Computational and Structural Biotechnology Journal 23 (2024) 309–315

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Software/Web server Article

DeepNeuropePred: A robust and universal tool to predict cleavage sites from neuropeptide precursors by protein language model

Lei Wang [a,b,1], Zilu Zeng [c,1], Zhidong Xue [a,d,*], Yan Wang [a,b,*]

[a] Institute of Medical Artificial Intelligence, Binzhou Medical University, Yantai, Shandong 264003, China
[b] School of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China
[c] Wuhan Children's Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei 430010, China
[d] School of Software Engineering, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China

ARTICLE INFO

Keywords:
Neuropeptides
Deep learning
Protein language model
Webserver

ABSTRACT

Neuropeptides play critical roles in many biological processes such as growth, learning, memory, metabolism, and neuronal differentiation. A few approaches have been reported for predicting neuropeptides that are cleaved from precursor protein sequences. However, these models for cleavage site prediction of precursors were developed using a limited number of neuropeptide precursor datasets and simple precursors representation models. In addition, a universal method for predicting neuropeptide cleavage sites that can be applied to all species is still lacking. In this paper, we proposed a novel deep learning method called DeepNeuropePred, using a combination of pre-trained language model and Convolutional Neural Networks for feature extraction and predicting the neuropeptide cleavage sites from precursors. To demonstrate the model's effectiveness and robustness, we evaluated the performance of DeepNeuropePred and four models from the NeuroPred server in the independent dataset and our model achieved the highest AUC score (0.916), which are 6.9%, 7.8%, 8.8%, and 10.9% higher than Mammalian (0.857), insects (0.850), Mollusc (0.842) and Motif (0.826), respectively. For the convenience of researchers, we provide a web server (http://isyslab.info/NeuroPepV2/deepNeuropePred.jsp).

1. Introduction

Neuropeptides are a diverse and complex class of signaling molecules that modulate nearly every physiological process and behavior in metazoans [1]. Typically consisting of less than 100 amino acids, they are produced from larger precursor molecules through a series of post-translational processing steps[2,3]. Neuropeptides exert their effects not only via the nervous system but also peripherally through the endocrine system, where they regulate physiological functions, including food intake, metabolism, reproduction, fluid homeostasis, cardiovascular function, energy homeostasis, stress control, pain perception, social behaviors, memory and learning, and circadian rhythm[4–7]. Consequently, neuropeptides are implicated in the pathogenesis of numerous diseases, and the neuropeptide signaling system represents a promising therapeutic target for the treatment of sleep disorders, autism, depression, heart failure, obesity, diabetes, high blood pressure, epilepsy, and other disorders[1,4,8,9]. Furthermore, neuropeptides serve as valuable biomarkers and diagnostic probes for prospective disease diagnosis and prognosis. Neuropeptides are enclosed within larger prohormones, most between 50 and 500 amino acids in length. These prohormones have a signal peptide at the N-terminus consisting of approximately 16–30 amino acid residues and can encode multiple structurally related or unrelated neuropeptides, or they can encode a single neuropeptide[10,11]. The sequence of prohormones can often be inferred from genetic information. However, it is generally difficult to predict biologically active peptides based on genetic information due to the many processing steps involved. Obtaining experimental verification for the final neuropeptide structure can be challenging due to the limited number of neurons that frequently produce specific neuropeptides, their intricate structures, and their low physiological concentrations[12].

With the development of the next-generation sequencing technology, more and more new genomes have been obtained and it is more important to identify precursors and detect neuropeptides. The

* Corresponding authors at: Institute of Medical Artificial Intelligence, Binzhou Medical University, Yantai, Shandong 264003, China.
E-mail addresses: zdxue@hust.edu.cn (Z. Xue), yanw@hust.edu.cn (Y. Wang).
[1] These authors contributed equally to this work

https://doi.org/10.1016/j.csbj.2023.12.004
Received 28 August 2023; Received in revised form 30 November 2023; Accepted 2 December 2023
Available online 5 December 2023

successful combination of mass spectrometry with genome sequencing has enabled the characterization of the insect peptidome for neuropeptides in Drosophila melanogaster[13–15] and Apis mellifera[16]. While the most ubiquitous sites for cleavage are these basic residues, there are also some additional amino acid combinations that can be used as cleavage sites in the prohormone[17]. Neuropeptide precursors contain certain structural information: (1) The signal peptide; (2) The basic motifs; (3) C-terminal amidation; (4) post-translational modifications[2]. Thus, due to the conserved pattern, it is possible to identify the cleavage sites of neuropeptide precursors. NeuroPred is a web application designed to predict cleavage sites at basic amino acid locations in neuropeptide precursor sequences, including Motif[12], Mammalian [18], Mollusc[19], and Insect[20] models. Southey et al. [12] proposed a Known Motif model comprised of several prevalent motifs associated with neuropeptide precursor cleavage. While the Motif method identified the most known cleavages, it also had a high rate of false positive prediction results. Hummon et al. [19] also predicted neuropeptide cleavage sites in mollusk precursors using the logistic regression model on combinations of amino acids and location information. Amare et al. [18] developed a binary logistic regression model to analyze mammalian neuropeptide precursors. The study revealed significant differences in processing between vertebrate and molluscan precursors, particularly in the processing of dibasic sites. NeuroPred-Insect model was trained on Apis mellifera and Drosophila melanogaster precursors using binary logistic regression, multi-layer perceptron and k-nearest neighbor models. All of the aforementioned neuropeptide cleavage site prediction tools are limited by small-scale data, resulting in unsatisfactory predictive accuracy.

In recent years, deep learning-based methods have been applied to different biological problems[21–24]. Compared with traditional machine learning models, deep learning-based methods can better capture the dataset distribution feature and reduce complex feature engineering processing. In the field of natural language processing (NLP), transfer learning through pre-trained language models has become ubiquitous. These models primarily learn context-based word embeddings, such as BERT[25]. With over a billion protein sequences in databases, a highly effective approach is to use self-supervised language models to learn latent information from unlabeled sequences. A pre-trained language model called ESM[26], based on the Transformer architecture, was trained to predict protein contact maps using over 680 million parameters. Protein language models have presented some exciting breakthroughs, enabling the discovery of protein structure and function solely from the evolutionary relationships that exist in sequence corpora. Recently, different researchers[27–30] have developed some deep language models based on large-scale protein sequence datasets. These protein language models have been widely used in related tasks of the protein, such as signal peptide prediction[31], protein subcellular localization[32], bitter peptide[33], neuropeptide prediction[34], protein domain prediction[35,36], and transmembrane protein prediction [37]. The use of transfer learning to obtain better feature representations for neuropeptide precursors is highly inspiring and holds great potential.

Currently, a universal and robust solution for predicting cleavage sites of neuropeptide precursors is still lacking. With more and more neuropeptide precursors being incorporated into protein public databases such as UniProt, this provides an opportunity for improved annotation of the cleavage site prediction of these neuropeptides. By collecting the neuropeptide precursors datasets from the UniProt database[38], we proposed a novel and robust model called DeepNeuropePred to detect the cleavage sites of neuropeptide precursors. Through a combination of pre-trained language models and Convolutional Neural Networks for feature extraction, our developed model achieved superior performance than four models from NeuroPred (Mammlian, Insects, Mollusc, and Motif). The DeepNeuropePred webserver is freely available at http://isyslab.info/NeuroPepV2/deepNeuropePred.jsp and source code is visible at https://github.com/ISYSLAB-HUST/DeepNeuropePred.

2. Materials and methods

2.1. Benchmark dataset

By searching the keywords such as “neuropeptide” from the UniProt/KB database and filtering these protein terms without the precursor of flags and the signal peptide, we collected 1194 complete reviewed precursors of neuropeptide. Here, we adopted 31 test precursors as the independent test dataset which was integrated into the UniProt database after 2014. To guarantee a fair comparison on the independent test dataset, the collected sequences of the training dataset that are similar to the test precursors at a threshold of 40% using CD-HIT[39] would be dropped. By the above steps, the remaining precursors from the training dataset included 717 precursors (training: validation=4:1). All training data and test data are freely available at https://github.com/isyslab-hust/DeepNeuropePred. It can be seen from Table S1 that the distribution ratio of cleavage sites and non-cleavage sites are not much different in the training set and the test set.

2.2. Feature extraction based on the protein language model

The emergence of pre-trained language models has propelled research on protein representation into an exciting new phase that eliminates the need for human annotation. This breakthrough allows for the acquisition of protein sequence representation from a vast pool of unannotated protein sequences, leading to significant improvements in downstream tasks. Additionally, utilizing pre-trained language models acts as an efficient regularization technique, reducing the risk of overfitting on limited training data.

In this context, we have adopted a transformer-based self-supervised language model called ESM[26] (version: esm1_t12_85M_UR50S), which boasts over 85 million parameters and 12 transformer layers. The esm1_t12_85M_UR50S model is capable of processing protein sequences as inputs, generating dynamic embeddings with a dimension of $L * 768$, where L represents protein length. The esm1_t12_85M_UR50S model have been developed based on the scaled dot-product attention. The attention function is defined as:

$$A = \mathrm{softmax}\left(1 \Big/ \sqrt{d}\, Q(h)K(h)^T\right)V(h) \tag{1}$$

Here the query Q, key K, and value V, are projections of the protein sequence to $n \times d$ matrices where n is the length of the protein sequence and d is the dimension of the outer product between Q and K. This outer product obtains an $n \times n$ attention map, which is rescaled and passed through the SoftMax function, thereby representing each position of the sequence in the output as a convex combination of the sequence of values V. The pre-trained language model (ESM-12) can obtain the global feature representation of the precursor neuropeptides because the input is the full length of the precursor sequence rather than the window sequence of the cleavage site.

2.3. Cleavage site local information enhancement with convolutional neural network

The processing procedure of the complete precursor sequence was followed as the Insect model of NeuroPred (Southey, et al., 2008). Consistent with NeuroPred, we also employ a window of length 18 around the cleavage site. Considering the strong local correlation of the neuropeptide cleavage site, we use a multi-scale convolutional neural network based on sparse connectivity to obtain the local vector representations of adjacent residues. The strong local correlation feature can be obtained by the followings:

$$c1 = Dropout(conv(z, h1, n1)) \tag{2}$$

$$c2 = Dropout(conv(c1, h2, n2)) \tag{3}$$

Where *conv* is the convolutional function, $h1$,$h2$ is the convolutional kernel size and $n1$, $n2$ is the number of convolutional kernels. The dropout function is a technique utilized to mitigate the risk of overfitting in models. It operates by randomly deactivating neurons and their corresponding connections, thereby preventing the network from relying too heavily on any one neuron. This strategy encourages all neurons to develop better generalization abilities.

2.4. Loss function

DeepNeuropePred takes the entire window sequence of L residues as input and outputs a probability score, where the score indicates whether the input site belongs to neuropeptide cleavage sites. The cross-entropy loss function is used as the optimization goal of the neural network. In our proposed model, the learning rate is 0.001, and the optimizer is Adam optimizer. The loss function is defined as:

$$Loss\ (\ y, \widehat{y}\ )\ =\ -\left(\mathrm{w_d y log}(\widehat{\mathrm{y}}) + \mathrm{w_b}(1-\mathrm{y})\log(1-\widehat{\mathrm{y}})\right) + \lambda\sum_{u} u^2 \tag{4}$$

where y and $\widehat{y}$ are the label and the predicted possibility scores, respectively; ground truth label y reflects if it belongs to the neuropeptide cleavage site (1) or non-neuropeptide cleavage site (0), and $\widehat{y}$ is the output of the DeepNeuropePred; λ is the regularization factor; u represents all the trainable parameters; w_d and w_b are the reciprocal of the number of neuropeptide cleavage sites and non-neuropeptide cleavage sites, which can reduce the imbalance of samples in the dataset. Finally, 0.5 is selected as the threshold, probability scores greater than 0.5 is considered neuropeptide cleavage site, and those less than 0.5 is considered non-neuropeptide cleavage site.

2.5. Evaluation metrics

To evaluate the model performance, we choose the following metrics: AUC (Area Under the Receiver Operating Characteristic Curve), ACC (accuracy), Precision, Recall, AUPR (area under the precision–recall curve) and MCC (Matthew's correlation coefficient). The relevant definitions are as follows:

$$precision = \frac{TP}{TP + FP} \tag{5}$$

$$recall = \frac{TP}{TP + TN} \tag{6}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$FPR = \frac{FP}{FP + TN} \tag{8}$$

$$F1 = \frac{2 * TP}{2 * TP + FP + FN} \tag{9}$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \tag{10}$$

TP (true positive), FP (false positive), TN (true negative) and FN (false negative) are obtained from the confusion matrix. The Receiver Operating Characteristic (ROC) curve has the FPR (false positive rate) and TPR (recall or true positive rate) as the horizontal and vertical coordinates, respectively, and the area under the ROC curve is the AUC score. The area under the curve with recall and precision as the horizontal and vertical coordinates is the AUPR score.

3. Result and discussion

3.1. Overview of the DeepNeuropePred framework

The model framework shown in Fig. 1 consists of 4 parts, pre-trained self-supervised language model, convolutional layers, average pooling layer, and position-wise fully-connected layers. Using a pre-trained language model could reduce the risk of overfitting on small training data, which is equivalent to a kind of regularization method. The pre-trained language model (esm1_t12_85M_UR50S, https://github.com/facebookresearch/esm) can obtain the global feature representation of the precursor because the input is the full length of the precursor sequence rather than the window sequence of the cleavage site. In addition, we integrated an advanced signal peptide prediction tool (SignalP 5.0 [40], https://services.healthtech.dtu.dk/services/SignalP-5.0/). Then, the processing procedure of the complete precursor sequence was followed as the Insect model of NeuroPred, and the candidate samples were fed into our model. Using the convolution layers with different scale kernels (1 and 3) could obtain local features of the windows of 18 amino acids in two different scales and the average pooling layer was used to obtain the glob representation of the windows. In the next stage, position-wise fully connected layers were used for mapping the embedding of cleavage sites and non-cleavage sites to the classification space.

3.2. Ablation study of DeepNeuropePred

To evaluate the contribution of the CNN component (two convolutional layers) and esm1_t12_85M_UR50S module, we set up the ablative configurations. When the esm1_t12_85M_UR50S module was removed, the neuropeptide precursor sequences were encoded using one-hot encoding. When the two CNN layers were removed, a linear layer was used as a replacement. Five-fold cross-validation was used to evaluate the performance of DeepNeuropePred, DeepNeuropePred w/o ESM, and DeepNeuropePred w/o CNN on the training dataset. The entire training dataset was randomly split into five subsets containing the same number of chains. The results of the ablative experiment of neuropeptide cleavage site prediction are shown in Table 1.

It is easy to see that the protein language model (esm1_t12_85M_UR50S) obviously plays a very critical role in the neuropeptide cleavage site prediction. The ACC and MCC scores of DeepNeuropePred w/o ESM and DeepNeuropePred increased from 0.81 to 0.91, and 0.63–0.75, respectively. DeepNeuropePred achieved a Precision score of 0.89, a Recall score of 0.881, and an F1-score of 0.87 which was about 10% (0.79),3.0% (0.83), and 6.0% (0.81) higher than DeepNeuropePred w/o ESM. These results illustrate that the protein language module was also important for the neuropeptide cleavage site prediction. When the input part of ESM was removed, the performance of DeepNeuropePred degraded significantly. The ACC and MCC scores of DeepNeuropePred w/o CNN and DeepNeuropePred increased from 0.86 to 0.91, and 0.67–0.75, respectively. From this ablation experiment, it can be proved that the semantic representation based on the pre-trained model such as ESM can improve the prediction of neuropeptide cleavage sites. To make a fair and robust comparison, we further compared DeepNeuropePred with other state-of-the-art methods (four models from NeuroPred) on the independent test set.

3.3. Comparison with the state-of-the-art methods

We compared DeepNeuropePred with the existing models such as the Motif, Mammalian, Mollusc, and Insect models from the NeuroPred server. It should be emphasized that the test neuropeptide precursors are less 40% identity with the training dataset, which ensures the fairness of the comparison. As shown in Table 2, we see that DeepNeuropePred achieved the highest accuracy of 0.87, followed by Motif (0.80), Mammalian (0.78), Insect (0.78), and Mollusc (0.77). For the AUPRC

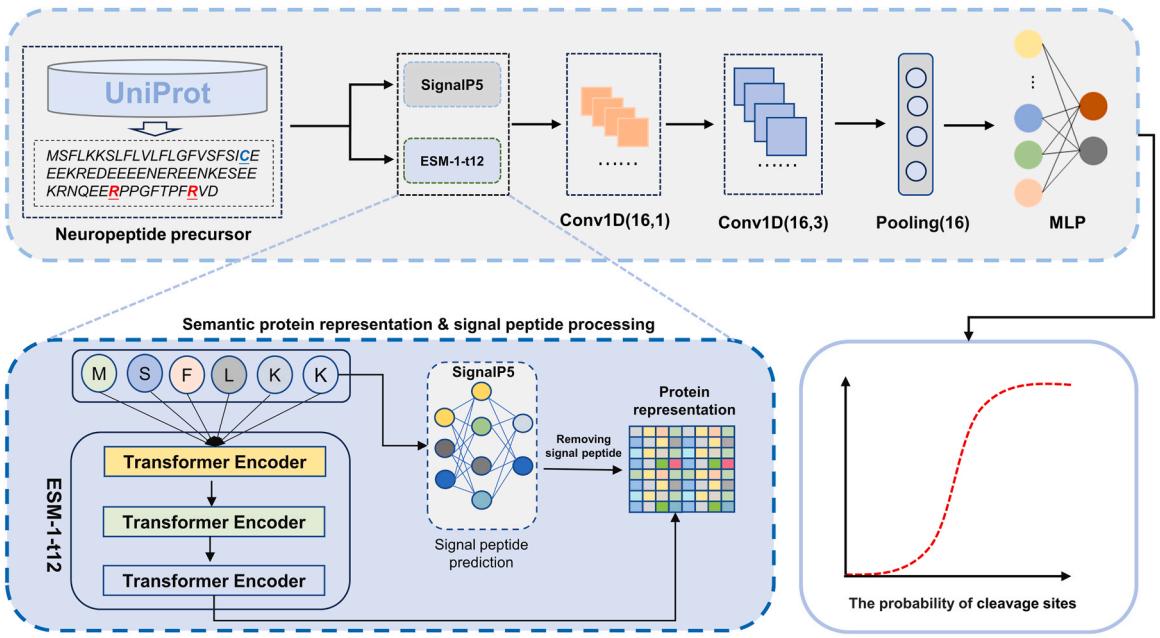

Fig. 1. The flowchart of DeepNeuropePred. The peptide precursor sequence from the UniProt database is initially fed into ESM-1-t12 to obtain a residue-level sequence representation, while SignalP 5.0 software is used to predict the signal peptide's position. A feature representation with a window length of 18 is input into the DeepNeuropePred neural network to ultimately predict the probability of cleavage site at that location.

Table 1
Five-fold cross-validation results of DeepNeuropePred, w/o ESM and w/o CNN.

Model	Precision	Recall	F1-score	MCC	ACC	AUPRC
w/o ESM	0.79	0.83	0.81	0.63	0.81	0.72
w/o CNN	0.84	0.81	0.82	0.67	0.86	0.74
DeepNeuropePred	**0.89**	**0.86**	**0.87**	**0.75**	**0.91**	**0.85**

Table 2
The performance metrics of DeepNeuropePred, Motif, Mammalian, Mollusc, and Insect.

Model	Precision	Recall	F1-score	MCC	ACC	AUPRC
Motif	0.74	0.83	0.78	0.56	0.80	0.48
Mammalian	0.70	0.75	0.72	0.45	0.78	0.56
Insect	0.69	0.74	0.71	0.43	0.78	0.49
Mollusc	0.71	0.79	0.75	0.49	0.77	0.64
DeepNeuropePred	**0.81**	**0.84**	**0.82**	**0.65**	**0.87**	**0.78**

score, DeepNeuropePred outperformed Motif by 30%, Mammalian by 22%, Insect by 29%, and Mollusc by 14%. Furthermore, DeepNeuropePred achieved the highest scores in Precision, Recall, MCC, and F1-score, indicating that it outperforms the four models from NeuroPred in predicting the cleavage sites of neuropeptides.

In all test proteins, the positive samples of the cleavage site are 68 and the negative samples are 247(Positive: Negative=1:4) which is an unbalanced test set. However, the AUC score is insensitive to unbalanced data sets, and the performance of the model can be well evaluated. As shown in Fig. 2, DeepNeuropePred achieved the highest AUC score of 0.916, followed by Mammalian (0.857), Insect (0.850), Mollusc (0.842) and Motif (0.826), which are 6.9%, 7.8%, 8.8% and 10.9% higher respectively. Intuitively, the Motif model was based on some specific patterns, and the generalization ability is insufficient compared to other methods (Mammalian, Insect, and Mollusc). Furthermore, Mammalian, Insect, and Mollusc were trained by the specific domain datasets and these models always had a good performance for the similar precursors. However, our model proved its robustness through independent test precursors. These results also demonstrated the strong representation

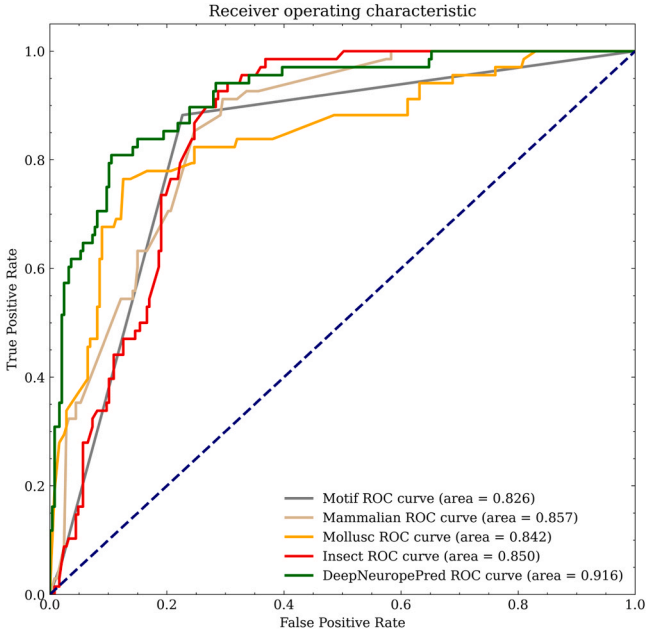

Fig. 2. The Area Under the Receiver Operating Characteristic Curve of DeepNeuropePred, Motif, Mammalian, Mollusc, and Insect.

capability of DeepNeuropePred for both neuropeptide precursors and cleavage sites.

3.4. *Visualization of features extracted by DeepNeuropePred*

DeepNeuropePred is expected to capture meaningful patterns between neuro-peptide and non-neuropeptide cleavage sites. To investigate whether the DeepNeuropePred model has learned to encode cleavage site classifying attributes in its feature representation, we used all datasets of cleavage sites and non-cleavage sites and projected the learned embeddings of the esm1_t12_85M_UR50S, convolutional layers

1 and 2 using t-distributed stochastic neighbor embedding (t-SNE) algorithm[41], which was decomposed into two dimensions. The parameters for t-SNE were set to a perplexity of 10 and 1000 iterations for the optimization. The results, as shown in Fig. 3, clearly indicate that the embedding vectors of the Transformer blocks are unable to distinguish between cleavage sites and non-cleavage sites. Different convolutional layers enhance the inter-class distance between them, which indicates that local information enhancement is important for predicting cleavage sites. It can be concluded from Fig. 3 that different layers can encode and capture features at different levels from the embedding feature of esm1_t12_85M_UR50S.

3.5. Case study comparing neuropeptide mass spectrometry data with DeepNeuropePred

Experimental Methods such as Mass Spectrometry play an important role in neuropeptide cleavage site discovery. To further validate the performance of DeepNeuropePred, we assessed its accuracy using neuropeptide mass spectrometry data. Liessem et al.[42] conducted transcriptome analysis of the central nervous system of Carausius morosus, identified 5 novel neuropeptide precursors (Table S2) in C. morosus through mass spectrometry. These neuropeptide precursors have no obvious homology with known neuropeptide precursors in other insects (Carausius neuropeptide-like precursor 1, HanSolin, PK-like1, PK-like2, RFLamide). The data processing pipeline followed that of the Insect model of NeuroPred, resulting in a total of 61 eligible sites, including 16 neuropeptide cleavage sites and 45 non-cleavage sites. After rigorous evaluation, DeepNeuropePred achieved an accuracy of 80.3% in predicting neuropeptide cleavage and non-cleavage sites. Further analysis of some failed cases revealed a few sites that were close to the selection threshold, with RFLamide having a predicted probability of 0.475 at the 165th amino acid and Pyrokinin-like 1 with a predicted probability of 0.382 at the 69th amino acid. These results suggest that DeepNeuropePred has taken these positions into consideration as potential cleavage sites.

3.6. Webserver of DeepNeuropePred

For the convenience of academic users, the DeepNeuropePred server is freely available at http://isyslab.info/NeuroPepV2/deepNeuropePred.jsp. Our web services are mainly built with Flask, Redis, and Celery. Flask (https://github.com/pallets/flask) and Celery (https://github.com/celery/celery) are two widely-used Python libraries that offer rich functionalities and APIs to help developers easily build complex asynchronous applications. The DeepNeuropePred webserver architecture is shown in Fig. 4. When using Flask and Celery to build asynchronous tasks, we first create a Flask backend program and configure Celery's message broker and result backend within the application. We then define a Celery task that takes two parameters and simulates the execution of a neural peptide cleavage site prediction task, with the Celery worker listening to the task queue. The Flask application is started and asynchronously calls the Celery task, immediately returning the task ID to the frontend. The task execution process occurs in the background, asynchronously executed by the Celery worker. Once the task is completed, the Celery worker stores the result in the result backend, which can be obtained by a frontend request. Throughout the process, Flask and Celery work hand-in-hand, with Flask being responsible for receiving client requests and asynchronously calling the Celery task, while Celery executes the task and stores the result in the result backend. This approach improves the response speed and performance of our application while ensuring the reliability and stability of tasks. For the DeepNeuropePred inference, the PyTorch framework (https://pytorch.org/) is used to construct the neural network.

4. Conclusion

In this study, we introduce DeepNeuropePred, a transfer learning method for detecting cleavage sites in neuropeptide precursors. Our approach was tested on a more rigorous dataset and offers the distinct advantage of being able to model the long-distance representation of the entire sequence, as opposed to just the local window feature of the cleavage site. In independent tests, DeepNeuropePred outperformed the existing model from the NeuroPred server. To the best of our knowledge, this is the first application of a transfer learning algorithm to neuropeptide cleavage site prediction. The use of transfer learning offers new possibilities for predicting cleavage sites in neuropeptide precursors with limited training datasets. While DeepNeuropePred has shown promising results compared to the four models of NeuroPred, the annotation data for neuropeptide cleavage sites remains insufficient. Due to the substantial need for labeled datasets in neural network models, the lack of data also hinders the improvement of model performance. In the future, we will develop new strategies to improve the accuracy of neuropeptide cleavage site prediction.

Funding

This work was supported by National Natural Science Foundation of China under Grant 62172172, Huazhong University of Science and

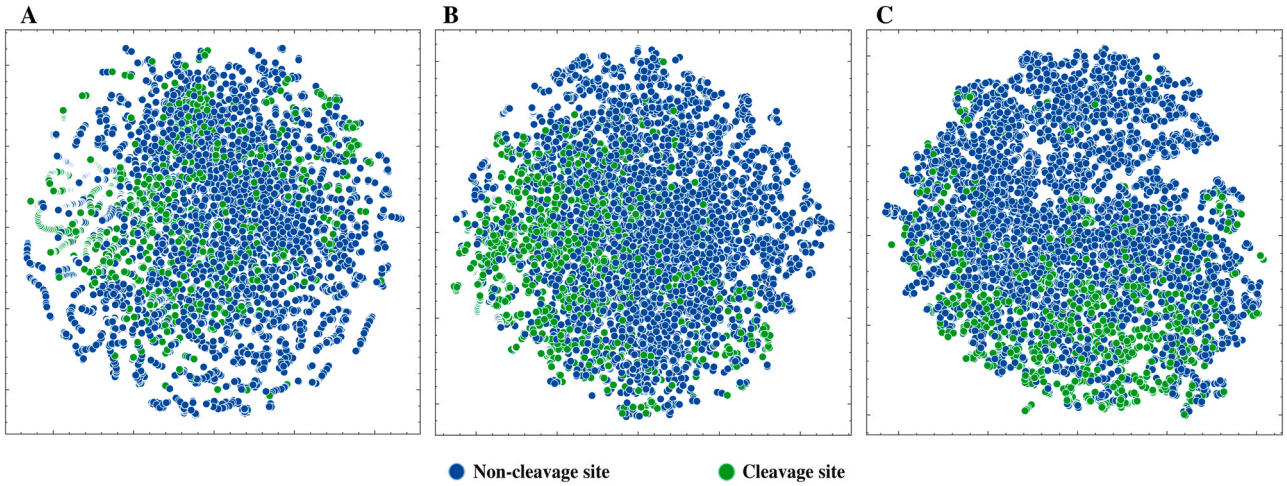

Fig. 3. The neuropeptide cleavage sites are represented in the output embeddings of the esm1_t12_85M_UR50S (A), convolutional layer 1(B), and convolutional layer 2(C), visualized here with t-SNE.

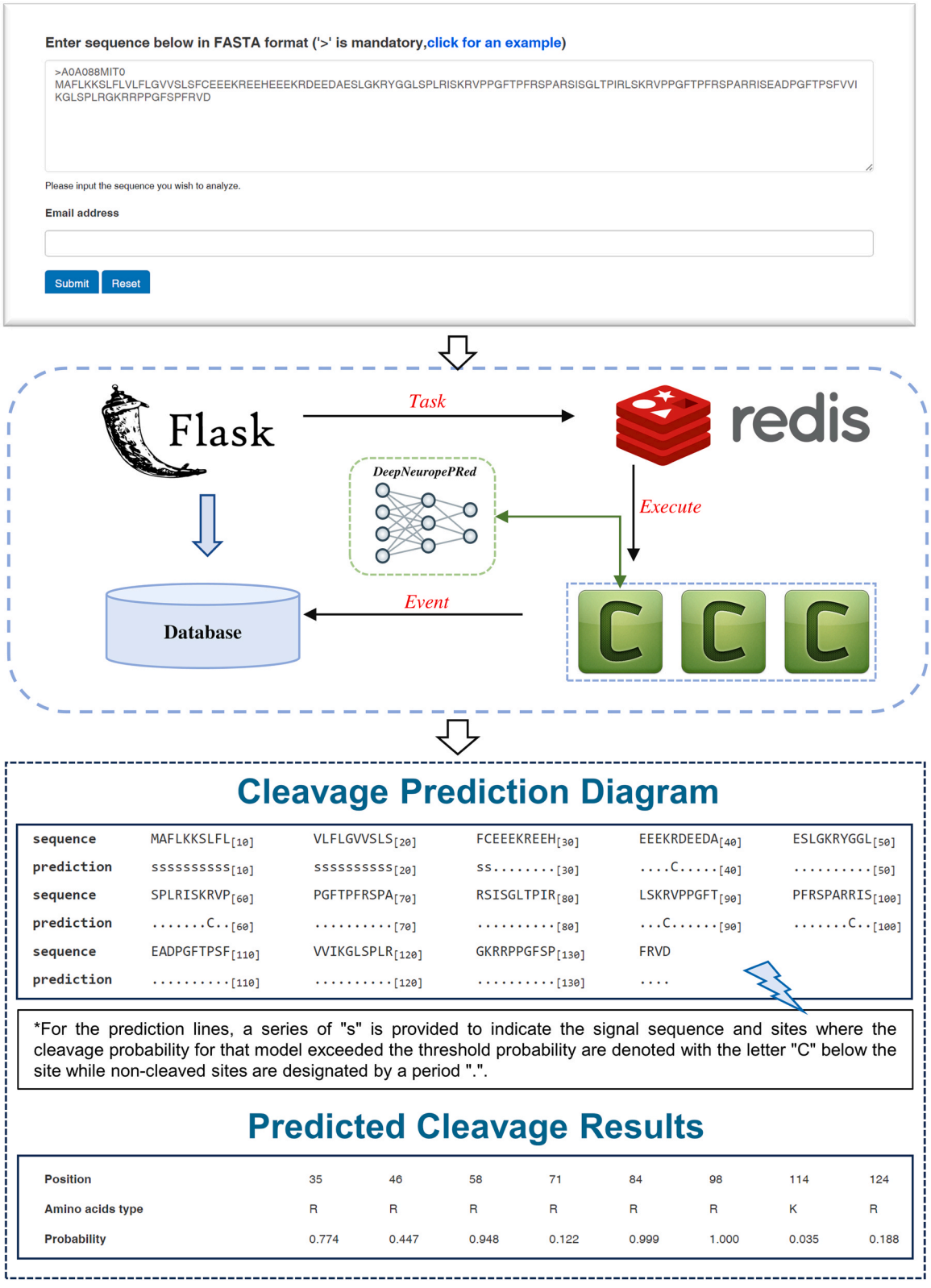

Fig. 4. The front-end and back-end architecture of the DeepNeuropePred webserver. The upper part shows the input information for the webpage, the middle part displays details about the website's backend, and the bottom part exhibits the output of the prediction results.

Technology Independent Innovation Fund-COVID-19 Special Project under Grant 2020kfyXGYJ060, and Scientific Research Start-up Foundation of Binzhou Medical University under Grant BY2020KYQD01.

CRediT authorship contribution statement

Conceptualization, L.W., Z.X. and Y.W.; Data curation, L.W. and Y. W.; Formal analysis, L.W., Z.X., Z.Z. and Y.W.; Funding acquisition, Z.X. and Y.W.; Methodology, L.W., Z.Z. and Y.W.; Project administration, Z.

X. and Y.W.; Software, L.W.; Writing – original draft, L.W., Z.X., Z.Z. and Y.W.; Writing – review & editing, L.W., Z.X., Z.Z. and Y.W. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

Thanks to the Facebook Research team for providing the pre-trained weights for the transformer protein language models.

Conflict of Interest

The authors declare no conflict of interest.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.12.004.

References

[1] Mendel HC, Kaas Q, Muttenthaler M. Neuropeptide signalling systems - an underexplored target for venom drug discovery. Biochem Pharm 2020;181:114129.
[2] Burbach JP. What are neuropeptides? Methods Mol Biol 2011;789:1–36.
[3] Wang Y, Wang M, Yin S, et al. NeuroPep: a comprehensive resource of neuropeptides. Database 2015;2015:bav038.
[4] Hokfelt T, Broberger C, Xu ZQ, et al. Neuropeptides–an overview. Neuropharmacology 2000;39:1337–56.
[5] Sobrino Crespo C, Perianes Cachero A, Puebla Jimenez L, et al. Peptides and food intake. Front Endocrinol 2014;5:58.
[6] Shahjahan M, Kitahashi T, Parhar IS. Central pathways integrating metabolism and reproduction in teleosts. Front Endocrinol 2014;5:36.
[7] Kormos V, Gaszner B. Role of neuropeptides in anxiety, stress, and depression: from animals to humans. Neuropeptides 2013;47:401–19.
[8] Nassel DR, Zandawala M. Recent advances in neuropeptide signaling in Drosophila, from genes to physiology and behavior. Prog Neurobiol 2019;179:101607.
[9] Nassel DR. Neuropeptides in the nervous system of Drosophila and other insects: multiple roles as neuromodulators and neurohormones. Prog Neurobiol 2002;68:1–84.
[10] Holmgren S, Jensen J. Evolution of vertebrate neuropeptides. Brain Res Bull 2001;55:723–35.
[11] Caers J, Verlinden H, Zels S, et al. More than two decades of research on insect neuropeptide GPCRs: an overview. Front Endocrinol 2012;3:151.
[12] Southey BR, Rodriguez-Zas SL, Sweedler JV. Prediction of neuropeptide prohormone cleavages with application to RFamides. Peptides 2006;27:1087–98.
[13] Baggerman G, Cerstiaens A, De Loof A, et al. Peptidomics of the larval Drosophila melanogaster central nervous system. J Biol Chem 2002;277:40368–74.
[14] Baggerman G, Boonen K, Verleyen P, et al. Peptidomic analysis of the larval Drosophila melanogaster central nervous system by two-dimensional capillary liquid chromatography quadrupole time-of-flight mass spectrometry. J Mass Spectrom 2005;40:250–60.
[15] Predel R, Wegener C, Russell WK, et al. Peptidomics of CNS-associated neurohemal systems of adult Drosophila melanogaster: a mass spectrometric survey of peptides from individual flies. J Comp Neurol 2004;474:379–92.
[16] Hummon AB, Amare A, Sweedler JV. Discovering new invertebrate neuropeptides using mass spectrometry. Mass Spectrom Rev 2006;25:77–98.
[17] Hummon AB, Huang HQ, Kelley WP, et al. A novel prohormone processing site in Aplysia californica: the Leu-Leu rule. J Neurochem 2002;82:1398–405.
[18] Amare A, Hummon AB, Southey BR, et al. Bridging neuropeptidomics and genomics with bioinformatics: prediction of mammalian neuropeptide prohormone processing. J Proteome Res 2006;5:1162–7.
[19] Hummon AB, Hummon NP, Corbin RW, et al. From precursor to final peptides: a statistical sequence-based approach to predicting prohormone processing. J Proteome Res 2003;2:650–6.
[20] Southey BR, Sweedler JV, Rodriguez-Zas SL. Prediction of neuropeptide cleavage sites in insects. Bioinformatics 2008;24:815–25.
[21] Shi Q, Chen W, Huang S, et al. Deep learning for mining protein data. Brief Bioinform 2021;22:194–218.
[22] He Y, Shen Z, Zhang Q, et al. A survey on deep learning in DNA/RNA motif mining. Brief Bioinform 2021;22.
[23] Xu J, Li F, Leier A, et al. Comprehensive assessment of machine learning-based methods for predicting antimicrobial peptides. Brief Bioinforma 2021;22:bbab083.
[24] Shi Q, Chen W, Huang S, et al. DNN-Dom: predicting protein domain boundary from sequence alone by deep neural network. Bioinformatics 2019;35:5128–36.
[25] Devlin J., Chang M.-W., Lee K. et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2019, 4171–4186.
[26] Rives A, Meier J, Sercu T, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci USA 2021;118.
[27] Elnaggar A, Heinzinger M, Dallago C, et al. ProtTrans: toward understanding the language of life through self-supervised learning. IEEE Trans Pattern Anal Mach Intell 2022;44:7112–27.
[28] Geffen Y, Ofran Y, Unger R. DistilProtBert: a distilled protein language model used to distinguish between real proteins and their randomly shuffled counterparts. Bioinformatics 2022;38:ii95–8.
[29] Alley EC, Khimulya G, Biswas S, et al. Unified rational protein engineering with sequence-based deep representation learning. Nat Methods 2019;16:1315–22.
[30] Bepler T, Berger B. Learning the protein language: evolution, structure, and function. Cell Syst 2021;12:654–69. e653.
[31] Teufel F, Armenteros JJA, Johansen AR, et al. SignalP 6.0 predicts all five types of signal peptides using protein language models. Nat Biotechnol 2022;40:1023.
[32] Thumuluri V, Almagro Armenteros JJ, Johansen AR, et al. DeepLoc 2.0: multi-label subcellular localization prediction using protein language models. Nucleic Acids Res 2022.
[33] Jiang J, Lin X, Jiang Y, et al. Identify bitter peptides by using deep representation learning features. Int J Mol Sci 2022;23.
[34] 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.
[35] Wang L, Zhong H, Xue Z, et al. Res-Dom: predicting protein domain boundary from sequence using deep residual network and Bi-LSTM. Bioinform Adv 2022;2:vbac060.
[36] Wang L., Wang Y. GNN-Dom: an unsupervised method for protein domain partition via protein contact map. In: Bioinformatics Research and Applications: 18th International Symposium, ISBRA 2022, Haifa, Israel, November 14–17, 2022, Proceedings. 2023, p. 286–294. Springer.
[37] Wang L, Zhong HL, Xue ZD, et al. Improving the topology prediction of a-helical transmembrane proteins with deep transfer learning. Comput Struct Biotechnol J 2022;20:1993–2000.
[38] UniProt C. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res 2021;49:D480–9.
[39] Fu L, Niu B, Zhu Z, et al. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 2012;28:3150–2.
[40] Almagro Armenteros JJ, Tsirigos KD, Sonderby CK, et al. SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat Biotechnol 2019;37:420–3.
[41] Van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res 2008:9.
[42] Liessem S, Ragionieri L, Neupert S, et al. Transcriptomic and neuropeptidomic analysis of the stick insect, Carausius Morosus. J Proteome Res 2018;17:2192–204.