BBPpred: Sequence-Based Prediction of Blood-Brain Barrier Peptides with Feature Representation Learning and Logistic Regression

Ruyu Dai[1], Wei Zhang[1], Wending Tang[1], Evelien Wynendaele[2], Qizhi Zhu[1], Yannan Bin[1], Bart De Spiegeleer[2], Junfeng Xia[1,]*

[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]Drug Quality and Registration (DruQuaR) Group, Faculty of Pharmaceutical Sciences, Ghent University, Harelbekestraat 72, 9000 Ghent, Belgium

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

Contents

1 **Supporting Figures** S-3

Figure S1. Prediction performance of different learning strategies. (A) ROC curves on the training dataset. (B) PR curves on the training dataset. (C) ROC curves on the test dataset. (D) PR curves on the test dataset.

2 **Supporting Tables**

S-4~7

Table S1: List of feature descriptors and their dimensions.

S-4~5

Table S2: List of feature importance scores measured by F-score

S-5~6

Table S3: Performances assessment of feature subsets based on SFS measured by AUC

S-6~7

Table S4: Performance assessment of the optimal feature representation (OF7) with different classifiers

S-7

Table S5: Performance assessment of the optimal feature representation (OF7) and the original seven feature representation on the training dataset.

S-7

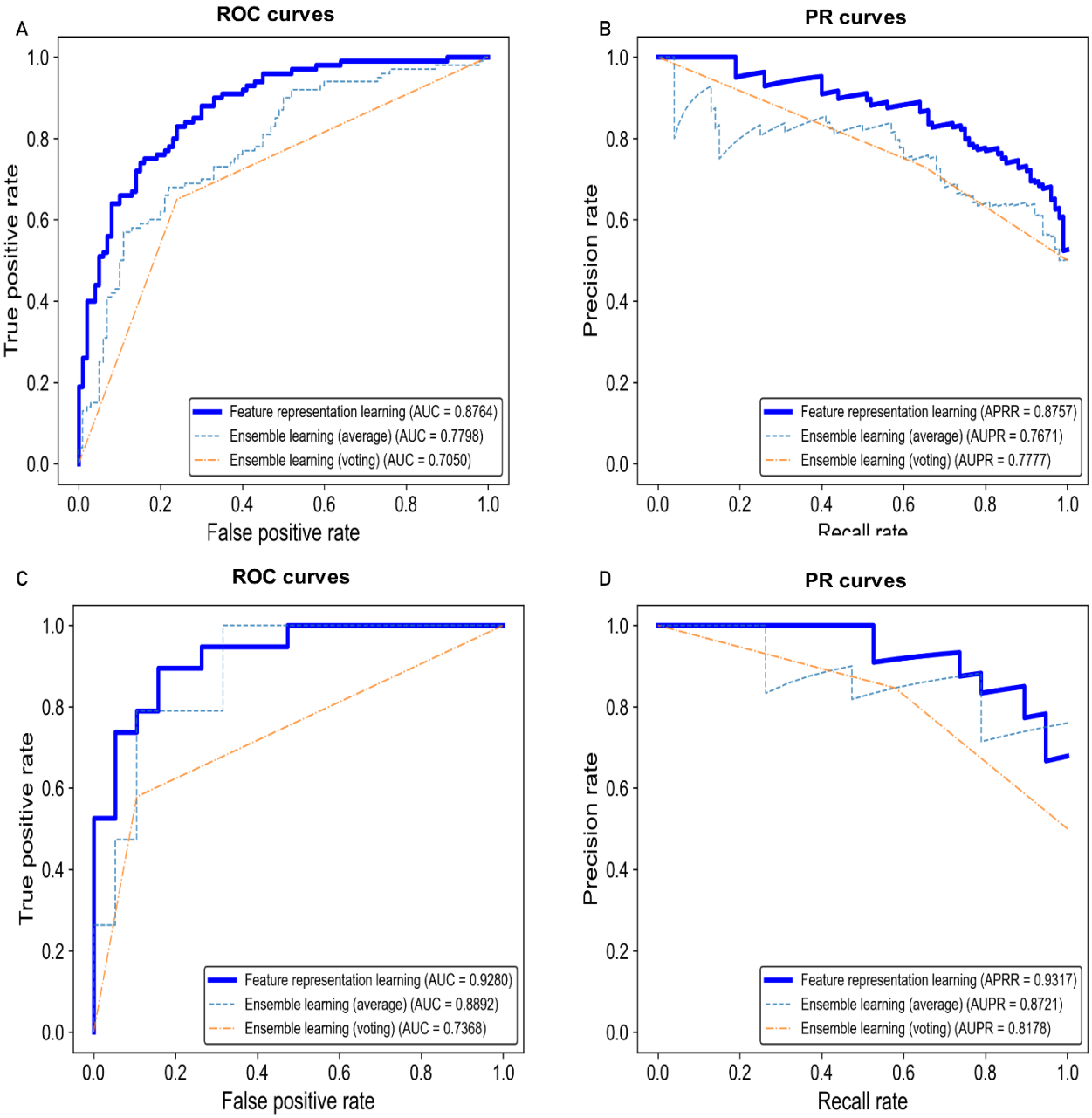

Figure S1. Prediction performance of different learning strategies. (A) ROC curves on the training dataset. (B) PR curves on the training dataset. (C) ROC curves on the test dataset. (D) PR curves on the test dataset.

Table S1. List of feature descriptors and their dimensions.

Feature descriptor	Feature type	Dimension	Feature descriptor	Feature type	Dimension
1	AAC (N=1)	20	53	DPC (CT=1)	400
2	AAC (CT=1)	20	54	DPC (CT=2)	400
3	AAC (CT=2)	20	55	DPC (CT=3)	400
4	AAC (CT=3)	20	56	DPC (CT=4)	400
5	AAC (CT=4)	20	57	DPC (CT=5)	400
6	AAC (CT=5)	20	58	DPC (NT=1)	400
7	AAC (NT=1)	20	59	DPC (NT=2)	400
8	AAC (NT=2)	20	60	DPC (NT=3)	400
9	AAC (NT=3)	20	61	DPC (NT=4)	400
10	AAC (NT=4)	20	62	DPC (NT=5)	400
11	AAC (NT=5)	20	63	DPC (NTCT=1)	400
12	AAC (NTCT=1)	20	64	DPC (NTCT=2)	400
13	AAC (NTCT=2)	20	65	DPC (NTCT=3)	400
14	AAC (NTCT=3)	20	66	DPC (NTCT=4)	400
15	AAC (NTCT=4)	20	67	DPC (NTCT=5)	400

16	AAC (NTCT=5)	20	68	GDPC	25
17	AAC (N=2)	20	69	GGAP (G=1)	400
18	ASDC	400	70	GGAP (G=2)	400
19	BIT12	12	71	GGAP (G=3)	400
20	BIT188	188	72	GGAP (G=4)	400
21	BIT20 (CT=1)	20	73	GTPC	125
22	BIT20 (CT=2)	40	74	IT (CT=2)	3
23	BIT20 (CT=3)	60	75	IT (CT=3)	3
24	BIT20 (CT=4)	80	76	IT (CT=4)	3
25	BIT20 (CT=5)	100	77	IT (CT=5)	3
26	BIT20 (NT=1)	20	78	IT (NT=2)	3
27	BIT20 (NT=2)	40	79	IT (NT=3)	3
28	BIT20 (NT=3)	60	80	IT (NT=4)	3
29	BIT20 (NT=4)	80	81	IT (NT=5)	3
30	BIT20 (NT=5)	100	82	IT (NTCT=1)	3
31	BIT20 (NTCT=1)	40	83	IT (NTCT=2)	3
32	BIT20 (NTCT=2)	80	84	IT (NTCT=3)	3
33	BIT20 (NTCT=3)	120	85	IT (NTCT=4)	3
34	BIT20 (NTCT=4)	160	86	IT (NTCT=5)	3

35	BIT20 (NTCT=5)	200	87	OLP (CT=1)	10
36	BIT21 (CT=1)	21	88	OLP (CT=2)	20
37	BIT21 (CT=2)	42	89	OLP (CT=3)	30
38	BIT21 (CT=3)	63	90	OLP (CT=4)	40
39	BIT21 (CT=4)	84	91	OLP (CT=5)	50
40	BIT21 (CT=5)	105	92	OLP (NT=1)	10
41	BIT21 (NT=1)	21	93	OLP (NT=2)	20
42	BIT21 (NT=2)	42	94	OLP (NT=3)	30
43	BIT21 (NT=3)	63	95	OLP (NT=4)	40
44	BIT21 (NT=4)	84	96	OLP (NT=5)	50
45	BIT21 (NT=5)	105	97	OLP (NTCT=1)	20
46	BIT21 (NTCT=1)	42	98	OLP (NTCT=2)	40
47	BIT21 (NTCT=2)	84	99	OLP (NTCT=3)	60
48	BIT21 (NTCT=3)	126	100	OLP (NTCT=4)	80
49	BIT21 (NTCT=4)	168	101	OLP (NTCT=5)	100
50	BIT21 (NTCT=5)	210	102	PAAC	24
51	CTD	63	103	QSO	48
52	CTF	343			

* Parameters NT, CT, and NTCT mean that we approach to extract subsequences with a certain length (range from 1 to 5) from the N-, C-terminus or both N- and C-terminus of a given peptide. In AAC feature descriptor, N = 1, 2 denotes the method types: 1 denotes the frequency and 2 denotes the occurrence number, and we extract N-, C-terminus or both N- and C-terminus subsequences (range from 1 to 5) when N=1. Parameter G is the distance between two amino acids considered.

Table S2. List of feature importance scores measured by F-score.

Rank	Feature descriptor	Score	Rank	Feature descriptor	Score
1	BIT20 (NT=2)	85.1371	53	CTF	28.7047
2	BIT12	80.6451	54	AAC (NTCT=1)	27.9036
3	AAC (NT=1)	77.7422	55	BIT21 (NTCT=1)	27.5486
4	DPC (NT=1)	77.6948	56	GGAP (G=1)	26.6082
5	BIT20 (NT=1)	77.5210	57	GGAP (G=3)	26.2747
6	BIT21 (NT=2)	71.8159	58	DPC (NTCT=1)	25.5273
7	OLP (NT=1)	69.0949	59	OLP (CT=5)	21.2451
8	AAC (NTCT=5)	67.9733	60	DPC (CT=4)	21.1493
9	DPC (NT=2)	67.7131	61	DPC (NTCT=3)	19.9785
10	BIT20 (NTCT=2)	66.6645	62	BIT20 (CT=5)	16.3670
11	BIT21 (NT=1)	64.1939	63	AAC (NT=3)51	16.2798
12	BIT21 (NT=4)	63.3164	64	DPC (NTCT=4)	13.1745
13	QSO	63.0428	65	OLP (CT=4)	12.2778
14	BIT21 (NT=3)	61.4583	66	DPC (NT=3)	11.9029
15	AAC (N=2)	60.9271	67	DPC (NTCT=5)	10.7736
16	PAAC	60.1124	68	DPC (CT=2)	10.5404
17	AAC (N=1)	57.6307	69	BIT20 (CT=2)	9.1259
18	CTD	57.5600	70	BIT21 (CT=5)	8.9847
19	BIT21 (NTCT=2)	56.3773	71	BIT20 (CT=4)	8.8244
20	BIT20 (NT=5)	56.1641	72	DPC (CT=3)	8.4833
21	OLP (NTCT=2)	55.9420	73	IT (CT=5)	8.3169

22	BIT21 (NT=5)	55.5526	74	DPC (NT=4)	7.0705
23	BIT20 (NT=4)	55.2334	75	DPC (CT=5)	6.3156
24	BIT20 (NTCT=5)	55.1754	76	OLP (CT=3)	6.1971
25	OLP (NTCT=5)	54.6377	77	OLP (CT=2)	5.8301
26	OLP (NT=5)	53.9839	78	BIT21 (CT=4)	5.6605
27	BIT20 (NT=3)	53.1029	79	AAC (CT=5)	5.4716
28	OLP (NT=3)	51.9596	80	BIT21 (CT=2	5.4324
29	AAC (NTCT=3)	51.8422	81	AAC (CT=4)	5.3418
30	BIT20 (NTCT=3)	51.8142	82	AAC (CT=2)	4.1429
31	OLP (NTCT=1)	51.7999	83	BIT20 (CT=3)	4.0855
32	OLP (NT=2)	51.6765	84	BIT21 (CT=3)	3.9771
33	OLP (NT=4)	51.2905	85	IT (NTCT=2)	3.9367
34	OLP (NTCT=4)	49.8852	86	DPC (NT=5)	3.5554
35	BIT20 (NTCT=4)	49.5620	87	IT (NTCT=5)	2.9152
36	OLP (NTCT=3)	49.3089	88	BIT21 (CT=1)	2.9057
37	BIT21 (NTCT=5)	49.2307	89	AAC (CT=3)	2.5862
38	BIT21 (NTCT=3)	48.7653	90	IT (CT=2)	2.5855
39	BIT188	48.0729	91	IT (NTCT=4)	2.5161
40	BIT20 (NTCT=1)	47.4533	92	IT (NT=2)	1.8547
41	BIT21 (NTCT=4)	47.0786	93	IT (NT=5)	1.4000
42	GGAP (G=4)	46.4878	94	IT (NT=4)	1.2126

43	AAC (NTCT=4)	45.3240	95	IT (NT=3)	0.7370
44	AAC (NT=2)	42.3880	96	IT (NTCT=3)	0.2223
45	ASDC	40.5044	97	DPC (CT=1)	0.1987
46	GTPC	38.7684	98	BIT20 (CT=1)	0.1987
47	GGAP (G=2)	38.4875	99	AAC (CT=1)	0.1987
48	GDPC	38.0546	100	IT (CT=3)	0.1649
49	AAC (NT=5)	34.0916	101	IT (CT=4)	0.0764
50	AAC (NTCT=2)	33.0712	102	OLP (CT=1)	0.0356
51	DPC (NTCT=2)	30.4244	103	IT (NTCT=1)	0.0058
52	AAC (NT=4)	28.7979			

Table S3. Performances assessment of feature subsets based on SFS measured by AUC.

Rank	Feature descriptor	AUC	Rank	Feature descriptor	AUC
1	BIT20 (NT=2)	0.8078	20	OLP (CT=2)	0.8700
2	BIT12	0.7645	21	AAC (CT=5)	0.8640
3	AAC (NTCT=5)	0.8528	22	BIT21 (CT=2)	0.8608
4	BIT21 (NT=4)	0.8420	23	AAC (CT=4)	0.8588
5	AAC (NTCT=3)	0.8488	24	AAC (CT=2)	0.8650
6	OLP (NTCT=1)	0.8628	25	IT (NTCT=2)	0.8533
7	AAC (NT=5)	0.8710	26	IT (NTCT=5)	0.8623
8	AAC (NTCT=2)	0.8545	27	BIT21 (CT=1)	0.8528
9	DPC (NTCT=2)	0.8563	28	AAC (CT=3)	0.8618
10	GGAP (G=2)	0.8430	29	IT (CT=2)	0.8648
11	OLP (CT=5)	0.8568	30	IT (NTCT=4)	0.8595
12	DPC (CT=4)	0.8535	31	IT (NT=2)	0.8590

13	DPC (NTCT=3)	0.8550	32	IT (NT=5)	0.8548
14	AAC (NT=3)	0.8620	33	IT (NT=4)	0.8558
15	DPC (NTCT=4)	0.8595	34	IT (NT=3)	0.8610
16	DPC (NT=3)	0.8630	35	IT (NTCT=3)	0.8613
17	DPC (NTCT=5)	0.8618	36	IT (CT=3)	0.8560
18	DPC (CT=2)	0.8588	37	IT (CT=4)	0.8525
19	IT (CT=5)	0.8553	38	IT (NTCT=1)	0.8598

Table S4. Performance assessment of the optimal feature representation (OF7) with different classifiers.

Classifier	**SEN**	**SPE**	**PRE**	**F1**	**MCC**	**ACC**
ERT	0.7900	0.7900	0.7900	0.7900	0.5800	0.7900
KNN	0.7400	0.8300	0.8132	0.7749	0.5723	0.7850
LR	0.7800	0.7700	0.7723	0.7761	0.5500	0.7750
MLP	0.7900	0.7700	0.7745	0.7822	0.5601	0.7800
RF	0.7800	0.7700	0.7723	0.7761	0.5500	0.7750
SVM	0.8100	0.7600	0.7714	0.7902	0.5707	0.7850
XGB	0.7500	0.7700	0.7653	0.7576	0.5201	0.7600

Table S5. Performance assessment of the optimal feature representation (OF7) and the original seven feature representation on the training dataset.

Feature descriptor	SEN	SPE	PRE	F1	MCC	ACC
OF7	0.7800	0.7700	0.7723	0.7761	0.5500	0.7750
BIT20 (NT=2)	0.7000	0.7000	0.7000	0.7000	0.4000	0.7000
BIT12	0.5800	0.7100	0.6667	0.6203	0.2925	0.6450

AAC (NTCT=5)	0.5500	0.6900	0.6395	0.5914	0.2424	0.6200
BIT21 (NT=4)	0.6600	0.6900	0.6804	0.6701	0.3502	0.6750
AAC (NTCT=3)	0.5500	0.6600	0.6180	0.5820	0.2113	0.6050
OLP (NTCT=1)	0.6700	0.7200	0.7053	0.6872	0.3905	0.6950
AAC (NT=5)	0.6200	0.6500	0.6392	0.6294	0.2701	0.6350