Supporting Information:

Prediction of Neuropeptides from Sequence Information Using Ensemble Classifier and Hybrid Features

Yannan Bin[1,2], Wei Zhang[1], Wending Tang[1], Ruyu Dai[1], Menglu Li[2], Qizhi Zhu[1], Junfeng Xia[1,2]*

[1]Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, Institutes of Physical Science and Information Technology, Anhui University, Hefei, Anhui 230601, China

[2]School of Computer Science and Technology, Anhui University, Hefei, Anhui 230601, China

E-mail: jfxia@ahu.edu.cn.

Contents

1 Supporting Dataset

In order to compare the performances of PredNeuroP and NeuroPIpred on predicting the insect NPs, we used another validation dataset (named NeuroPIpred_VD) downloaded from NeuroPIpred (only insect NPs, https://webs.iiitd.edu.in/raghava/neuropipred/download.php).[1] There are 175 positive insect NPs and 175 negative non-insect NPs in the validation dataset NeuroPIpred_VD.

References

1. Agrawal, P., et al., NeuroPIpred: a tool to predict, design and scan insect neuropeptides. *Scientific Reports*, **2019**, *9*, 5129.

2 Supporting Figures

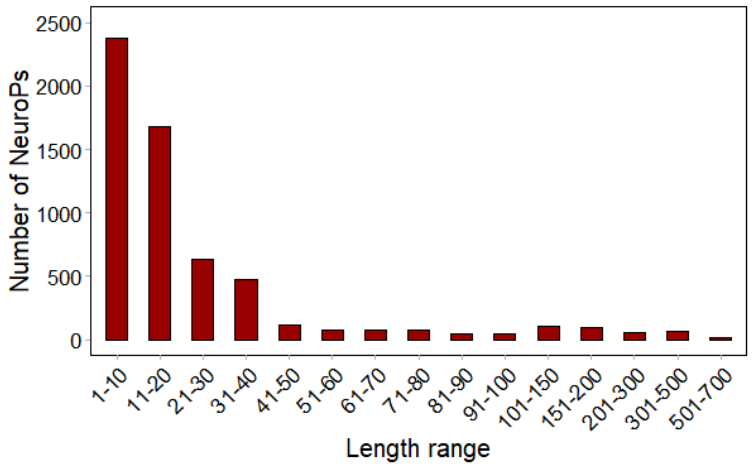

Figure S1: Length distribution of 5,948 experimentally validated NPs in NeuroPep database.

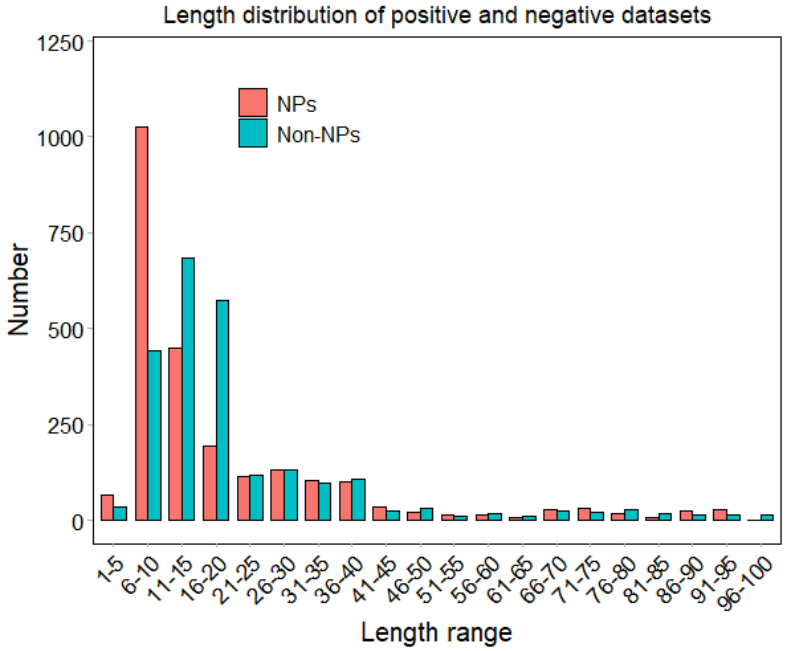

Figure S2: Length distribution of sequences in positive and negative datasets (2,425 NPs and non-NPs, respectively).

NP sequence

S	T	R	V	M	A	H	L	P	L	R	L

Binary profile

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	T
S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
V	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
M	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
A	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
H	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
L	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
L	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
L	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0

Figure S3 Generation of binary profile

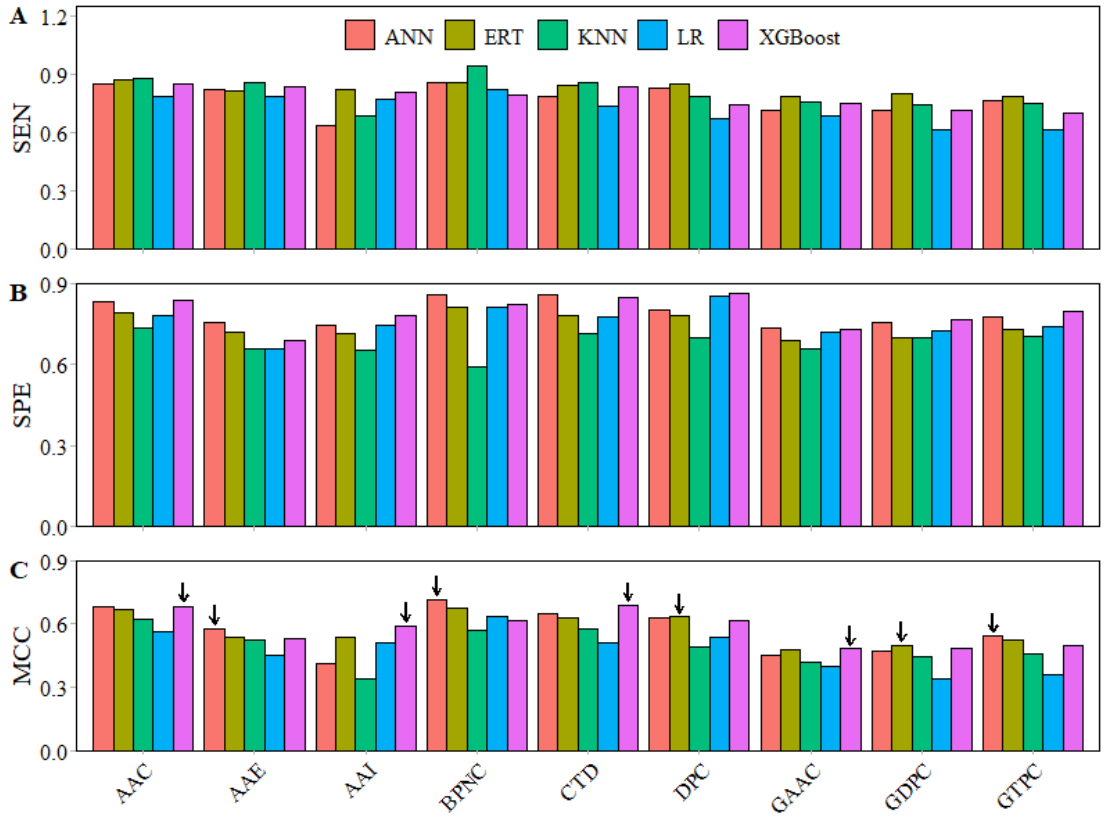

Figure S4: Predictive performance comparison of different ML models with nine feature groups on training dataset with 10-fold cross-validation. The arrow represents the highest-MCC of ML model in the feature group.

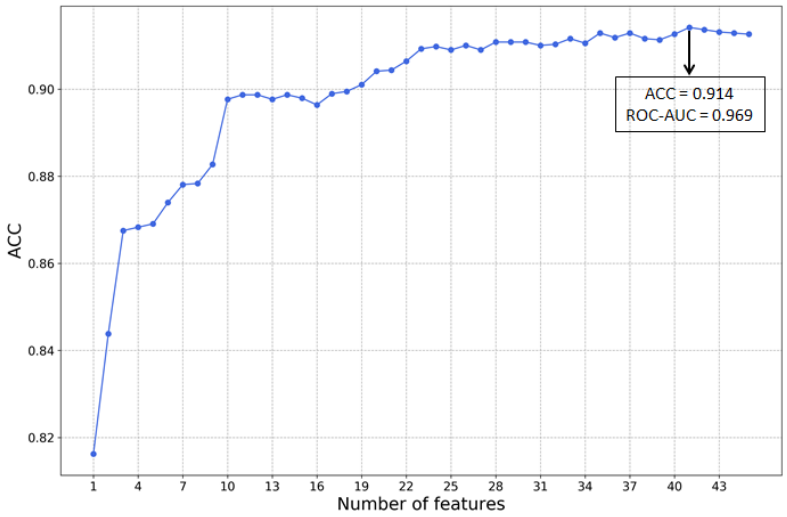

Figure S5: The performances of base-learners using mRMR and SFS in term of ACC.

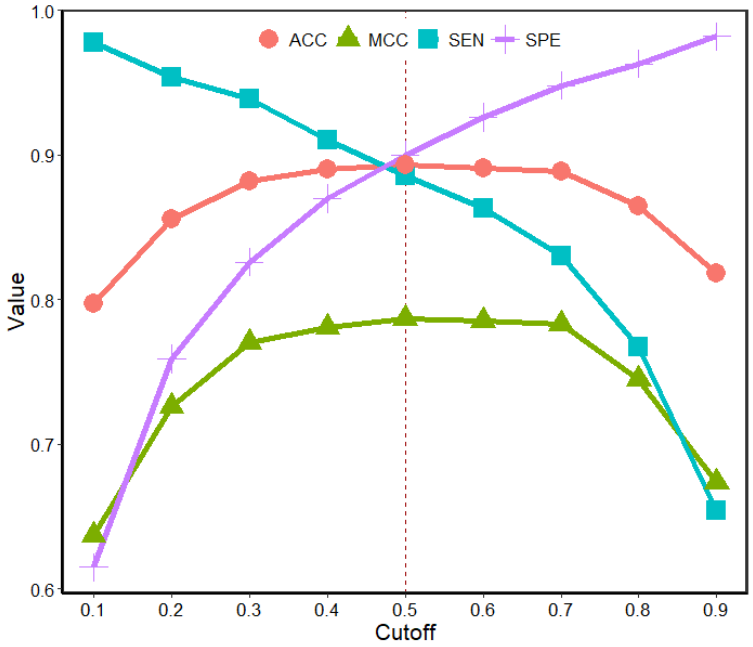

Figure S6: The performances of PredNeuroP with different cutoffs on the training dataset.

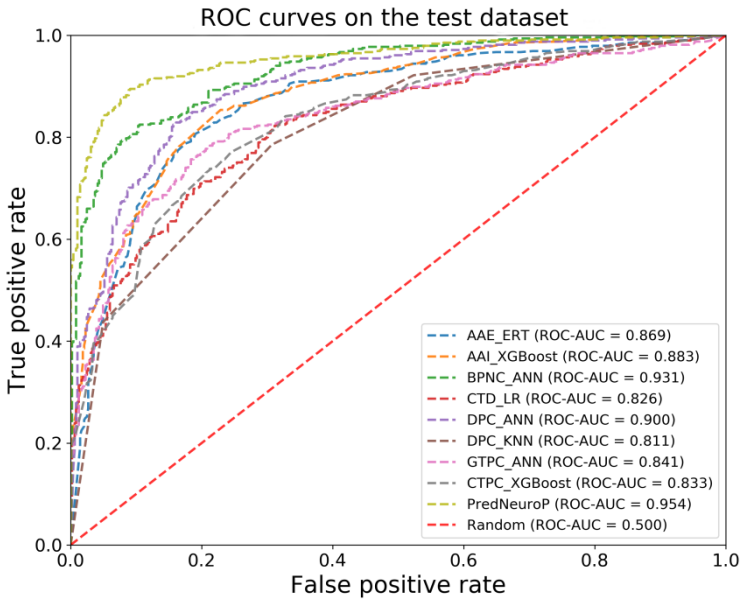

Figure S7: The performances of PredNeuroP and the selected base-learners on the test dataset.

3 Supporting Tables

Table S1: Overview of the nine feature groups

Feature category	Feature name	Abbreviation	Dimension
Composition-based feature	Amino acid composition (full, NT5 and CT5 sequences*)	AAC	60
	Dipeptide composition	DPC	400
Binary profile-based feature	Binary profiling feature (NT5 and CT5 sequences)	BPNC	200
Physicochemical property-based feature	Amino acid index (full, NT5 and CT5 sequences)	AAI	36
	Grouped amino acid composition of (full, NT5 and CT5 sequences)	GAAC	15
	Grouped dipeptide composition	GDPC	25
	Grouped tripeptide composition	GTPC	125
	Composition-transition-distribution	CTD	147
Position-based feature	Amino acid entropy (full, NT5 and CT5 sequences)	AAE	60

* NT5, the first 5 residues from N-terminus in peptide sequence; CT5, the last 5 residues from C-terminus in peptide sequence.

Table S2: Details of 12 AAindex properties in AAI features

AAindex ID	Property	Derivation
EISD840101	Hydrophobicity	For the property, $C = \frac{1}{N}\sum_{n=1}^{N} AI_n$ C is the average property (hydrophobicity, hydrophilicity, and so forth) of a given peptide, N is the length of the peptide, AI is the Index value for the property (hydrophobicity, hydrophilicity, and so forth) for the n^{th} amino acid in the peptide.
HOPT810101	Hydrophilicity	
CHAM810101	Steric parameter	
EISD860101	Solvation	
KYTJ820101	Hydropathy	
MITS020101	Amphiphilicity	
DAWD720101	Size	
GRAR740102	Polarity	
BIGC670101	Residue volume	
FAUJ880109	Hydrogen	For the property, $C = \sum_{n=1}^{N} AI_n$, C is the sum of Index values of all residues in the given peptide.
KLEP840101	Net charge	
FASG760101	Weight	

Table S3: Details of the five amino acid categories in features of GAAC, GDPC and GTPC

Physicochemical property	Amino acid
Aliphatic group	G, A, V, L, M, I
Aromatic group	F, Y, W
Positive charge group	K, R, H
Negative charged group	D, E
Uncharged group	S, T, C, P, N, Q

Tables S4: Details of the three amino acid categories in different physicochemical attributes for CTD feature

Physicochemical property	Group 1	Group 2	Group 3
Hydrophobicity	R, K, E, D, Q, N	G, A, S, T, P, H, T	C, V, I, L, M, F, W
Normalized Van der Waals volume	G, A, S, C, T, P, D	N, V, E, Q, I, L	M, H, K, F, R, Y, W
Polarity	L, I, F, W, C, M, V, Y	P, A, I, G, S	H, Q, R, K, N, E, D
Polarizability	G, A, S, D, T	C, P, N, V, E, Q, I, L	K, M, H, F, R, Y, W
Charge	K, R	A, N, C, Q, G, H, I, L, M, F, P, S, T, W, Y, V	D, E
Secondary structures	E, A, L, M, Q, K, R, H	V, I, Y, C, W, F, T	G, N, P, S, D
Solvent accessibility	A, L, F, C, G, I, V, W	R, K, Q, E, N, D	M, P, S, T, H, Y

Table S5: Motifs in full, NT5 and CT5 sequences

No.	NT5 sequence (Frequency)	CT5 sequence (Frequency)	Full sequence (Frequency)
1	HAD (0.91%)	MRF (3.96%)	**FLRF** (3.63%)
2	DEI (0.78%)	**FLRF** (3.63%)	**FMRF** (2.52%)
3	HADG (0.78%)	SFG (2.60%)	**NFLRF** (2.23%)
4	VNF (0.78%)	**FMRF** (2.52%)	**YSFG** (1.94%)
5	YGGF (0.78%)	IRF (2.35%)	YSFGL (1.89%)
6	CNT (0.66%)	NFLR (2.27%)	DFM (1.86%)
7	QHW (0.66%)	**NFLRF** (2.10%)	DFMR (1.77%)
8	QHWS (0.66%)	SFGL (2.10%)	DFMRF (1.77%)
9	SDP (0.66%)	YSF (1.98%)	FGPR (1.53%)
10	WSY (0.66%)	**YSFG** (1.94%)	AFGL (1.44%)

The common motifs in full and CT5 sequences are highlighted in bold.

Table S6: The performance comparison of PredNeuroP and NeuroPIpred on the validation dataset NeuroPIpred_VD

Tool	SEN	SPE	ACC	MCC	ROC-AUC
PredNeuroP	0.549	0.834	0.696	0.401	0.763
NeuroPIpred	0.823	0.851	0.837	0.670	0.910