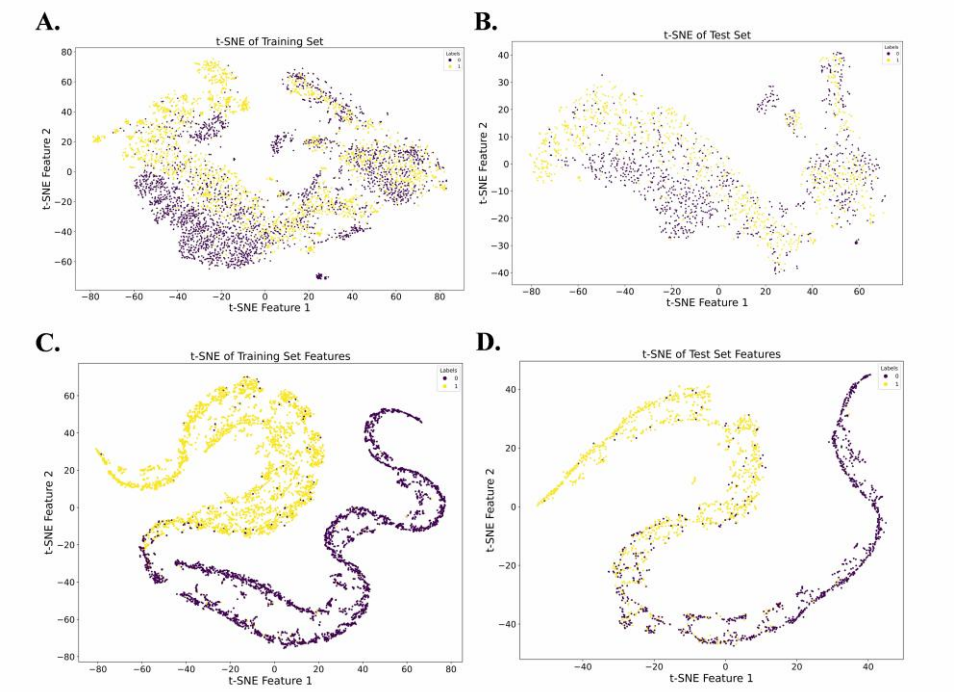

Figure 3. Dimensionality reduction analysis plots before training (first row) and after training (second row).

2.4 Conformal Prediction for Analysis of Sequence Performance

Neuropeptides are a class of biologically active small molecules increasingly being revealed by research for their potential in drug design and disease treatment. However, neuropeptides in nature often limit specific biological activity and stability, making sequence optimization a critical step in enhancing their clinical application value.

In the sequence optimization section, this paper adopts a strategy of iterative mutations of sequences through integrating deep models. The model integrates predictions from several models established in the previous sections, optimizing the understanding and prediction accuracy of the sequences. The optimization employs a random mutation strategy, randomly selecting a position in the sequence and replacing it with any amino acid from all possible options. Although this method lacks meticulous exploration compared to systematic mutation, adding randomness helps find advantageous mutations that systematic methods may overlook. This process is repeated at subsequent positions in the sequence to ensure a comprehensive search of the entire sequence space and identify key mutation points that enhance the activity and stability of neuropeptides.

Table 2. Neuropeptide analysis for optimized strategy generation.

Predicted Label	Credibility	Sequences
1	0.8253	FLGGQPYW
1	0.7876	SLDSMGARYGFLPYGRF
1	0.4870	FSQAQGKVDMPLPRQFTSS RSSERWAPKS
1	0.3219	ECWSQAEDCSDGH-CCAGRSFSKNCRPYGGD

The conformal prediction results show a series of credibility values for the neuropeptides generated by the optimization strategy. The first sequence in Table , FLGGQPYW, has a credibility value of 0.8253, indicating that the model believes the sequence is highly

likely to be a neuropeptide. Biological experiments have indeed proven that it has a significant effect on promoting neuron cell line proliferation and synapse regeneration.

2.5 Neuropeptide Prediction Critical Site Analysis and Biological Characteristics

Active neuropeptides are derived from pro-neuropeptide precursors through protein hydrolysis processing[23]. As another mechanism to increase the complexity of mammalian neuropeptides[24], a single precursor molecule can be cleaved differently, thereby producing different peptide groups in other cell types[25]. The analysis of neuropeptides through mass spectrometry peptidomics reveals the diversity of neuropeptides; there are hundreds to thousands of different peptides. The unique biological activity of each neuropeptide is defined by its primary amino acid sequence, which is mediated by peptidergic receptors. Neuropeptide signal transduction achieves neurotransmission along with classical small-molecule neurotransmitters[26]. The cleavage process of neuropeptide precursors usually occurs on basic amino acid residues, especially at lysine (K) and arginine (R) sites.

To address the problem of lengthy neuropeptide precursors being challenging to analyze, this paper adopts a method of trimming the neuropeptide precursor sequences based on a sliding window strategy. This method combines the precise and efficient processing capabilities of bioinformatics, setting a 39-residue-length window from the N-terminus to the C-terminus in the neuropeptide precursor sequence, sliding with an amino acid as the step length, thus producing (N-39)+1 possible cleavage peptide segments. This processing method allows the extraction and analysis of all possible neuropeptide candidate fragments in the entire neuropeptide precursor sequence[27].

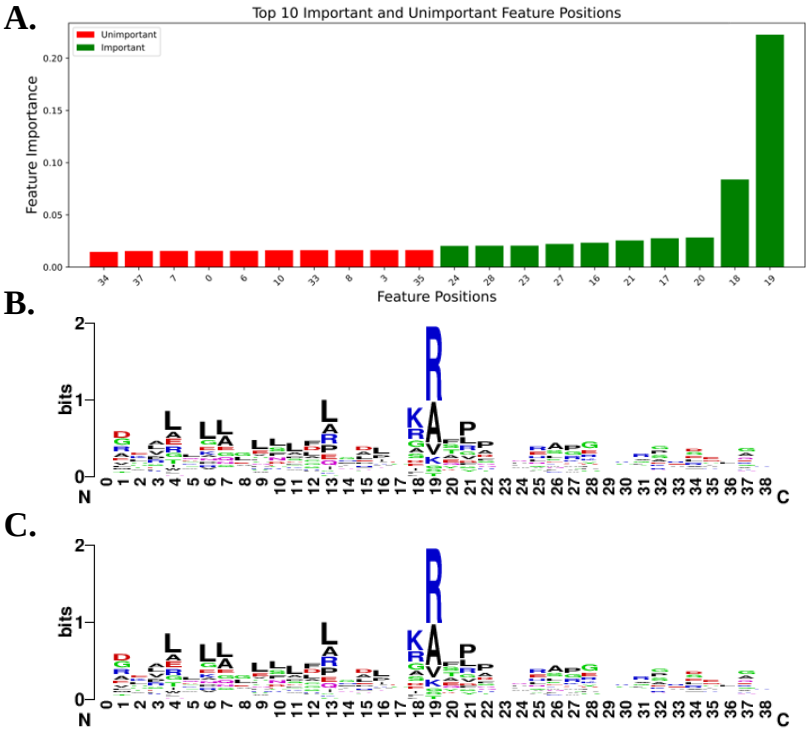

Figure 4. Neuropeptide prediction critical site analysis. (A) Ranking of crucial cleavage sites in the back ten and top 10. (B) Key cleavage sites of proto neuro peptides. (C) Optimize the critical cleavage sites of neuropeptides.

This paper trained a bioinformatics model using the random forest algorithm to identify critical cleavage sites contributing most to neuropeptides' positive and negative sample classification. Label encoding was used in model training to handle protein sequences, and then key sites were identified through the random forest classifier's feature importance scoring. The feature importance revealed in Figure A shows that sites 21, 17, 20,

18, and 19 are significant in the neuropeptide precursor sequences generated by sequence optimization.

Further sequence analysis using WebLogo, the results in Figure B show that the occurrence frequency of lysine (K) at the 18th site is relatively high. The 19th site commonly sees arginine (R) and alanine (A), and the frequencies in Figure 5C match the cleavage sites of essential amino acids in the original sequences of existing neuropeptides.

These observations are consistent with previous literature[28], indicating that at the NPs cleavage and maturation process, essential amino acids at the site (especially the KR site or the sole R site) and specific non-basic amino acids (such as A at the 19th site) play critical roles. These sites comply with natural neuropeptides' physical cleavage and biosynthetic rules and enhance the necessity of maintaining critical amino acid positions unchanged in sequence optimization.

By calculating and comparing multiple physicochemical properties of generated sequences and sequences in the training set, this paper can more comprehensively evaluate the biological characteristics of generated sequences. These indicators include Eisenberg Hydrophobicity, Gravy (overall hydrophobic value), Isoelectric Point, Hydrophobic Ratio, Instability Index, Charge, Charge Density, and Aliphatic Index. The comparison of these physicochemical indicators between generated sequences and training set sequences in Figure 5 shows that their distribution trends are highly similar. By comparing the median positions in the violin plots, it can be found that the neuropeptide sequences produced by the model have the same trend as the original sequences on these characteristics, proving that the model's generation capability maintains the original physiological and chemical properties while retaining the original biological functions.

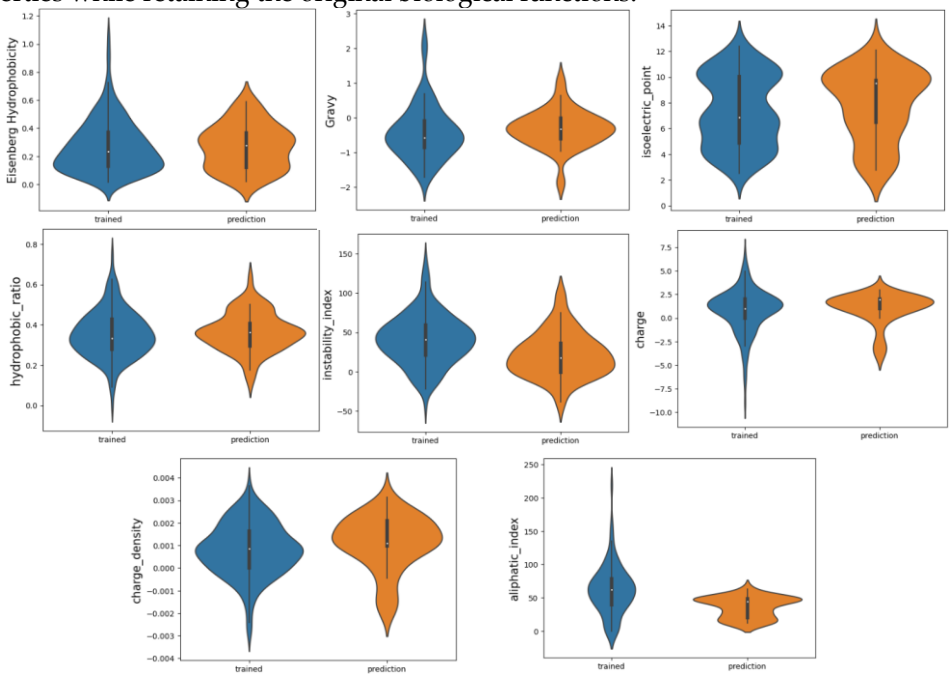

Figure 5. Optimize the resulting neuropeptide physicochemical profile.

3. Discussion

Neuropeptides, as pivotal biomolecules, play a fundamental role in decoding and regulating neural signal transmission processes. These bioactive molecules are widely distributed throughout the central and peripheral nervous systems, participating in various physiological processes such as stress response and nerve repair facilitation. To address the challenge of neuropeptide prediction, we have developed NeuroSeq. This integrated learning network architecture enhances prediction accuracy and efficiency by automatically capturing intricate patterns within amino acid sequences. Moreover, our approach incorporates average probability voting to improve prediction effectiveness. We propose

a method for optimizing neuropeptide sequence selection and evaluating reliability based on integrated deep-learning techniques to generate high-probability neuropeptides. Experimental results demonstrate superior performance compared to existing methods while facilitating accelerated identification of neuropeptides and promoting peptide-based treatments for nervous system disorders. The prediction of neuropeptides, vital signaling molecules in the nervous system, has been a persistent challenge due to their inherent complexity and the limitations of traditional computational methods. This study introduced NeuroSeq, an integrated learning framework designed to significantly enhance the accuracy and efficiency of neuropeptide prediction and aid in the discovery of novel therapeutic candidates. Our approach successfully addresses several key limitations of previous models, such as reliance on manual feature engineering, incomplete feature extraction, and the need for more robust optimization and reliability assessment strategies.

NeuroSeq's strength lies in its multifaceted architecture, combining pre-trained protein language models (ProteinBERT) with advanced Convolutional Neural Networks (CNNs) and Temporal Convolutional Networks (TCNs) to automatically discern complex patterns within amino acid sequences. Furthermore, the integration of machine learning classifiers with deep learning models through an optimized average probability voting ensemble method demonstrably improved prediction validity. This ensemble strategy, weighting deep learning at 0.7 and machine learning at 0.3, proved optimal in our experiments. Our results clearly indicate that NeuroSeq surpasses existing state-of-the-art models, including PredNeuroP, NeuroPred-FRL, NeuroPpred-Fuse, and NeuroPred-PLM, across multiple key performance metrics such as accuracy (0.918), precision (0.915), recall (0.922), F1-score (0.919), and Matthews Correlation Coefficient (MCC) (0.837). The particularly high recall rate (0.922) underscores NeuroSeq's potential as a powerful tool for high-coverage screening of neuropeptides, minimizing the chances of missing potentially functional candidates. The improved MCC (0.837), a balanced measure of classification performance, further highlights the robustness and reliability of NeuroSeq over its predecessors. The t-SNE visualization compellingly demonstrated the model's capacity to learn highly discriminative features, leading to a clearer separation between neuropeptide and non-neuropeptide sequences in the feature space after training.

A significant contribution of this work is the novel pipeline for neuropeptide sequence optimization and reliability assessment using conformal prediction. This allowed for the generation of high-confidence neuropeptide candidates. Notably, one such optimized sequence, FLGGQPYW, predicted with a high credibility score of 0.8253, was subsequently validated experimentally, demonstrating a significant effect on promoting neuron cell line proliferation and synapse regeneration. This tangible result bridges the gap between computational prediction and practical therapeutic discovery, showcasing the real-world applicability of NeuroSeq.

Further enhancing the biological relevance of our model, the critical site analysis identified specific amino acid positions (17, 18, 19, 20, 21) within precursor sequences that are crucial for distinguishing neuropeptides. The frequent occurrence of lysine (K) at site 18 and arginine (R) or alanine (A) at site 19 aligns with known enzymatic cleavage preferences in neuropeptide processing, such as the importance of basic residues (K, R) at cleavage sites. This consistency with established biological understanding reinforces the model's ability to capture biologically meaningful patterns rather than just statistical correlations. Moreover, the physicochemical properties of the neuropeptide sequences generated by NeuroSeq closely mirrored those of the training set, suggesting that our model preserves essential biological characteristics while optimizing for predicted activity.

While NeuroSeq demonstrates considerable advancements, some limitations should be acknowledged. The reimplementation of the NeuroPred-FRL model was based on literature descriptions due to the unavailability of its source code, which might introduce variations from the original model's performance. The random mutation strategy employed for sequence optimization, while beneficial for exploring a wider sequence space

and potentially identifying overlooked advantageous mutations, may not be as exhaustive as systematic mutation approaches. The experimental validation, though promising, was conducted on a limited set of predicted peptides.

Future research should focus on several avenues. Expanding the training dataset to include a more diverse range of neuropeptides from various species and functional classes could further enhance NeuroSeq's generalizability. More extensive experimental validation of the predicted and optimized neuropeptides is crucial to translate these computational findings into tangible therapeutic leads. Exploring alternative or more advanced deep learning architectures and ensemble techniques could yield further improvements in prediction accuracy. Additionally, enhancing the interpretability of the deep learning components, beyond the current critical site analysis, could provide deeper insights into the sequence determinants of neuropeptide identity and function. Integration of multi-omics data, such as gene expression or protein-protein interaction data, could also provide a more holistic approach to neuropeptide discovery.

NeuroSeq represents a significant step forward in the computational prediction of neuropeptides. Its integrated learning framework, combining advanced neural network architectures, ensemble methods, and a robust sequence optimization and reliability assessment pipeline, offers superior performance and practical utility. This work not only accelerates the identification of novel neuropeptides but also paves the way for the rational design of peptide-based therapeutics for a range of neurological disorders, thereby holding substantial promise for advancing both neuroscience research and clinical applications.

4. Materials and Methods

4.1 Dataset construction

To build a more accurate and efficient neuropeptide prediction model, this study comprehensively compares existing predictors such as PredNeuroP, NeuroPred-FRL, and NeuroPred-PLM, considering their respective datasets, including Swiss-Prot[29], Neuropep[30], and Neuropep2.0[31]. Considering data comprehensiveness and update frequency, neuropeptide sequences from the Neuropep2.0 database were ultimately selected for this study.

To ensure fairness, a series of consistent screening steps were followed: neuropeptides with lengths between 5 and 100 were retained, and subsequently, CD-HIT[32] with a threshold of 0.9 was applied to exclude sequences with a homology exceeding 90%. These steps resulted in 4460 neuropeptides, from which 892 were randomly selected for the test set. Negative samples with a similar length distribution to the positive samples were extracted from the UniProt database, yielding 4460 negative samples, with 892 sequences randomly chosen for the non-neuropeptide dataset. The training set totaled 7316 samples, while the test set comprised 1784.

4.2 Workflow of the NeuroSeq Model

The model architecture of NeuroSeq is shown in Figure 6. The NeuroSeq model, adopting an ensemble approach, combines multiple models to overcome their respective limitations while preserving their strengths. The main operational steps are as follows: (i) Randomly sample data to train both machine learning and deep learning models; (ii) Select candidate threshold values between 0 and 1 for each model and calculate the final performance metrics after assigning different thresholds for voting, choosing the thresholds that yield the best performance; (iii) Obtain the optimal ensemble model. The ensemble framework incorporates a dynamic voting system based on prediction result weights, ultimately determining a threshold allocation of 0.3 for machine learning and 0.7 for deep learning, yielding the final model.

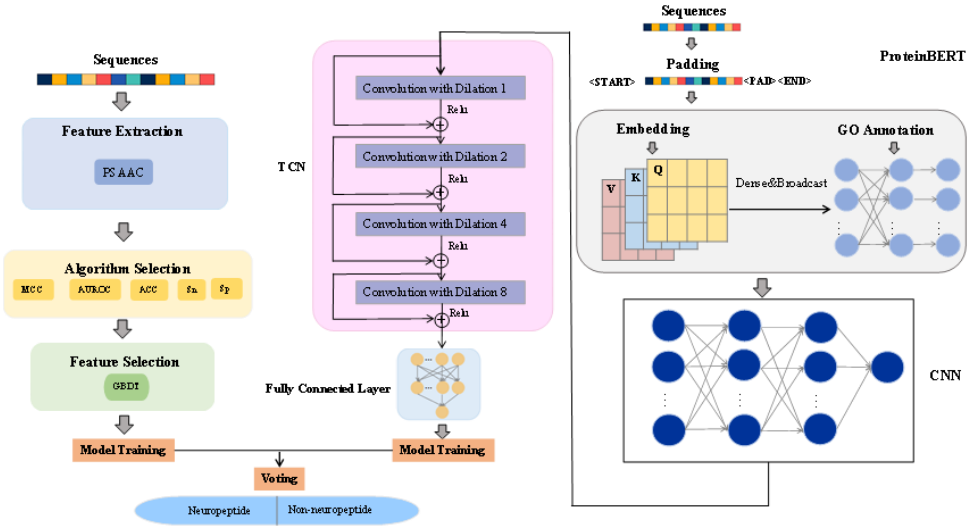

Figure 6. The model architecture of NeuroSeq.

4.3 Feature Representation of Peptides

In this work, we utilize three deep learning models, BERT, CNN, and TCN, along with three feature encoding methods: CTD, PSAAC, and ASDC. Detailed descriptions of each method are provided below:

4.3.1 BERT

BERT, short for Bidirectional Encoder Representations from Transformers, is a Transformer-based architecture that revolutionizes natural language processing tasks[33]. Its key aspect lies in its pre-training strategy, which primarily involves Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). During pre-training, BERT leverages large text corpora to establish text representations through unsupervised learning. The MLM task randomly hides words from the input sequence. It predicts them based on the context, explicitly selecting 15% of the text segments for processing, with 80% of the chosen segments masked, 10% replaced randomly, and the remaining 10% kept unchanged. The NSP task, on the other hand, predicts whether a sentence is a logical continuation of another, helping BERT understand language from two perspectives: intra-sentence semantic relationships and inter-sentence logical coherence. Regarding input representation, BERT employs WordPiece embeddings, which break words down into finer-grained subunits, enhancing sensitivity to morphological variations. Special separators like [CLS] and [SEP] are introduced to distinguish sentences, and positional embeddings capture the relative or absolute positions of words in the sequence. Pre-training of BERT was conducted on two massive text corpora, BooksCorpus and English Wikipedia, totaling 3.3 billion words. This diverse corpus covers a broad range of topics and endows the model with excellent generalization capabilities for downstream tasks.

ProteinBERT[34] emerges as a pre-trained protein language model that captures intricate relationships within protein sequences. Doing so significantly reduces the complexity and difficulty of learning new tasks. When processing protein sequences, ProteinBERT uses 26 tokens to represent 20 standard amino acids, along with selenium cysteine (U), undefined amino acid (X), amino acid (OTHER), and three additional markers (START, END, and PAD). For sequence identification, START and END markers are added before the first and after the last amino acid, respectively. The PAD token is incorporated into sequences shorter than the selected mini-batch length. ProteinBERT underwent pre-training on 106 million proteins, encompassing the entire known protein universe, with two concurrent tasks. The first task involves bidirectional language modeling of protein sequences, while the second predicts Gene Ontology (GO) annotations to capture diverse protein functions. GO annotations comprise a manually curated set of 45K terms defined

at the whole-protein level, spanning the entire protein space across all organisms. They cover molecular functions, biological processes, and subcellular locations.

4.3.2 CNN

Among the various deep learning architectures, the Convolutional Neural Network (CNN) is a prevalent model in the field. CNNs leverage a combination of convolutional layers, pooling layers[35], and fully connected layers to hierarchically extract and learn features from input data for specific tasks. When handling sequence data, CNNs excel due to their remarkable ability to capture local features. Our network design incorporates two core one-dimensional convolutional layers, each followed by a max-pooling layer. Each convolutional layer is equipped with 1024 filters, a kernel size of 4, and is activated by the ReLU function. Moreover, L2 regularization with a strength of 0. 02 is applied to every convolutional layer, mitigating overfitting and enhancing the model's generalization capabilities during learning. With a pooling size of 2, the subsequent max-pooling layers reduce dimensionality complexity, highlight the most critical features, and minimize computational resource requirements.

4.3.3 TCN

Temporal Convolutional Network (TCN)[36], an improvement upon the CNN deep learning architecture, emerges as a novel deep learning model primarily designed for processing sequential data. In numerous sequence modeling tasks, TCN outperforms existing recurrent networks such as Long Short-Term Memory (LSTM)[37], Gated Recurrent Unit (GRU)[38], and Recurrent Neural Network (RNN)[39], offering superior accuracy and simplicity. In our architecture, TCN captures and exploits long-term dependencies between sequences by employing multiple causal convolutional layers and residual blocks. Specifically, it utilizes 256 filters with a kernel size of six and manages the receptive field size through a series of dilation coefficients [1, 2, 4, 8].

4.3.4 CTD

The CTD representation method categorizes amino acids in protein sequences based on various attributes such as their chemical properties, charge, or hydrophobicity. In the CTD approach, amino acids are grouped according to these attributes, and an analysis is conducted on the Composition, Transition between amino acid properties, and Distribution of these properties across the entire sequence. The Composition (C) is represented as a 21-dimensional vector, calculated using the formula:

$$C(a) = \frac{N(a)}{L}, \qquad a = 1, 2, 3$$

Where N(a) denotes the total number of amino acids in category a. Transition (T) is encoded as a 21-dimensional vector indicating the occurrence of an A-type residue followed by a B-type residue, or vice versa. It is computed as follows:

$$\begin{cases} T(a,b) = \dfrac{N(a,b) + N(b,a)}{L-1} \\ T(a,c) = \dfrac{N(a,c) + N(b,c)}{L-1} \\ T(b,c) = \dfrac{N(b,c) + N(c,b)}{L-1} \end{cases}$$

In the above formula, N(a, b) represents the number of dipeptides in the sequence. The distribution descriptor consists of five values for three classes based on the fraction of the entire sequence where the first residue of a given group is located. Spatial Distribution (D) is encoded as a 105-dimensional vector. In summary, the CTD encoding results in a 147-dimensional vector.

4.3.5 PSAAC

PSAAC aims to enhance the expressiveness of the standard AAC by incorporating additional information such as sequence patterns, the importance of specific amino acid positions, and physicochemical characteristics. It is calculated as follows:

$$f(a,i) = \frac{N(a,i)}{M(i)}$$

In this formula, N(a, i) represents the frequency of amino acid a at position i, and M(i) denotes the total number of sequences considered at position i. By integrating this supplementary information, PSAAC provides a deeper understanding of sequence characteristics and enables more accurate functional predictions.

4.3.6 ASDC

ASDC describes the correlation between a given amino acid property in a sequence and the same or different properties at other positions within the sequence. The adaptive k-skip-n-gram model, an enhanced version of the skip-n-gram model, adapts to sequences of varying lengths in the dataset and incorporates more distance information into the features. The calculation method for ASDC is as follows:

$$FV = \left\{ \frac{N\left(\alpha_{m_1}\alpha_{m_2}\ldots\alpha_{m_n}\right)}{N\left(T_{\text{skipgram}}\right)} \middle| \; 1 \le m_1 \le 20, \le m_1 \le 20, \ldots, \le m_n \le 20 \right\}$$

$$T_{\text{skipgram}} = \cup_{a=1}^{k} \text{Skip}(\text{DT} = a)$$

Where . This method considers n residuals with distances ranging from 1 to k, where k is the length of each sequence. ASDC incorporates auto-covariance and cross-covariance analyses, determining whether there are statistical correlations between amino acid properties separated by specific distances in the sequence. This approach aids in predicting protein function or structural characteristics.

4.4 Reliability Assessment

Traditional methods for quantifying reliability, such as Platt scaling, PAC learning frameworks, and Bayesian approaches, often rely on substantial amounts of data and specific assumptions (e.g., normal distribution) to estimate prior probabilities. However, these assumptions are frequently challenging in real-world applications due to the scarcity of training samples and issues like measurement noise. Therefore, exploring a more robust and precise method for quantifying reliability becomes crucial.

The theory of Conformal Prediction (CP) proposed by Shafer and Vovk[40] offers a new perspective. At its core, CP assigns accurate confidence levels to prediction outcomes. Given a significance level ε, CP constructs a prediction interval that covers the actual label with at least a 1-ε confidence level. In machine learning classification tasks, based on the training set data Z, the classifier learns a mapping from feature space X to label space Y. By introducing ε as an additional parameter, CP furnishes a prediction interval for new instances, ensuring the accurate label coverage within it with a confidence of at least 1-ε.

We can assess the consistency between sample labels and model predictions using non-conformity measures for each training sample. Zhan et al. employed Conformal Predictive methods to quantify reliability, achieving control over prediction uncertainty[41]. This is accomplished by comparing the model's probability of predicting the sample's label with the maximum probability assigned to an incorrect label. When encountering a new sample, we can calculate its p-value based on. Prior non-conformity measures represent the probability of the sample having a particular potential label at a given significance level ε. The p-value reflects how well the label matches observed data under a predetermined confidence level.

Notably, unlike the predictive probabilities output by the commonly used Softmax function in multiclass classification, CP does not require the sum of p-values across the label space to equal 1. This is because CP focuses on the conformity between prediction

outcomes and the existing training observations rather than directly modeling conditional probabilities P(Y|X). Hence, training observations independently derive each potential label's non-conformity measure. CP provides two key reliability metrics for decision-making: credibility and confidence. Credibility reflects the maximum consistency between predicted labels and training observations, while confidence indicates trust in the optimal predicted label after discounting other possible predictions.

We establish a threshold (e.g.,0.5) to evaluate prediction reliability, considering a prediction reliable only if its associated p-value exceeds this threshold. To ensure high confidence in the correct label significantly surpassing support for other potential labels, we also mandate that the highest p-value must be at least three times that of the second-highest. Such criteria ensure high prediction reliability and alleviate potential ambiguity in label selection.

5. Conclusions

This study introduced NeuroSeq, an innovative integrated learning framework designed to address the persistent challenges in accurately and efficiently predicting neuropeptides. By synergistically combining pre-trained protein language models with advanced deep learning architectures (CNNs and TCNs) and an optimized ensemble learning strategy, NeuroSeq has demonstrated superior performance over existing methodologies.

Our findings highlight NeuroSeq's capability to automatically capture complex sequence patterns, leading to significantly improved predictive accuracy, as evidenced by an accuracy of 0.918 and an MCC of 0.837. The framework's novel approach to neuropeptide sequence optimization, guided by conformal prediction for reliability assessment, was successfully validated through experimental results, where an optimized peptide demonstrated promotion of neuron cell line proliferation and synapse regeneration. Furthermore, the identification of critical amino acid sites for neuropeptide processing, consistent with known biological mechanisms, underscores the model's ability to learn biologically relevant features.

In conclusion, NeuroSeq represents a substantial advancement in the computational toolkit for neuropeptide research. It not only enhances the precision of neuropeptide identification but also provides a robust platform for the generation and validation of novel peptide candidates. This work paves the way for accelerated discovery in neuropharmacology and holds significant promise for the development of new peptide-based therapeutics for neurological disorders, thereby making a valuable contribution to both fundamental neuroscience and translational medicine.

Author Contributions: Conceptualization, Mengpei Zhang and Yang Wang; methodology, Rui Wang; software, Yifan Luo; validation, Lijiang Huang, Minfang Zhou and Yingzi Feng; formal analysis, Yunlong Zhou; investigation, Rui Wang and Yifan Luo; resources, Yang Wang and Mengpei Zhang; data curation, Yifan Luo; writing—original draft preparation, Rui Wang; writing—review and editing, Yang Wang, Mengpei Zhang, Rui Wang and Yifan Luo; visualization, Yifan Luo; supervision, Yang Wang and Mengpei Zhang; project administration, Yang Wang and Mengpei Zhang; funding acquisition, Yang Wang and Mengpei Zhang. All authors have read and agreed to the published version of the manuscript.

Funding: Yang Wang thanks the Research Fund of Wenzhou Institute, UCAS (WIUCASQD2021043). Mengpei Zhang thanks Key Projects of the Joint Medical Research Center between the Wenzhou Institute, UCAS and Wenzhou Medical University Xiangshan Hospital.

Data Availability Statement: We encourage all authors of articles published in MDPI journals to share their research data. In this section, please provide details regarding where data supporting reported results can be found, including links to publicly archived datasets analyzed or generated during the study. Where no new data were created, or where data is unavailable due to privacy or ethical restrictions, a statement is still required. Suggested Data Availability Statements are available in section "MDPI Research Data Policies" at https://www.mdpi.com/ethics.

Acknowledgments: The High-Performance Computing Center of Wenzhou Institute, UCAS supports the work of Machine Learning Computational.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Burbach JP. *What are neuropeptides?* Methods Mol Biol 2011,789:1-36.
2. Mishra G, Townsend KL. *The metabolic and functional roles of sensory nerves in fatty tissues.* Nat Metab 2023,5(9):1461-1474.
3. Gupta PR, Prabhavalkar K. *Combination therapy with neuropeptides for the treatment of anxiety disorder.* Neuropeptides 2021,86:102127.
4. Behl T, Madaan P, Sehgal A, et al. *Demystifying the Neuroprotective Role of Neuropeptides in Parkinson's Disease: A Newfangled and Eloquent Therapeutic Perspective.* Int J Mol Sci 2022,23(9):4565.
5. Sachkova MY. *Neuropeptides at the origin of neurons*. Nat Ecol Evol 2022,6(10):1410-1411.
6. Anapindi KDB, Romanova EV, Checco JW, et al. *Mass Spectrometry Approaches Empowering Neuropeptide Discovery and Therapeutics*. Pharmacol Rev 2022,74(3):662-679.
7. Hook V, Lietz CB, Podvin S, et al. *Diversity of Neuropeptide Cell-Cell Signaling Molecules Generated by Proteolytic Processing Revealed by Neuropeptidomics Mass Spectrometry.* J Am Soc Mass Spectrom 2018,29(5):807-816.
8. Phetsanthad A, Vu NQ, Yu Q, et al. *Recent advances in mass spectrometry analysis of neuropeptides.* Mass Spectrom Rev 2023,42(2):706-750.
9. Wang L, Zeng Z, Xue Z, et al. *DeepNeuropePred: A robust and universal tool to predict cleavage sites from neuropeptide precursors by protein language model.* Comput Struct Biotechnol J 2023,23:309-315.
10. Qiang X, Zhou C, Ye X, et al. *CPPred-FL: a sequence-based predictor for large-scale identification of cell-penetrating peptides by feature representation learning.* Brief Bioinform 2020,21(1):11-23.
11. Ahmed S, Muhammod R, Khan ZH, et al. *ACP-MHCNN: an accurate multi-headed deep-convolutional neural network to predict anticancer peptides.* Sci Rep2021,11(1):23676.
12. Han B, Zhao N, Zeng C, et al. *ACPred-BMF: bidirectional LSTM with multiple feature representations for explainable anticancer peptide prediction.* Sci Rep2022,12(1):21915.
13. Karsenty S, Rappoport N, Ofer D, et al. *NeuroPID: a classifier of neuropeptide precursors.* Nucleic Acids Res 2014,42:W182-W186.
14. Kang J, Fang Y, Yao P, et al. *NeuroPP: a tool for the prediction of neuropeptide precursors based on optimal sequence composition*. Interdiscip Sci 2019,11:108–114.
15. Southey BR, Amare A, Zimmerman TA, et al. *NeuroPred: a tool to predict cleavage sites in neuropeptide precursors and provide the masses of the resulting peptides.* Nucleic Acids Res 2006,34(Web Server issue):W267-W272.
16. Agrawal P, Kumar S, Singh A, et al. *NeuroPIpred: a tool to predict, design, and scan insect neuropeptides*. Sci Rep 2019,9(1):5129.
17. Bin Y, Zhang W, Tang W, et al. *Prediction of neuropeptides from sequence information using ensemble classifier and hybrid features.* J Proteome Res 2020,19:3732–3740.
18. Jiang M, Zhao B, Luo S, et al. *NeuroPpred-Fuse: an interpretable stacking model for prediction of neuropeptides by fusing sequence information and feature selection methods.* Brief Bioinform 2021,22(6):bbab310.
19. Hasan MM, Alam MA, Shoombuatong W, et al. *NeuroPred-FRL:an interpretable prediction model for identifying neuropeptides using feature representation learning.* Brief Bioinform 2021,22(6):bbab167.
20. Liu Y, Wang S, Li X, et al. *NeuroPpred-SVM: A New Model for Predicting Neuropeptides Based on Embeddings of BERT.* J Proteome Res 2023,22(3):718-728
21. Wang L, Huang C, Wang M, et al. *NeuroPred-PLM: an interpretable and robust model for neuropeptide prediction by protein language model.* Brief Bioinform 2023,24(2):bbad077.
22. Wu R, Ding F, Wang R, et al. *High-Resolution De Novo Structure Prediction from Primary Sequence.* bioRxiv 2022.
23. Hook V, Lietz CB, Podvin S, Cajka T, Fiehn O. *Diversity of Neuropeptide Cell-Cell Signaling Molecules Generated by Proteolytic Processing Revealed by Neuropeptidomics Mass Spectrometry.* J Am Soc Mass Spectrom 2018,29(5):807-816.
24. Jiang Z, Lietz CB, Podvin S, et al. *Differential Neuropeptidomes of Dense Core Secretory Vesicles (DCSV) Produced at Intravesicular and Extracellular pH Conditions by Proteolytic Processing.* ACS Chem Neurosci 2021,12(13):2385-2398.
25. Salio C, Lossi L, Ferrini F, et al. *Neuropeptides as synaptic transmitters.* Cell Tissue Res 2006,326(2):583-598.
26. Hook V, Kind T, Podvin S, et al. *Metabolomics Analyses of 14 Classical Neurotransmitters by GC-TOF with LC-MS Illustrates Secretion of 9 Cell-Cell Signaling Molecules from Sympathoadrenal Chromaffin Cells in the Presence of Lithium*. ACS Chem Neurosci 2019,10(3):1369-1379.
27. Wang Y, Kang J, Li N, et al. *NeuroCS: A Tool to Predict Cleavage Sites of Neuropeptide Precursors.* Protein Pept Lett 2020,27(4):337-345.

28. Falth, M. et al. *Neuropeptidomics strategies for specific and sensitive identification of endogenous peptides*. Cell Proteomics 2007,6:1188–1197.
29. Boutet E, Lieberherr D, Tognolli M, et al. *UniProtKB/Swiss-Prot*. Methods Mol Biol 2007,406:89-112.
30. Wang Y, Wang M, Yin S, et al. *NeuroPep: a comprehensive resource of neuropeptides.* Oxford 2015,bav038.
31. Wang M, Wang L, Xu W, et al. *NeuroPep 2.0: An Updated Database Dedicated to Neuropeptide and Its Receptor Annotations.* J Mol Biol 2024,436(4):168416.
32. Li W, Godzik A. *Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.* Bioinformatics 2006,22(13):1658-1659.
33. Ruffolo J, Chu L, Mahajan S, et al. *Fast, accurate antibody structure prediction from deep learning on a massive set of natural antibodies*. Nat Commun 2023,14(1):2389.
34. Brandes N, Ofer D, Peleg Y, et al. *ProteinBERT: a universal deep-learning model of protein sequence and function.* Bioinformatics 2022,38(8):2102-2110.
35. Bouvrie J . Notes on Convolutional Neural Networks. Neural Nets 2006.
36. Devkota P, Mohanty SD, Manda P. *A Gated Recurrent Unit-based architecture for recognizing ontology concepts from biological literature.* BioData Min 2022,15(1):22.
37. Hochreiter S, Schmidhuber J. *Long Short-Term Memory.* Neural Computation 1997,9(8):1735-1780.
38. Chung J, Gulcehre C, Cho K H, et al. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. Eprint Arxiv 2014.
39. Zaremba W, Sutskever I, Vinyals O. Recurrent Neural Network Regularization. Eprint Arxiv 2014.
40. Vovk, Vladimir, Alexander Gammerman. Algorithmic Learning in a Random World.2005.
41. Zhan X, Wang F, Gevaert O. *Reliably Filter Drug-Induced Liver Injury Literature With Natural Language Processing and Conformal Prediction*. IEEE J Biomed Health Inform 2022,26(10):5033-5041.