




NeuroPred-PLM: an interpretable and robust model for neuropeptide prediction by protein language model

Lei Wang , Chen Huang, Mingxia Wang, Zhidong Xue  and Yan Wang 

Corresponding authors. Yan Wang, Institute of Medical Artificial Intelligence, Binzhou Medical University, Yantai, Shandong 264003, China.

E-mail: yanw@hust.edu.cn; Zhidong Xue, School of Software Engineering, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China.

E-mail: zdxue@hust.edu.cn

Abstract

Neuropeptides are a diverse and complex class of signaling molecules that regulate a variety of biological processes. Neuropeptides provide many opportunities for the discovery of new drugs and targets for the treatment of a wide range of diseases, and thus, computational tools for the rapid and accurate large-scale identification of neuropeptides are of great significance for peptide research and drug development. Although several machine learning-based prediction tools have been developed, there is room for improvement in the performance and interpretability of the proposed methods. In this work, we developed an interpretable and robust neuropeptide prediction model, named NeuroPred-PLM. First, we employed a language model (ESM) of proteins to obtain semantic representations of neuropeptides, which could reduce the complexity of feature engineering. Next, we adopted a multi-scale convolutional neural network to enhance the local feature representation of neuropeptide embeddings. To make the model interpretable, we proposed a global multi-head attention network that could be used to capture the position-wise contribution to neuropeptide prediction via the attention scores. In addition, NeuroPred-PLM was developed based on our newly constructed NeuroPep 2.0 database. Benchmarks based on the independent test set show that NeuroPred-PLM achieves superior predictive performance compared with other state-of-the-art predictors. For the convenience of researchers, we provide an easy-to-install PyPi package (<https://pypi.org/project/NeuroPredPLM/>) and a web server (<https://huggingface.co/spaces/isyslab/NeuroPred-PLM>).

Keywords: neuropeptide prediction; interpretable model; deep learning; protein language model

INTRODUCTION

Neuropeptides are diverse and complex classes of signaling (transmitter) molecules that modulate almost every physiological process and behavior in living species [1]. They generally consist of less than 100 amino acids, produced from larger precursor molecules by a series of post-translational processing [2, 3]. Neuropeptides not only act via nervous system but also act peripherally through endocrine systems, where they regulate various 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]. Therefore, they are implicated in multiple disease processes, and neuropeptide signaling system is a therapeutic target for the treatment of sleep disorders, autism, depression, heart failure, obesity, diabetes, high blood pressure, epilepsy and other diseases [1, 4, 8, 9].

Neuropeptides are also valuable biomarkers and diagnostic probes for prospective disease diagnosis and prognosis.

Traditional experimental methods, such as mass spectrometry and liquid chromatography [10–12], can accurately identify new neuropeptides, and these methods are expensive and time-consuming. So far, several computational methods for predicting neuropeptides have been developed [13–17]. Methods that use genomic approaches to identify neuropeptides rely on sequences that are known to have a common pattern in neuropeptide families (e.g. the RFamide motif), which does not work for some neuropeptide families that lack a distinct pattern [17, 18]. Agrawal *et al.* [13] proposed a machine learning-based computational method, NeuroPIpred, which predicted neuropeptides in insects by using different machine learning methods combined with physicochemical properties and structural features. Recently, Bin *et al.* [14] proposed an integrated method tool called PredNeuroP

Lei Wang is a PhD candidate at the School of Life Science and Technology, Huazhong University of Science and Technology. He is also a research assistant at the institute of medical artificial intelligence, Binzhou Medical University, China. His research interests primarily focus on protein representation learning and structure prediction.

Chen Huang is a master student at the School of Software Engineering, Huazhong University of Science and Technology. His research interests primarily focus on deep learning.

Mingxia Wang is a research assistant at the institute of medical artificial intelligence, Binzhou Medical University, China. Her research interests primarily focus on bioinformatics methodology research.

Zhidong Xue is a professor at the School of Software Engineering, Huazhong University of Science and Technology. His research interests primarily focus on machine learning, bioinformatics, biomedical big data analysis and protein structure prediction.

Yan Wang is a professor at the Institute of Medical Artificial Intelligence, Binzhou Medical University, China. Her research interests primarily focus on bioinformatics, biomedical big data analysis and protein structure prediction.

Received: September 30, 2022. **Revised:** January 3, 2023. **Accepted:** February 15, 2023

© The Author(s) 2023. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

for identifying neuropeptides based on a two-layer stacked framework. Hasan et al. [15] developed a machine learning-based meta-predictor called NeuroPred-FRL, with feature representation learning. Jiang et al. [16] developed an interpretable two-layer stacking model, NeuroPred-Fuse, for predicting neuropeptides by fusing various sequence-derived features and feature selection methods.

In recent years, deep learning-based methods have been widely used on complex biology problems [19–22]. Compared with traditional machine learning models, deep learning-based methods can better learn the distribution of data and do not require complex feature engineering processing. However, prediction methods based on deep learning have two main limitations: (i) deep learning-based methods require a large amount of labeled data to fit the data distribution, but the number of annotations for non-redundant neuropeptides is relatively small; (ii) algorithms based on deep learning or machine learning are poorly interpretable. Recently, pre-trained models based on protein sequences have developed very rapidly [23, 24], which can well obtain the contextual information of protein sequences, no longer need to design complex protein feature engineering processing methods and have been used to solve many downstream problems of biological sequence prediction such as protein secondary structure prediction [25], protein subcellular localization [26] and protein domain boundary [27]. In recent years, biological information algorithms based on deep learning or machine learning often focus on performance and ignore the interpretability of the algorithm, which brings challenges to understanding the relationship between algorithms and biological problems [28]. NeuroPred-FRL employs SHapley Additive exPlanations (SHAP) [29] to explain the relationship of each feature to positive or negative class prediction. This method assigns each feature an SHAP score representing its impact on the predictions of the trained model. NeuroPred-Fuse is a feature selection-based method that identifies the most important features that are beneficial for neuropeptide classification. Machine learning-based neuropeptide prediction methods (NeuroPred-FRL and NeuroPred-Fuse) use feature selection to explore feature combinations that contribute significantly to neuropeptide prediction, but this interpretability is based on artificially constructed features, which is difficult to identify the contribution of amino acids embedding features in different position of a certain neuropeptide sequence.

To overcome the above problems, we proposed an interpretable and robust neuropeptide prediction method NeuroPred-PLM based on protein language models and our newly updated neuropeptide database (NeuroPep 2.0, unpublished manuscript). NeuroPep 2.0 nearly doubled the number of neuropeptide entries in the first version of the database, based on which all the current neuropeptide prediction tools were developed. The protein language model (ESM) [23] was used to alleviate the problem of sequence representation of small samples. Next, we used a multi-scale convolutional neural network to obtain local correlation features of neuropeptides. To enhance the interpretability of neuropeptide prediction models, we adopted a multi-head attention mechanism to capture the position-wise contribution to neuropeptide prediction via the attention scores. The independent test shows that NeuroPred-PLM achieves better performance than other state-of-the-art predictors. To facilitate the community-wide use of our proposed method, we built a user-friendly software package at <https://github.com/ISYSLAB-HUST/NeuroPred-PLM> and a web server (<https://huggingface.co/spaces/isysslab/NeuroPred-PLM>).

MATERIALS AND METHODS

Datasets

We mainly collected experimentally validated neuropeptides from the NeuroPep 2.0 database. A total of 11 282 experimentally validated neuropeptide sequences were obtained, of which 5333 sequences were newly added data in version 2.0. Then, a series of filter steps were adopted. First, neuropeptides with lengths between 5 and 100 were retained. Second, CD-HIT (version:4.8.1) [30] with a threshold of 0.9 was applied to the remaining samples to exclude the sequences that have more than 90% identity with other sequences. In data processing, the length of the neuropeptide (5–100) and the CD-HIT parameter (0.9) are consistent with the SOTA methods such as NeuroPred-FRL and PredNeuroP. After the above filtering steps, 4463 neuropeptides were still retained. To ensure a fair comparison, the independent neuropeptide test data were all from the newly added neuropeptides in NeuroPep 2.0, and 444 neuropeptides were randomly selected, accounting for 10% of the total 4463 neuropeptides. The screening method of negative samples was consistent with NeuroPred-FRL. Negative samples were extracted with a length distribution similar to the positive samples from the UniProt database [31], which resulted in 4463 negative samples. The non-neuropeptide test sets were randomly selected from the 4463 negative samples, and 444 sequences (10%) were randomly selected. All the training data and test data can be freely accessed at <https://github.com/ISYSLAB-HUST/NeuroPred-PLM/tree/main/dataset>.

Methods

The model architecture of NeuroPred-PLM is shown in Figure 1. First, the semantic representation of peptides is obtained through a 12-layer protein language model (ESM). Next, a projection layer is used to project the features of the high-dimensional space into the low-dimensional space. Then, multi-scale-based convolutional layers can extract locally relevant features of neuropeptides. A global multi-head attention network is used to focus on the position-wise contributions of neuropeptides. Finally, the classification layer maps the global feature representation to the classification space and outputs the predicted probability scores of neuropeptides. The functionality of the three core components (ESM blocks, Convolutional layer and Global multi-head attention layer) in NeuroPred-PLM is introduced in detail in Table S2 available online at <http://bib.oxfordjournals.org/>.

Feature extraction based on the protein language model

The development of pre-training language models has brought the research in protein representation to a new stage without manual labeling. The representation of protein sequences can be learned from massive unlabeled protein sequences, and downstream tasks can be significantly improved. Using pre-training language models can reduce the risk of overfitting on small training data, which is equivalent to a regularization method.

Here, we adopt a transformer-based self-supervised language model called ESM which includes more than 85 million parameters (12-layer transformer). The ESM model can accept a protein sequence and generate dynamic embedding with $L * 768$, where L is the peptide length. When the model is optimized, the first 9 layers of the ESM backbone and the embedding layers are frozen, and the last three layers can be fine-tuned.

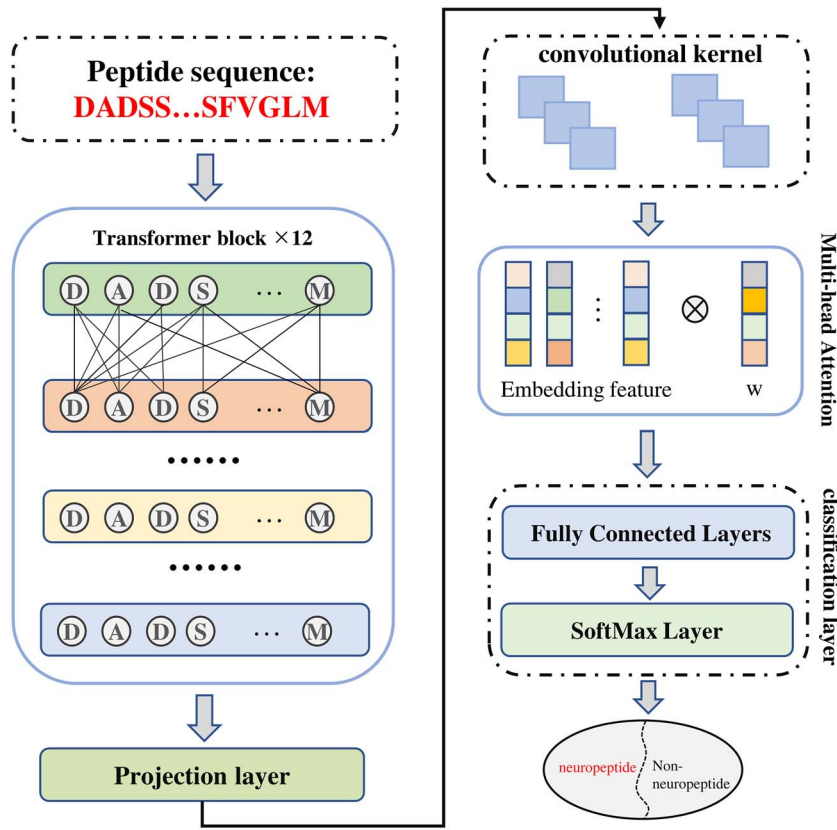


Figure 1. The flowchart of NeuroPred-PLM.

Dimension reduction with projection layer and convolutional neural network

The feature embedding dimension of the protein language model is too large (up to 768) to be suitable for fine-tuning models with small biological samples. Therefore, it is indispensable to reduce the feature dimension of the protein language model. Traditional methods of dimension reduction include feature selection, matrix factorization, manifold learning and neural networks. To build an end-to-end differentiable prediction model, we adopted a learnable projection matrix to reduce the dimension of the embedding. Given the embedding feature of the language model $u \in R^{m \times l}$, we can obtain new embeddings via

$$z = P * u + b, \quad (1)$$

where $P \in R^{n \times m}$ is the learnable projection matrix, $b \in R^n$ is the bias and $z \in R^{n \times l}$. Considering the strong local correlation of the peptide sequence, 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 following formula:

$$c1 = \text{Dropout}(\text{conv}(z, k1, n1)) \quad (2)$$

$$c2 = \text{Dropout}(\text{conv}(c1, k2, n2)), \quad (3)$$

where **conv** is the convolutional function, $k1, k2$ is the convolutional kernel size and $n1, n2$ is the number of convolutional kernels. The dropout function is used to reduce the risk of overfitting of the model, which works by randomly disabling neurons and their corresponding connections, prevents the network from

becoming overly dependent on a single neuron and forces all neurons to learn to generalize better [32].

Global multi-head attentive neural network and classification layer

In the field of natural language processing, position-wise pooling is generally used to obtain the global representation of the sentence level, but this method ignores the position-wise contribution of neuropeptides. In view of the above problems, we proposed to use the global attention-based network to obtain the position-wise contribution to neuropeptide prediction via the attention scores. The global feature of attention-based networks can be obtained by the following formula:

$$\alpha = \text{softmax}(w^T c2) \quad (4)$$

$$\hat{Z} = \sum_i^L \alpha \cdot c2, \quad (5)$$

where $w \in R^{h \times n}$ is the learnable global vector, $\alpha \in R^{h \times L}$ is the attention score, the number of the head is h and \hat{Z} is the global representation of peptides. To project the \hat{Z} vector into the feature space of neuropeptides and non-neuropeptides, the classification layer consists of one-layer fully connected layer and SoftMax layer. The predicted neuropeptide probabilities are normalized between 0 and 1 after the final SoftMax layer.

Loss function

NeuroPred-PLM takes the entire peptide sequence of L residues as input and outputs a probability score, where the score indicates whether the input peptide sequence belongs to neuropeptide. The cross-entropy loss function is used as the optimization goal of the neural network, and the label smoothing technology [33] is used

to reduce the risk of overfitting of the model. The loss function is defined as

$$H(y, p) = \sum_{k=1}^K -y_k^{LS} \log(p_k) \quad (6)$$

$$y_k^{LS} = y_k(1 - \alpha) + \alpha/K, \quad (7)$$

where K is 2, α is the label smoothing of parameter, y and p are the label and the predicted possibility score, respectively; ground truth label y reflects if it belongs to neuropeptide (1) or non-neuropeptide (0) and p is the output of the model. In this work, 0.5 was selected as the threshold, probability scores greater than 0.5 were considered as neuropeptides and those less than 0.5 were considered as non-neuropeptides.

Performance metrics

To evaluate the performance of the model, seven common metrics including precision (Pre), recall (Rec), F1 score (F1), sensitivity (SN), specificity (SP), accuracy (ACC) and Matthew's correlation coefficient (MCC) were calculated and defined as follows:

$$\text{Pre} = \frac{TP}{TP + FP} \quad (8)$$

$$\text{Rec} = \frac{TP}{TP + FN} \quad (9)$$

$$\text{F1} = 2 \frac{\text{Pre} * \text{Rec}}{\text{Pre} + \text{Rec}} \quad (10)$$

$$\text{SN} = \frac{TP}{TP + FN} \quad (11)$$

$$\text{SP} = \frac{TP}{FP + TN} \quad (12)$$

$$\text{ACC} = \frac{TP + TN}{TP + TN + FP + FN} \quad (13)$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}, \quad (14)$$

where TP, TN, FP and FN denote the numbers of true positive, true negative, false positive and false negative samples, respectively.

RESULTS AND DISCUSSION

Hyperparameters tuning and cross-validation

NeuroPred-PLM was built on Python 3.8, einops 0.4, cudatoolkit 11.3 and PyTorch 1.9.0. All model experiments were trained on the NVIDIA GeForce GTX 1080 Ti (11GB). Considering the limitations of GPU memory, the last three layers of the 12-layer ESM (version: 0.4.1) backbone were finetuned with a small learning rate ($2e-6$). Some hyperparameters are needing to be optimized, including projection embedding dimension, convolutional filters, dropout probability, the number of heads, the label smoothing of parameters and the fully connected layer dimension. All hyperparameters (Table S1 available online at <http://bib.oxfordjournals.org/>) were tuned with Neural Network Intelligence (<https://github.com/microsoft/nni>). The grid search was performed according to the hyperparameters in Table S1 available online at <http://bib.oxfordjournals.org/> and the accuracy results of all hyperparameter combinations can be seen in Figure S1 available online at <http://bib.oxfordjournals.org/>. In this work, the optimized hyperparameters are as follows: projection dimension is 24, convolution filters are (24, 3) and (12, 1), dropout probability is 0.1, the number of heads is 12, the label smoothing of the parameter is 0.1 and the dimension of the fully connected layers are 144 and 12, respectively. For the NeuroPred-PLM

backbone, a small learning rate ($5e-4$) was applied with the Adam optimizer [34].

Given that the number of neuropeptides is not particularly sufficient, to prevent overfitting of the model, we used 10-fold cross-validation to evaluate the performance of the model. First, we randomly divide all the training set sample data into 10 parts, randomly select 9 parts as the training set each time and the remaining 1 part as the validation set. The optimal model and parameters are evaluated based on the loss function, and the optimal model is used for the evaluation of independent validation. The results of the 10-fold cross-validation are shown in Table 1. The average accuracy of neuropeptide prediction reached 92.7%, the MCC score reached 0.854, the precision reached 0.926, the recall rate reached 0.928, the F1 score reached 0.927 and the specificity reached 0.926. According to the standard deviation analysis, there is no obvious difference in the model evaluation metrics obtained by training with different fold datasets, indicating that NeuroPred-PLM has good robustness.

Comparison with the state-of-the-art methods for neuropeptide prediction

To further assess the performance of neuropeptide prediction tools, we compared NeuroPred-PLM with the state-of-the-art methods (PredNeuroP, NeuroPred-FRL and NeuroPpred-Fuse) on the independent test set. It should be emphasized that the neuropeptide sequences of the independent test set are all from the newly added entries in the NeuroPep 2.0, which ensures the fairness of the comparison. Among the methods compared, NeuroPpred-Fuse is currently the best neuropeptide prediction model, and to exclude the influence of training data size on the results, we retrained NeuroPpred-Fuse with the same training set as NeuroPred-PLM. Since the source code of NeuroPred-FRL was not provided, we obtained test results from the NeuroPred-FRL web server. Here, PredNeuroP was used as the baseline model. As shown in Table 2, we see that NeuroPred-PLM achieved the highest accuracy of 0.922, followed by NeuroPpred-Fuse (0.905), PredNeuroP (0.864) and NeuroPred-FRL (0.861). The MCC, Recall and F1 scores of NeuroPred-PLM were about 3.2, 3.3 and 1.7% higher than the second-best NeuroPpred-Fuse, respectively. NeuroPred-FRL achieved the highest Precision (0.960), but its Recall was the lowest among the four methods. Overall, the performance of NeuroPred-PLM is significantly better than the other three models.

Visualization of features extracted by NeuroPred-PLM

NeuroPred-PLM can be expected to capture meaningful patterns between neuropeptides and non-neuropeptides. To investigate whether the NeuroPred-PLM model has learned to encode properties of neuropeptides classification in its representations, we use the test set of neuropeptides and non-neuropeptides and project the learned embedding of the Transformer block, projection layer, convolutional layer and multi-head attentive neural network of the NeuroPred-PLM into two dimensions by applying the t-distributed stochastic neighbor embedding (t-SNE) algorithm. The results are shown in Figure 2. It is obvious that the embedding vector of the Transformer block is difficult to distinguish between neuropeptides and non-neuropeptides. The projection layer and the convolutional layer enhance the class spacing between them. Furthermore, the multi-head attentive neural network can clearly capture higher order differences, making it easier to distinguish neuropeptides and non-neuropeptides. It can

Table 1. Performance of NeuroPred-PLM with 10-fold cross-validation on the training set

Fold	MCC	ACC	Pre	Rec	F1	SP	SN
1	0.864	0.932	0.915	0.953	0.933	0.910	0.953
2	0.862	0.930	0.906	0.953	0.929	0.910	0.953
3	0.833	0.917	0.916	0.919	0.932	0.915	0.919
4	0.838	0.919	0.931	0.900	0.921	0.937	0.900
5	0.869	0.934	0.951	0.921	0.933	0.948	0.921
6	0.868	0.934	0.940	0.928	0.914	0.940	0.928
7	0.858	0.929	0.924	0.940	0.921	0.918	0.940
8	0.836	0.918	0.926	0.902	0.937	0.933	0.902
9	0.838	0.919	0.920	0.918	0.913	0.920	0.918
10	0.873	0.936	0.934	0.941	0.934	0.932	0.941
average	0.854±0.014	0.927 ±0.0068	0.926 ±0.012	0.928 ±0.017	0.927 ±0.0079	0.926 ±0.012	0.928 ±0.017

Table 2. Performance comparisons of NeuroPred-PLM with the three representative state-of-the-art methods on the independent test set

Methods	ACC	Pre	Rec	F1	MCC
PredNeuroP	0.864	0.935	0.782	0.852	0.738
NeuroPred-FRL	0.861	0.960	0.757	0.847	0.740
NeuroPpred-Fuse	0.905	0.906	0.908	0.907	0.813
NeuroPred-PLM	0.922	0.907	0.941	0.924	0.845

be clearly concluded that different layers can encode and capture different levels of features from ESM embeddings.

Model interpretation

Deep learning algorithms are promising in many other scientific fields, especially in biological data analysis, because they are very good at discovering complex structures in high-dimensional data. However, deep learning methods are considered as black-box models and lack interpretability in the field of bioinformatics. Numerous studies have shown that different neuropeptide families have different conserved patterns such as the NPY family and PDH family. It is difficult to explain the different positional contributions of neuropeptides by feature selection methods alone. However, NeuroPred-PLM adopts a global multi-head attention mechanism that can capture the position-specific conservation of different neuropeptide families. Here, we take the pigment-dispersing hormone family (PDH) and the Neuropeptide Y family (NPY) as examples to explore the interpretability of the NeuroPred-PLM model.

The neuropeptides of the PDH family consist of 18 amino acids and at least 50% of the amino acid sequence appears to be conserved. These neuropeptides of the PDH family may be involved in the regulation of the insect circadian system [35, 36]. As early as 1988, Rao *et al.* [37] found that many positions of the PDH family have a great influence on its activity, such as positions 3, 4, 11 and 13. To explore whether NeuroPred-PLM could capture these position specificities affecting the activity of neuropeptides of PDH family, we visualized the multi-head attention scores of different positions by heatmap. As shown in Figure 3, different attention heads capture the contribution of various positions along the peptide sequence, among which head 8 and head 12 have a high weight at position 4, head 5 has a high weight near position 11, and head 4, head 5 and head 8 have a high weight on position 13. In addition, we found that positions 2 and 15 in the PDH family may also have an effect on neuropeptide activity based on attention scores.

The NPY family comprises three kinds of peptides—namely NPY [38], peptide YY (PYY) [39] and pancreatic polypeptide (PP) [40]—that act as hormones and/or neurotransmitters/neuromodulators. The peptides of the NPY family consist of 36 amino acid residues [41]. They are highly conserved across species, and the NPY family plays key roles in regulating appetite and food intake, regulating mood and anxiety disorders, regulating stress responses and ethanol intake [42–44]. To figure out the interpretability of the NeuroPred-PLM model for conserved sites in the NPY family, we visualized multi-head attention scores based on non-redundant peptide sequences of length 36 (Figure 4). Head 5 has a high attention score at positions 5 and 35, Pedragosa-Badia *et al.* [45] reported that replacing position 5 (Pro) of NPY with Ala resulted in a 600-fold loss of affinity and Eckard *et al.* [46] reported that position 35 (Arg) of NPY was very important for the neuropeptide receptor affinity. Head 12 has a high attention score at position 31, and Cabrele *et al.* [47] found that the substitution of position 31(Leu) of NPY with Ala resulted in a 1000-fold lower affinity. In addition, Head 4 at position 15 and head 8 at position 16 also learned very important features, which may also reveal other features of NPY family. With the help of two examples of NPY and PDH families, it can be clearly seen that NeuroPred-PLM can learn the amino acid position specificity of neuropeptide family and discover key residues in different families, which shows that the method is highly interpretable via attention scores.

Software and web server development

NeuroPred-PLM provides an independent software package, which can be installed in the PyPI repository (<https://pypi.org/project/NeuroPredPLM/>), supports GPU acceleration mode and can be used for large-scale neuropeptide screening. For the convenience of researchers, we also provide a simple web server based on the hugging face space, which can be accessed directly through <https://huggingface.co/spaces/isyslab/NeuroPred-PLM>. To facilitate batch submission of neuropeptide prediction tasks, we provide a restful API interface. To use GPU accelerated mode,

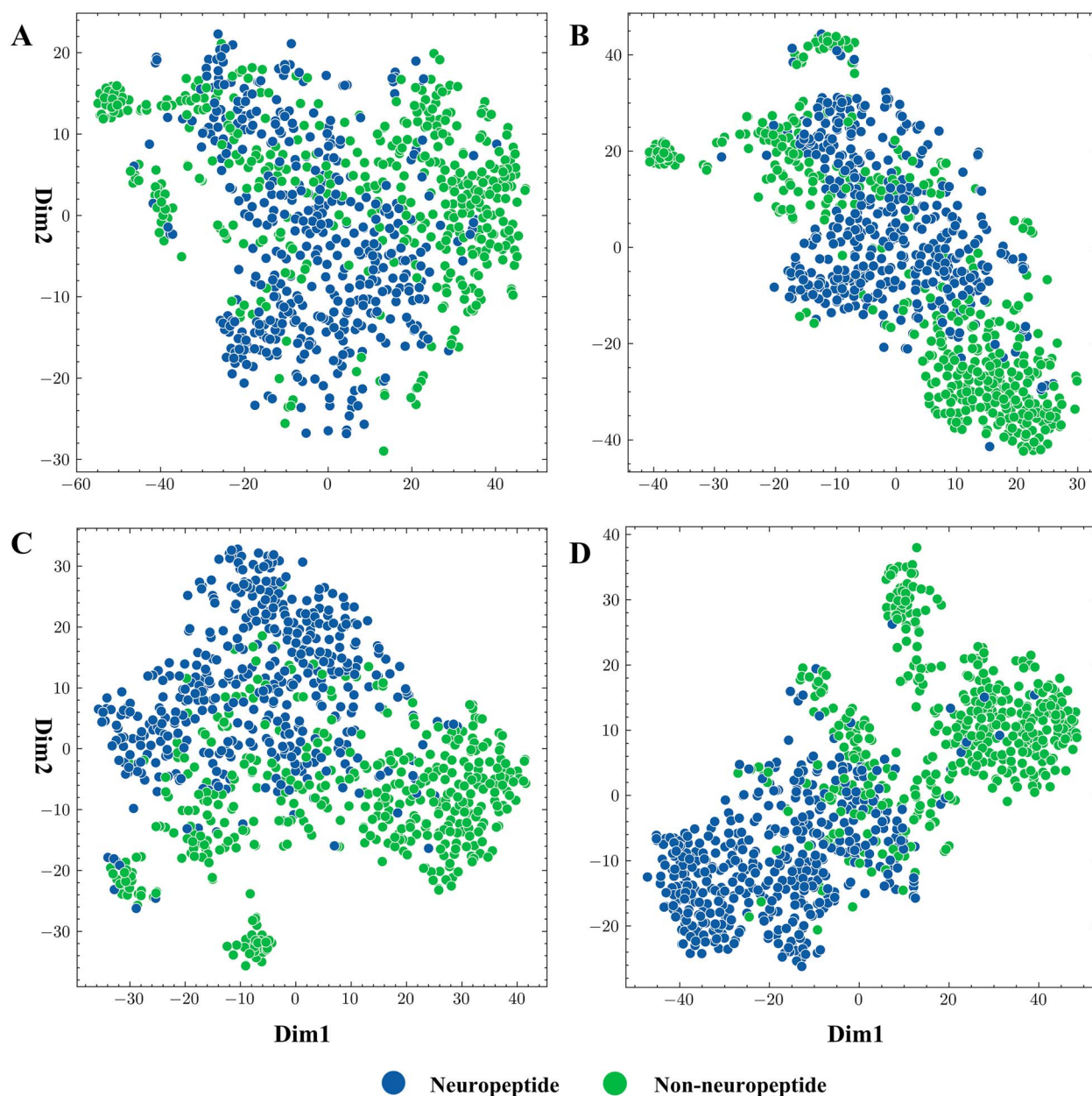


Figure 2. The test set of neuropeptides and non-neuropeptides are represented in the output embeddings of the Transformer (A), projection layer (B), convolutional layer (C) and multi-head attentive neural network (D), visualized here with t-SNE.

we provide a colab notebook (https://colab.research.google.com/github/ISYSLAB-HUST/NeuroPred-PLM/blob/master/notebook/NeuroPred_PLM_test.ipynb).

DISCUSSION AND CONCLUSION

In this work, we proposed an interpretable and robust method, NeuroPred-PLM, using a pre-trained self-supervised language model called ESM, multi-scale convolutional neural networks and global multi-head attentive neural network for neuropeptide prediction. Based on the test results, the proposed NeuroPred-PLM showed superior performance in predicting neuropeptides compared with the state-of-the-art methods.

Several existing machine learning-based neuropeptide prediction models rely on complex feature engineering. PredNeuroP contained two feature groups: (i) amino acid composition (AAC),

NT5 and CT5 sequences and (ii) dipeptide composition (DPC). NeuroPred-Fuse used six different feature encoding schemes to encode the peptide sequences into a feature vector, which included AAC, DPC, GGAP, CTD, ASDC and PSAAC. NeuroPred-FRL employed the 11 feature encodings (BE, AAI, Kmer-AC, KgapAC, PrAC, TPPC, KgapAP, CTF, QSO, GDPC and GTPC). These traditional feature representation methods have some defects, and the pre-designed feature representation methods cannot dynamically represent the semantic information of neuropeptides. Currently, protein representation learning has developed rapidly and is widely used for downstream tasks in many biological fields. Protein language models use self-supervised neural networks to learn protein semantic information, which has a huge advantage in that it can automatically extract protein representations for neuropeptide prediction and can greatly reduce the complexity of feature engineering. In our method, we adopt the ESM model to

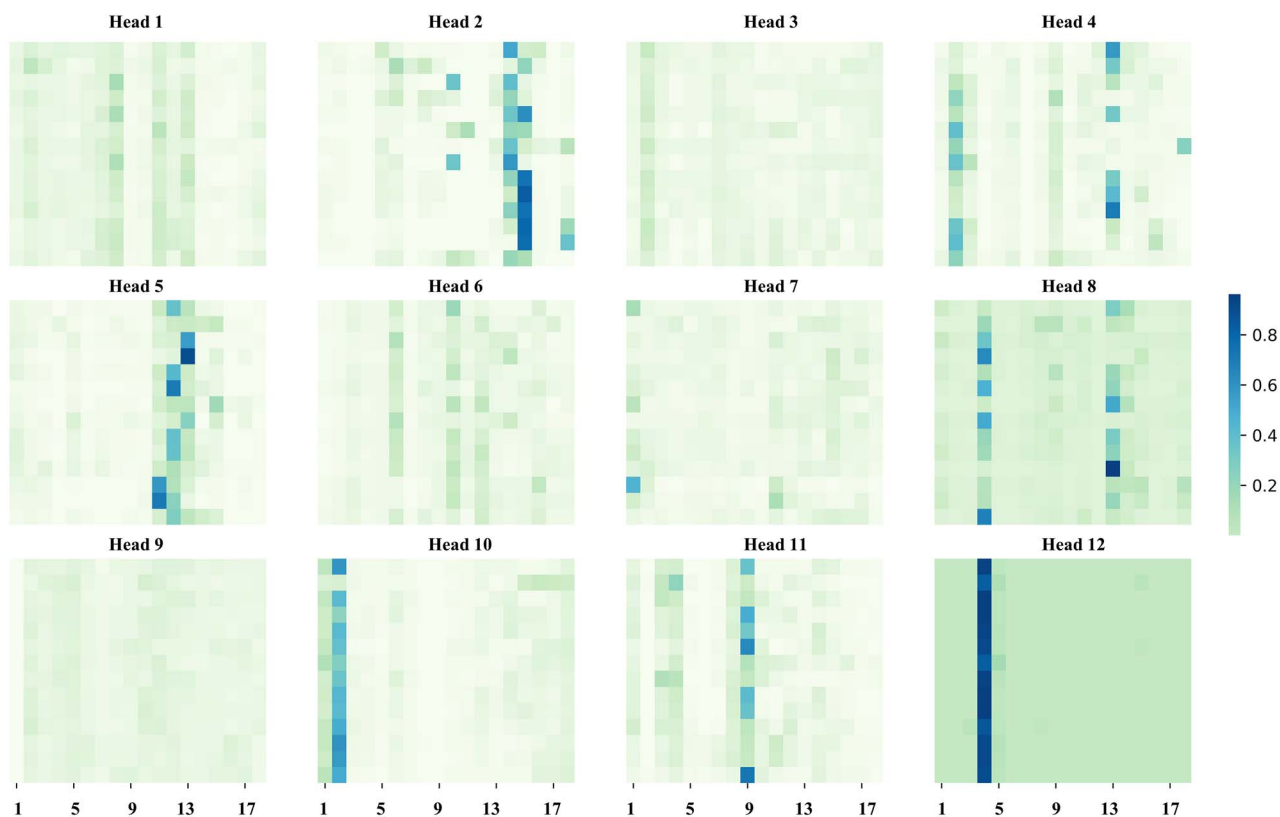


Figure 3. The heatmap of the multi-head attention scores in the PDH family.

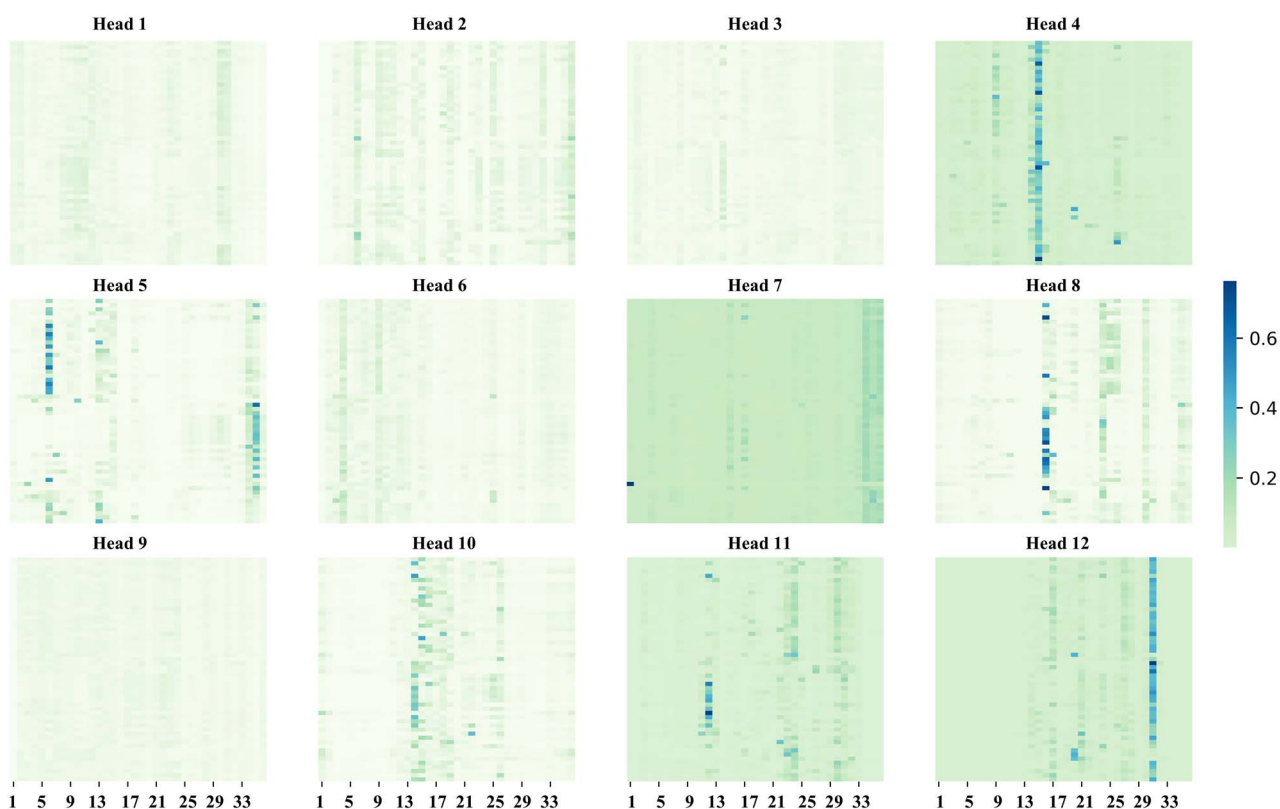


Figure 4. The heatmap of the multi-head attention scores in the NPY family.

extract the dynamic semantic representation of neuropeptides. Compared with traditional feature extraction methods, this method using pre-trained models can well incorporate the contextual information of neuropeptides.

Besides, NeuroPed-PLM also outperforms other machine learning-based neuropeptide prediction methods in interpretability. Machine learning-based neuropeptide prediction methods (NeuroPred-FRL and NeuroPred-Fuse) only explain what features

are meaningful for neuropeptide prediction from the perspective of the contribution of artificially designed features, but it is difficult to identify which amino acids contribute most to the prediction. Our method could capture the contribution of amino acids embedding features in different positions of a certain neuropeptide sequence through the global multi-head attention mechanism. Through the analysis of two neuropeptide families (PDH and NPY family), it is proved that NeuroPred-PLM can well explore the important position residues of neuropeptides. Furthermore, attention mechanism will be helpful for the development of selective neuropeptide receptors, exploring the contribution of key positions to the binding affinity of neuropeptides and regulating ligand preference for receptors, which are of great value for neuropeptide-based drug design.

In addition to the superior performance and interpretability, NeuroPred-PLM provides a PyPi package, which can provide great scalability for advanced bioinformaticians. To facilitate the study of neuropeptides, we also provide a simple web service, which greatly reduces the threshold for using NeuroPred-PLM. A tutorial on how to use NeuroPred-PLM can be found in the colab notebook. In terms of accessibility and scalability, NeuroPred-PLM is currently the most comprehensive tool. In the future, we intend to improve the performance of neuropeptide prediction by developing new deep-learning methods.

Key Points

- NeuroPred-PLM uses a deep learning algorithm based on protein language model to accurately predict neuropeptides.
- Independent test experiments show that NeuroPred-PLM achieves significantly better performance than existing machine learning-based methods.
- NeuroPred-PLM adopts the global multi-head attention mechanism, which can well explain the contribution of amino acid embedding features at different positions to neuropeptide prediction via attention scores.
- We further provide an easy-to-install software package and a web server.

DATA AVAILABILITY

The data and code underlying this article are available for download from GitHub at <https://github.com/ISYSLAB-HUST/NeuroPred-PLM>.

SUPPLEMENTARY DATA

Supplementary data are available online at <https://academic.oup.com/bib>.

ACKNOWLEDGMENTS

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

FUNDING

National Natural Science Foundation of China (61772217 and 62172172), Scientific Research Start-up Foundation of Binzhou Medical University (BY2020KYQD01).

References

1. Mendel HC, Kaas Q, Muttenthaler M. Neuropeptide signalling systems - an underexplored target for venom drug discovery. *Biochem Pharmacol* 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 (Oxford)* 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 (Lausanne)* 2014;**5**:58.
6. Shahjahan M, Kitahashi T, Parhar IS. Central pathways integrating metabolism and reproduction in teleosts. *Front Endocrinol (Lausanne)* 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. Boonen K, Landuyt B, Baggerman G, et al. Peptidomics: the integrated approach of MS, hyphenated techniques and bioinformatics for neuropeptide analysis. *J Sep Sci* 2008;**31**:427–45.
11. Secher A, Kelstrup CD, Conde-Frieboes KW, et al. Analytic framework for peptidomics applied to large-scale neuropeptide identification. *Nat Commun* 2016;**7**:11436.
12. Fricker LD, Lim J, Pan H, et al. Peptidomics: identification and quantification of endogenous peptides in neuroendocrine tissues. *Mass Spectrom Rev* 2006;**25**:327–44.
13. Agrawal P, Kumar S, Singh A, et al. NeuroPIpred: a tool to predict, design and scan insect neuropeptides. *Sci Rep* 2019;**9**:5129.
14. 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**:3732–40.
15. Hasan MM, Alam MA, Shoombuatong W, et al. NeuroPred-FRL: an interpretable prediction model for identifying neuropeptide using feature representation learning. *Brief Bioinform* 2021;**22**:bbab167.
16. 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**:bbab310.
17. Corbière A, Vaudry H, Chan P, et al. Strategies for the identification of bioactive neuropeptides in vertebrates. *Front Neurosci* 2019;**13**:948.
18. Nathoo AN, Moeller RA, Westlund BA, et al. Identification of neuropeptide-like protein gene families in *Caenorhabditis elegans* and other species. *Proc Natl Acad Sci* 2001;**98**:14000–5.
19. Shi Q, Chen W, Huang S, et al. Deep learning for mining protein data. *Brief Bioinform* 2019;**22**:194–18.
20. He Y, Shen Z, Zhang Q, et al. A survey on deep learning in DNA/RNA motif mining. *Brief Bioinform* 2021;**22**:22.
21. Xu J, Li F, Leier A, et al. Comprehensive assessment of machine learning-based methods for predicting antimicrobial peptides. *Brief Bioinform* 2021;**22**:bbab083.
22. Yan J, Bhadra P, Li A, et al. Deep-AmPEP30: improve short antimicrobial peptides prediction with deep learning. *Mol Ther Nucleic Acids* 2020;**20**:882–94.

23. 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 U S A* 2021;**118**:e2016239118.
24. Elnaggar A, Heinzinger M, Dallago C, et al. ProtTrans: towards cracking the language of life's code through self-supervised deep learning and high performance computing. *IEEE Trans Pattern Anal Mach Intell* 2021;**44**:7112–27.
25. Hoie MH, Kiehl EN, Petersen B, et al. NetSurfP-3.0: accurate and fast prediction of protein structural features by protein language models and deep learning. *Nucleic Acids Res* 2022;**50**:W510–5.
26. Thummuluri V, Almagro Armenteros JJ, Johansen AR, et al. DeepLoc 2.0: multi-label subcellular localization prediction using protein language models. *Nucleic Acids Res* 2022;**50**:W228–34.
27. Wang L, Zhong H, Xue Z, et al. Res-Dom: predicting protein domain boundary from sequence using deep residual network and bi-LSTM. *Bioinformatics Advances* 2022;**2**:vbac060.
28. Talukder A, Barham C, Li X, et al. Interpretation of deep learning in genomics and epigenomics. *Brief Bioinform* 2021;**22**:22.
29. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Long Beach, California, USA: Curran Associates Inc., 2017;4768–77.
30. Fu L, Niu B, Zhu Z, et al. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012;**28**:3150–2.
31. UniProt C. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* 2021;**49**:D480–9.
32. Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res* 2014;**15**:1929–58.
33. Müller R, Kornblith S, Hinton GE. When does label smoothing help? *Adv Neural Inf Process Syst* 2019;**32**:4694–4703.
34. Kingma DP, Ba J. Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.2014*;2014;**6980**.
35. Homberg U, Würden S, Dirksen H, et al. Comparative anatomy of pigment-dispersing hormone-immunoreactive neurons in the brain of orthopteroid insects. *Cell Tissue Res* 1991;**266**:343–57.
36. Helfrich-Forster C, Homberg U. Pigment-dispersing hormone-immunoreactive neurons in the nervous system of wild-type *Drosophila melanogaster* and of several mutants with altered circadian rhythmicity. *J Comp Neurol* 1993;**337**:177–90.
37. Rao KR, Riehm JP. Pigment-dispersing hormones: a novel family of neuropeptides from arthropods. *Peptides* 1988;**9** Suppl 1: 153–9.
38. Tatemoto K, Siimesmaa S, Jörnvall H, et al. Isolation and characterization of neuropeptide Y from porcine intestine. *FEBS Lett* 1985;**179**:181–4.
39. Tatemoto K. Isolation and characterization of peptide YY (PYY), a candidate gut hormone that inhibits pancreatic exocrine secretion. *Proc Natl Acad Sci* 1982;**79**:2514–8.
40. Adrian T, Ferri GL, Bacarese-Hamilton A, et al. Human distribution and release of a putative new gut hormone, peptide YY. *Gastroenterology* 1985;**89**:1070–7.
41. Vona-Davis L, McFadden D. NPY family of hormones: clinical relevance and potential use in gastrointestinal disease. *Curr Top Med Chem* 2007;**7**:1710–20.
42. Stanley BG, Kyrkouli SE, Lampert S, et al. Neuropeptide Y chronically injected into the hypothalamus: a powerful neurochemical inducer of hyperphagia and obesity. *Peptides* 1986;**7**:1189–92.
43. Heilig M. The NPY system in stress, anxiety and depression. *Neuropeptides* 2004;**38**:213–24.
44. Thiele TE, Marsh DJ, Bernstein IL, et al. Ethanol consumption and resistance are inversely related to neuropeptide Y levels. *Nature* 1998;**396**:366–9.
45. Pedragosa-Badia X, Stichel J, Beck-Sickinger AG. Neuropeptide Y receptors: how to get subtype selectivity. *Front Endocrinol* 2013;**4**:5.
46. Eckard CP, Cabrele C, Wieland HA, et al. Characterisation of neuropeptide Y receptor subtypes by synthetic NPY analogues and by anti-receptor antibodies. *Molecules* 2001;**6**:448–67.
47. Cabrele C, Beck-Sickinger AG. Molecular characterization of the ligand–receptor interaction of the neuropeptide Y family. *J Pep Sci* 2000;**6**:97–122.