ORIGINAL RESEARCH
published: 17 May 2022
doi: 10.3389/fgene.2022.845747

BBPpredict: A Web Service for Identifying Blood-Brain Barrier Penetrating Peptides

Xue Chen[1], *Qianyue Zhang*[1], *Bowen Li*[1], *Chunying Lu*[1], *Shanshan Yang*[1], *Jinjin Long*[1], *Bifang He*[1]*, *Heng Chen*[1]* and *Jian Huang*[2]*

[1]*Medical College, Guizhou University, Guiyang, China,* [2]*School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, China*

Blood-brain barrier (BBB) is a major barrier to drug delivery into the brain in the treatment of central nervous system (CNS) diseases. Blood-brain barrier penetrating peptides (BBPs), a class of peptides that can cross BBB through various mechanisms without damaging BBB, are effective drug candidates for CNS diseases. However, identification of BBPs by experimental methods is time-consuming and laborious. To discover more BBPs as drugs for CNS disease, it is urgent to develop computational methods that can quickly and accurately identify BBPs and non-BBPs. In the present study, we created a training dataset that consists of 326 BBPs derived from previous databases and published manuscripts and 326 non-BBPs collected from UniProt, to construct a BBP predictor based on sequence information. We also constructed an independent testing dataset with 99 BBPs and 99 non-BBPs. Multiple machine learning methods were compared based on the training dataset via a nested cross-validation. The final BBP predictor was constructed based on the training dataset and the results showed that random forest (RF) method outperformed other classification algorithms on the training and independent testing dataset. Compared with previous BBP prediction tools, the RF-based predictor, named BBPpredict, performs considerably better than state-of-the-art BBP predictors. BBPpredict is expected to contribute to the discovery of novel BBPs, or at least can be a useful complement to the existing methods in this area. BBPpredict is freely available at http://i.uestc.edu.cn/BBPpredict/cgi-bin/BBPpredict.pl.

Keywords: blood-brain barrier, random forest (RF), nested cross-validation, computational method, blood-brain barrier penetrating peptides (BBPs)

OPEN ACCESS

Edited by:
Chuan Dong,
Wuhan University, China

Reviewed by:
Leyi Wei,
Shandong University, China
Zunnan Huang,
Guangdong Medical University, China

Correspondence:
Bifang He
bfhe@gzu.edu.cn
Heng Chen
hchen13@gzu.edu.cn
Jian Huang
hj@uestc.edu.cn

Specialty section:
This article was submitted to
Computational Genomics,
a section of the journal
Frontiers in Genetics

Received: *30 December 2021*
Accepted: *30 March 2022*
Published: *17 May 2022*

Citation:
Chen X, Zhang Q, Li B, Lu C, Yang S, Long J, He B, Chen H and Huang J (2022) BBPpredict: A Web Service for Identifying Blood-Brain Barrier Penetrating Peptides.
Front. Genet. 13:845747.
doi: 10.3389/fgene.2022.845747

1 INTRODUCTION

Blood-brain barrier (BBB) highly protects the central nervous system (CNS) (Nance et al., 2022), preventing 98% of small molecules and 100% of large molecules from entering the brain (Sánchez-Navarro et al., 2017). It is the main obstacle for drug delivery into the brain (Banks, 2016). Therefore, exploring methods for drugs to penetrate BBB is a research hotpot in the development of drugs for CNS disorders (Terstappen et al., 2021).

Blood-brain barrier penetrating peptides (BBPs) can cross the BBB through various mechanisms without destroying the integrity of BBB (Van Dorpe et al., 2012; Oller-Salvia et al., 2016). It has been reported that partial BBPs can transfer drugs into the brain, which provides a new avenue for the development of drugs for CNS diseases (Zhou et al., 2021). Furthermore, because of their characteristics of easy synthesis, satisfactory effect, low toxicity and wide selectivity (Muttenthaler et al., 2021), BBPs show broad application prospects as carriers or therapeutic agents for CSN diseases treatment (Zhou et al., 2021). Nonaka et al. reported that IF7, an annexin A1-binding peptide, could overcome BBB and deliver chemotherapeutics to target brain tumors (Nonaka et al., 2020). Xie and coworkers demonstrated that d-peptide ligand of angiopep-2 modified nanoprobes could cross BBB and locate glioma sites (Xie et al., 2021). Lim and collaborators found that dNP2 peptide could penetrate BBB and deliver ctCTLA-4 protein to ameliorate autoimmune encephalomyelitis in mouse models (Lim et al., 2015). Kurzrock and Drappatz et al. showed that ANG1005 or GRN1005, a conjugate of angiopep-2 and paclitaxel, has reached clinical study for the treatment of glioma (Kurzrock et al., 2012; Drappatz et al., 2013).

There have been two BBP databases published to date, Brainpeps (Van Dorpe et al., 2012) and B3Pdb (Kumar et al., 2021b), since BBPs became candidates for developing peptide agents for managing CNS disorders. These studies are undoubtedly a strong boost to the development of medications for CNS diseases. However, the discovery of BBPs by wet-lab experiment is time-consuming and complex, and only hundreds of BBPs have been identified experimentally to date. Construction of computational methods for the identification of BBPs is very valuable for developing therapeutics for CSN diseases. Machine learning methods have been successfully applied to the classification of various peptides, such as cell-penetrating peptides (Wei et al., 2017a; Wei et al., 2017b; Kumar et al., 2018), antimicrobial peptides (Bhadra et al., 2018), anticancer peptides (Li and Wang, 2016). There are also two BBP predictors, BBPpred (Dai et al., 2021) and B3Pred (Kumar et al., 2021a), have published successively for identifying BBPs. BBPpred is based on logistic regression to identify BBPs, while B3Pred uses random forest (RF) to predict BBPs. Considering the low sample complexity of these two classifiers, the performance of computational models for identifying BBPs can be improved.

In this work, we collected more BBPs from existing databases (Van Dorpe et al., 2012; Kumar et al., 2021b) and published literatures to construct a new BBP predictor named BBPpredict, which is an online web service and freely available at http://i.uestc.edu.cn/BBPpredict/cgi-bin/BBPpredict.pl. By comparing the results of the nested five-fold cross-validation and independent testing dataset of various machine learning predictors, the RF-based model showed the best prediction performance. Thus, BBPpredict was implemented by using RF. We expect BBPpredict will help researchers find more novel BBPs.

2 MATERIALS AND METHODS

2.1 Datasets

In this work, we selected experimentally validated BBPs as candidate positive samples that were collected from Brainpeps (Van Dorpe et al., 2012), B3Pdb(Kumar et al., 2021b), public datasets of BBPpred (Dai et al., 2021) and B3Pred (Kumar et al., 2021a), and other published literatures from PubMed with query "(((Brain [Title/Abstract]) OR (blood–brain barrier [Title/Abstract])) AND peptide [Title/Abstract]) AND (transport [Title/Abstract] OR transfer [Title/Abstract] OR permeation [Title/Abstract] OR permeability [Title/Abstract])", covering the period 2011–2021. BBPs were then preprocessed as follows: 1) the repetitive sequences were eliminated; 2) peptide sequences with ambiguous residues ("X", "B" and "Z", etc.) were deleted (He et al., 2016). Finally, 425 BBPs were remained as positive samples. We also collected 1,304 non-BBPs that were obtained by the following three steps: 1) collect initial sequences from UniProt with the query "peptides length: [5 TO 50] NOT blood brain barrier NOT brain NOT brainpeps NOT b3pdb NOT permeation NOT permeability NOT venom NOT toxin NOT transmembrane NOT transport NOT transfer NOT membrane NOT neuro NOT hemolysis AND reviewed: yes" (Dai et al., 2021), 2) remove redundant sequences by using CD-HIT (sequence identity cut-off of 10%) (Dai et al., 2021), 3) exclude the peptide sequences with ambiguous residues ("X", "B," and "Z", etc.).

TABLE 1 | List of training dataset and independent testing dataset.

Dataset	Number of BBPs	Number of Non-BBPs
Training dataset	326	326
Independent testing dataset	99	99

2.2 Training and Independent Testing Datasets

To evaluate the performance of our predictor and existing predictors (BBPpred and B3Pred), 99 BBPs that collected through published literatures and 99 non-BBPs randomly selected from candidate negative samples construct an independent testing dataset that was completely independent of the training dataset of the three predictor models (BBPpred, B3Pred and our proposed BBPpredict) (**Table 1**). The remaining 326 BBPs were used as the positive training dataset. To balance the sample size for training, we randomly selected 326 non-BBPs as the negative training dataset (**Table 1**), whose length distribution is the same as the positive training dataset. All datasets are available for download from http://i.uestc.edu.cn/BBPpredict/download.html.

2.3 Feature Extraction

Feature extraction refers to the transformation of peptide sequences into fixed-length feature vectors, which is an indispensable step for the construction of predictors. In this study, we selected five feature encoding methods, including amino acid composition (AAC), dipeptide composition (DPC), composition of k-spaced amino acid group pairs (CKSAAGP, k = 3), pseudo-amino acid composition (PAAC) and grouped amino acid composition (GAAC) to extract the characteristics of peptide sequence. Here we set the length of a peptide to be N, and all feature extraction methods are based on 20 natural amino acids (i.e., "ACDEFGHIKLMNPQRSTVWY"). Feature extraction was implemented by an in-house script.

2.3.1 Amino Acid Composition

AAC calculates the frequency of each amino acid in the peptide sequence (Bhasin and Raghava, 2004). It can be calculated as:

$$f(i) = \frac{N(i)}{N}, i \in \{A, C, D, ...Y\} \tag{1}$$

where *N(i)* is the number of the amino acid type *i*.

2.3.2 Dipeptide Composition

DPC gives 400 descriptors (i.e."$AA, AC, AD, \ldots YY$ ") (Saravanan and Gautham, 2015). It is defined as:

$$D(r, s) = \frac{N_{rs}}{N-1}, r, s \in \{A, C, D, ...Y\} \tag{2}$$

where *Nrs* is the number of the dipeptide consisting of amino acids *r* and *s* in the peptide sequence.

2.3.3 Grouped Amino Acid Composition

For the GAAC encoding, 20 natural amino acids are firstly divided into five categories according to their physicochemical properties: amino acid groups g1 (GAVLMI), g2 (FYW), g3 (KRH), g4 (DE) and g5 (STCPNQ). Group g1 belongs to the aliphatic group, g2 aromatic group, g3 positive charge group, g4 negative charged group and g5 uncharged group, respectively. GAAC represents the frequency of each amino acid group (Lee et al., 2011) and can be described as:

$$\begin{aligned} f(g) &= \frac{N(g_i)}{N}, i \in \{g1, g2, g3, g4, g5\} \\ N(g_i) &= \sum N(i), i \in \{g1, g2, g3, g4, g5\} \end{aligned} \tag{3}$$

where $N(g_i)$ is the number of amino acids in group g, *N(i)* is the number of the amino acid type *i*.

2.3.4 Composition of *K*-Spaced Amino Acid Group Pairs

CKSAAGP is based on CKSAAP (Chen et al., 2007a; Chen et al., 2007b, 2008; Chen et al., 2009) descriptor and GAAC descriptor, which calculates the frequency of *k*-spaced group pairs. And the detailed calculation of CKSAAGP can refer to (Chen et al., 2018). In this study, we set k as three by default. And when k = 0, CKSAAGP can be calculated as:

$$\left(\frac{N_{g1g1}}{N_{total}}, \frac{N_{g1g2}}{N_{total}}, \frac{N_{g1g3}}{N_{total}}, \ldots \frac{N_{g1g5}}{N_{total}}\right)25 \tag{4}$$

Where N_{total} describes *N*-1, N_{gg} is the number of 0-spaced group pairs.

2.3.5 Pseudo-Amino Acid Composition

PAAC describes the information of two residues order and properties in the peptide sequence. The computation of PAAC is available in (Chou, 2001; 2005).

After feature extraction, each peptide was encoded by a 550-dimensional feature vector, which was generated by concatenating five types of feature vector.

2.4 Feature Scoring and Selection

Generally, not all features make contribution to the model construction. Partial features make remarkable contributions, while some others make slight contributions (He et al., 2019). Therefore, feature selection is a very vital step for accomplishing a classifier model with promising classification performance (Zhao et al., 2016). In this study, F-score method was employed to estimate each feature's contribution. The feature with a greater F-score implies its larger contribution for prediction model. We conducted the following procedures to select more informative features from the 550 features that were extracted from the training dataset. In the first stage, we evaluated the five-fold cross-validation performance of top 92, 184, 275, 367, 458, 550 features for various classification algorithms. In the five-fold cross-validation, the training dataset was equally divided into five subsets, among these five subsets, a subset was used as the testing-set and the other four subsets as the training-set. The division of top 92, 184, 275, 367, 458, 550 features based on the training-set was determined by making (count_max-count_min)/6 as the cut-off point of feature division, where "count_max" represents the maximum dimension of feature (550 features), and "count_min" is the minimum dimension of feature (1 feature). In the second stage, according to the five-fold cross-validation results of different classification algorithms, we obtained the number of features n with the highest accuracy. In the third stage, we selected top n features from the 550 features extracted from the training dataset and ranked by F-score in descending order to construct the final model.

2.5 Classification Model Construction

Eight traditional machine learning algorithms, including decision tree (DT), RF, k-nearest neighbors (KNN), adaptive boosting (AdaBoost), gentle adaptive boosting (GentleBoost), adaptive logistic regression (LogitBoost), linear support vector machine (linearSVM) and radial basis function (RBF) kernel SVM (rbfSVM) were used to build the predictive models based on the features selected by feature selection (see in **Supplementary Table S3**), respectively. LIBSVM 3.24 (http://www.csie.ntu.edu.tw/~cjlin/libsvm/) was utilized to accomplish linearSVM and rbfSVM (Chang and Lin, 2011). DT, RF, KNN, AdaBoost, GentleBoost and LogitBoost are respectively implemented by MATLAB R2021a built-in functions fitcTree, TreeBagger, fitcknn and fitcEnmbles. To compare with deep learning method, a long-short term memory (LSTM) network that realized based on Keras 2.3.1 (tensorflow 2.1.0 as backend) package of python 3.6 was also utilized to construct the classification model (Hochreiter and Schmidhuber, 1997). The LSTM classification model consisted of one LSTM layer with eight hidden neurons. The non-linear activation function hyperbolic tangent (tanh) was applied to LSTM layer. It should be noted that for LSTM, the vectored sequence of peptide was utilized as classification features and no feature

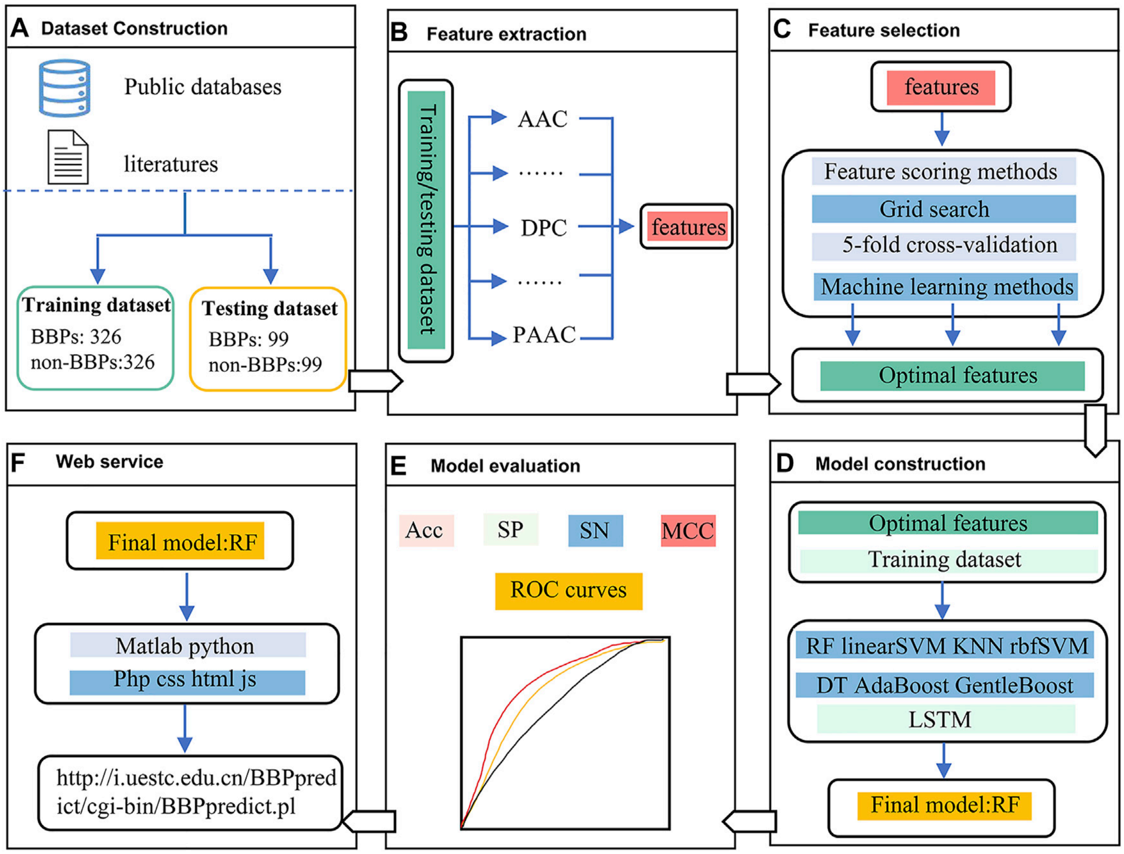


FIGURE 1 | The framework of BBPpredict. **(A)**. Dataset Construction. **(B)**. Feature extraction. **(C)**. Feature selection. **(D)**. Model construction. **(E)**. Model evaluation. **(F)**. Web service.

selection was applied. The pseudo code for final model construction can be found in the **Supplementary Material**.

2.6 Prediction Assessment

Five evaluation indexes, including accuracy (ACC), sensitivity (SN), specificity (SP), Matthews correlation coefficient (MCC) and the area under the receiver operating characteristic (ROC) curve (AUC), were utilized to quantify the performance of each predictive model. The first four indicators are calculated as follows:

$$SN = \frac{TP}{TP + FN} \tag{5}$$

$$SP = \frac{TN}{TN + FP} \tag{6}$$

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

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{8}$$

where *TP* describes the number of genuine BBPs which are predicted as BBPs. *FN* represents the number of genuine BBPs that are identified as non-BBPs. Denote *TN* as the number of true non-BBPs classified as non-BBPs and *FP* the number of true non-BBPs identified as BBPs. *SN* and *SP* primarily assess the ability of a predictive model to identify positive and negative samples respectively, while *ACC* and *MCC* investigate the comprehensive capacity of a prediction model to classify both positive and negative samples (Wang et al., 2019). The AUC score is often utilized to judge the merits and demerits of classifiers. In this study, we selected the optimal predictive model according to the AUC value. The model construction and evaluation were performed at a computational server (Sugon I840-G20, Dawning Information Industry Co., LTD., Beijing, China).

2.7 Reproducible Analysis

Data analysis reproducibility plays a vital role for achieving an independent verification of the analysis results (Walzer and Vizcaíno, 2020). In this work, we constructed 100 testing datasets and corresponding training datasets to verify the robustness of the construction method of the BBP predictor. To avoid high similarity between the independent testing dataset and the testing dataset of the reproducible analysis, here each testing dataset consisted of 50 BBPs randomly selected from candidate positive samples (114 BBPs) that are independent of the training datasets of BBPpred and B3Pred and 50 non-BBPs with the same selection rules with BBPs. The model building process based on 100 reconstructed datasets for different classification algorithms (RF, rbfSVM, linearSVM, etc.) is consistent with the above method. The

TABLE 2 | The prediction performances of different classifiers in nested five-fold cross-validation.

Scoring Method	Classifier	SN(%)	SP(%)	ACC(%)	MCC	AUC
F-score	**RF**	**79.14**	**84.66**	**81.90**	**0.6390**	**0.9030**
	KNN	76.69	80.98	78.83	0.5772	0.7883
	rbfSVM	78.83	83.13	80.98	0.6202	0.8872
	linearSVM	75.77	83.13	79.45	0.5906	0.8690
	DT	71.78	74.54	73.16	0.4634	0.7357
	LSTM	65.23	75.38	70.31	0.4083	0.7313
	AdaBoost	77.91	80.67	79.29	0.5861	0.8615
	GentleBoost	77.30	80.06	78.68	0.5738	0.8582
	LogitBoost	79.14	82.21	80.67	0.6138	0.8680

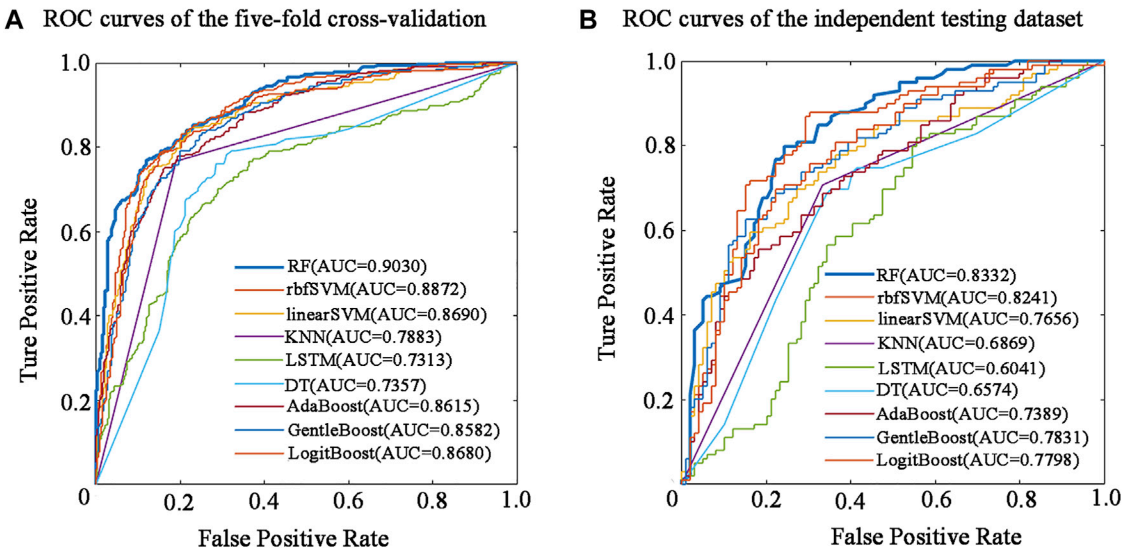


FIGURE 2 | Performance evaluation of different predictors in five-fold cross-validation and independent testing dataset. **(A)** ROC curves of the five-fold cross-validation. **(B)** ROC curves of the independent testing dataset.

result of the reproducibility analysis can be found in the **Supplementary Material**.

3 RESULT

3.1 Overall Workflow

The framework of this study is depicted in **Figure 1**. In the first stage, two benchmark datasets, including a training dataset and an independent testing dataset, were constructed. In the second stage, five feature extraction methods were utilized to encode each peptide sequence, and then a 550-dimensional feature vector was generated. In the third stage, feature scoring methods and grid search with five-fold cross-validation strategy was used for feature selection. In the fourth stage, multiple machine learning methods were employed to build different models. In the fifth stage, we evaluated the predictive performance of the nine models by using a nested five-fold cross-validation and an independent testing dataset, respectively. Finally, the RF model outperformed other models was selected as the final model, which was implemented into a web server.

3.2 Performance of Nine Classifiers in Nested Five-Fold Cross-Validation

The performance of the nine predictive models in the nested five-fold cross-validation is shown in **Table 2**, and the ROC curves are illustrated in **Figure 2A**. For a detailed description of nested five-validation cross-validation, please refer to the **Supplementary Material**. In **Table 2**, RF model outperformed the other eight machine learning models. All five evaluation metrics reached the highest level. It has an AUC score of 0.9030, ACC value of 81.90%, MCC value of 0.6390, SN value of 79.14% and SP value of 84.66% (see **Table 2**). Moreover, compared with the eight conventional machine learning classifiers, the performance of LSTM is not satisfactory. Except for SP, the values of the other four evaluation metrics of LSTM model were the lowest. The overall performance of traditional machine learning algorithms is generally better than LSTM.

TABLE 3 | The prediction performances of different classifiers in the independent testing dataset.

Scoring Method	Classifier	SN(%)	SP(%)	ACC(%)	MCC	AUC
F-score	**RF**	**76.77**	**77.78**	**77.27**	**0.5455**	**0.8332**
	rbfSVM	78.79	73.74	76.26	0.5259	0.8241
	KNN	70.71	66.67	68.69	0.3740	0.6869
	DT	69.70	61.62	65.66	0.3142	0.6574
	linearSVM	64.65	74.75	69.70	0.3960	0.7656
	LSTM	58.59	63.64	61.11	0.2225	0.6041
	AdaBoost	64.65	68.69	66.67	0.3336	0.7389
	GentleBoost	74.75	66.67	70.71	0.4155	0.7831
	LogitBoost	67.68	77.78	72.73	0.4569	0.7798

TABLE 4 | Comparison of datasets for three predictors.

	BBPpred	B3Pred	BBPpredict
Data source	Positive: Brainpeps, PepBank, articles, SATPdb Negative: UniProt	Positive: B3Pdb Negative: UniProt	Positive: Brainpeps, B3Pdb, BBPpred, B3Pred, articles Negative: UniProt
Article search deadline		22 July 2020	Nov. 2021
Article number	7	271	300
Positive sample number	119 (training:100, testing: 19)	269 (training:215, testing: 54)	425 (training:326, testing: 99)
Negative sample number	119 (training:100, testing: 19)	2,690 (training: 2,152, testing:538)	425 (training:326, testing: 99)
Peptide length	5–50	6–30	5–50

3.3 Performance of Nine Classifiers on the Independent Testing Dataset

To determine the final model for constructing BBPpredict, performance evaluation on the independent testing dataset is much more convincing than five-fold cross-validation. According to the steps in the method section, nine classification models are established by using the training dataset. The independent testing dataset was then utilized to test the performance of these models. As depicted in **Table 3** and **Figure 2B**, in term of AUC score, the RF model also performed best, with a score of 0.8332, higher than rbfSVM, linearSVM, KNN, DT, GentleBoost, AdaBoost, LogitBoost and LSTM classifiers by 0.0091, 0.0676, 0.1463, 0.1758, 0.0501, 0.0943, 0.0534 and 0.2291 respectively. In terms of accuracy and MCC, the RF classifier also achieved impressive values, with scores of 77.27% and 0.5455, which are better than other eight classifier algorithm predictors. Furthermore, the LSTM classifier had the weakest generalization ability. In addition, results of the reproducibility analysis for nine classifiers are highly consistent with the above results (see **Supplementary Table S9**).

3.4 Performance of the Predictions Under the Combinations of RF With Three Feature Scoring Methods

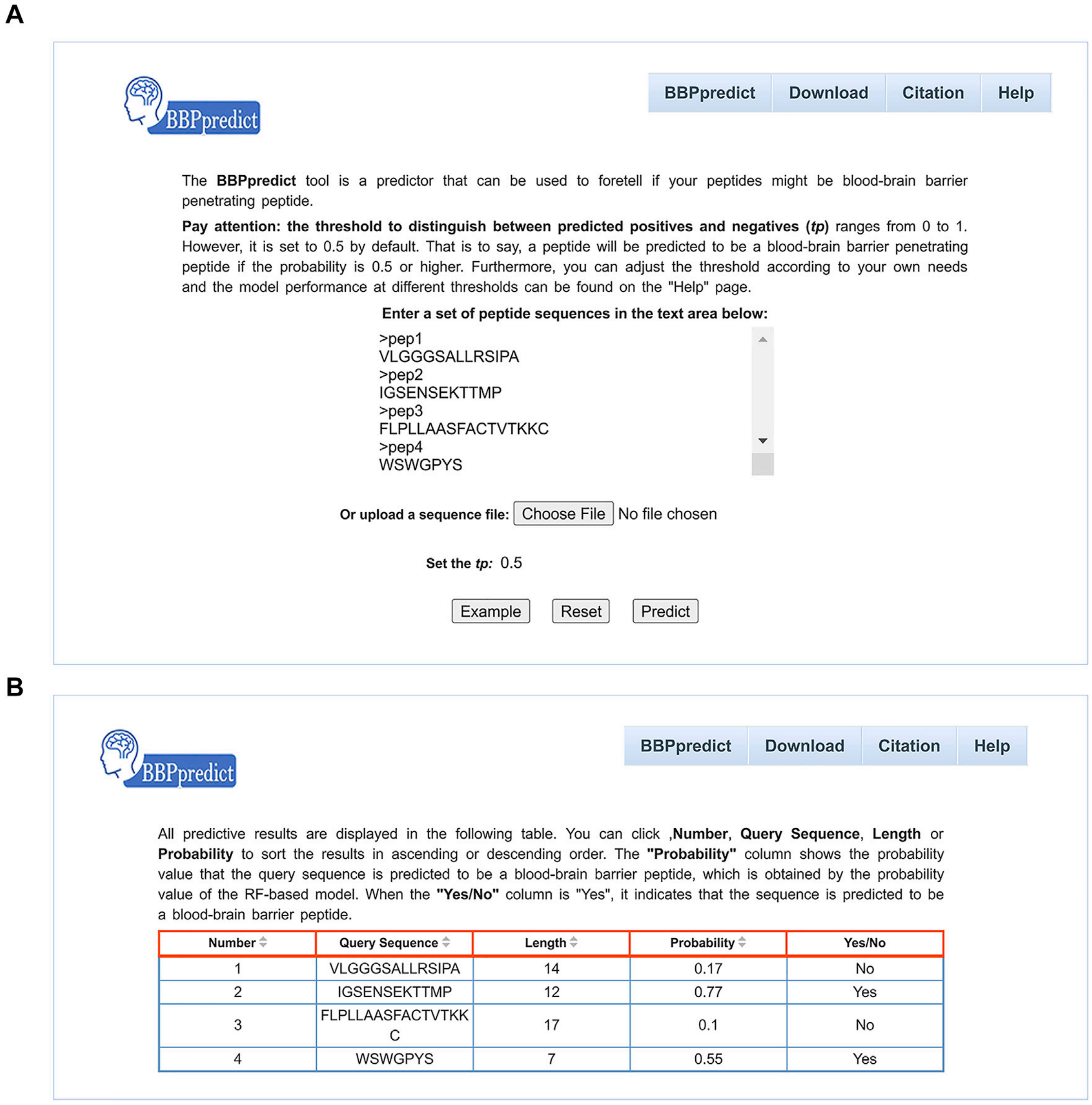
We also used the RF algorithm with optimal features selected by Pearson and Lasso feature scoring methods to construct prediction model. As shown in **Supplementary Tables S4,5**, the model under the combination of RF and F-score achieved the second highest AUC value in the nested five-fold cross-validation and the highest AUC value in the independent testing dataset. Therefore, we finally chose the combination of RF and F-score to build the final model based on 184 features and tree depth of 63.

TABLE 5 | The prediction performances of different predictors.

Predictor	SN(%)	SP(%)	ACC(%)	MCC
BBPpredict	**76.77**	**77.78**	**77.27**	**0.5455**
BBPpred	67.68	65.66	66.67	0.3334
B3Pred	70.71	64.65	67.68	0.3542

3.5 Prediction Performance of Existing Predictors

There are two published predictors for identifying BBPs, B3Pred and BBPpred. These predictors and our predictor are based on peptide sequence information. The comparison of datasets of existing predictors and our proposed predictor can be seen in **Table 4** (Detailed comparison can be found in **Supplementary Table S8**). To be fair, an independent testing dataset, which is completely independent of three predictors' training datasets, was used to compare their performance. As shown in **Table 5**, compared with the existing BBPs predictors, our predictor achieved a promising performance (ACC = 77.27%, SN = 76.77%, SP = 77.78% and MCC = 0.5455), it outperformed BBPpred and B3Pred, higher than them by 10.6% and 9.59% in accuracy, severally, with MCC increasing 0.2121 and 0.1913, respectively. There were remarkable improvements in sensitivity and specificity (see **Table 5**). The above results demonstrate that BBPpredict is more capable of distinguishing between BBPs and non-BBPs than BBPpred and B3Pred.



The **BBPpredict** tool is a predictor that can be used to foretell if your peptides might be blood-brain barrier penetrating peptide.

Pay attention: the threshold to distinguish between predicted positives and negatives (*tp*) ranges from 0 to 1. However, it is set to 0.5 by default. That is to say, a peptide will be predicted to be a blood-brain barrier penetrating peptide if the probability is 0.5 or higher. Furthermore, you can adjust the threshold according to your own needs and the model performance at different thresholds can be found on the "Help" page.

Enter a set of peptide sequences in the text area below:

```
>pep1
VLGGGSALLRSIPA
>pep2
IGSENSEKTTMP
>pep3
FLPLLAASFACTVTKKC
>pep4
WSWGPYS
```

Or upload a sequence file: Choose File No file chosen

Set the *tp*: 0.5

All predictive results are displayed in the following table. You can click ,**Number**, **Query Sequence**, **Length** or **Probability** to sort the results in ascending or descending order. The **"Probability"** column shows the probability value that the query sequence is predicted to be a blood-brain barrier peptide, which is obtained by the probability value of the RF-based model. When the **"Yes/No"** column is "Yes", it indicates that the sequence is predicted to be a blood-brain barrier peptide.

Number ⇕	Query Sequence ⇕	Length ⇕	Probability ⇕	Yes/No
1	VLGGGSALLRSIPA	14	0.17	No
2	IGSENSEKTTMP	12	0.77	Yes
3	FLPLLAASFACTVTKK C	17	0.1	No
4	WSWGPYS	7	0.55	Yes

FIGURE 3 | Web interface of BBPpredict. **(A)** The query sequences and threshold of the probability value (tp) are required to be submitted in the input interface. **(B)** The result page returned from BBPpredict.

TABLE 6 | Performance of BBPpredict in the independent testing dataset when tp changes.

tp	SN (%)	SP (%)	ACC (%)	MCC
0.1	100	11.11	55.56	0.2425
0.2	98.99	29.29	64.14	0.3944
0.3	94.95	44.44	69.70	0.4564
0.4	86.87	64.65	75.76	0.5284
0.5	76.77	77.78	77.27	0.5455
0.6	58.59	82.83	70.71	0.4269
0.7	45.45	90.91	68.18	0.4082
0.8	36.36	96.97	66.67	0.4191
0.9	13.13	97.98	55.56	0.2100
0.95	5.05	97.98	51.51	0.0820

3.6 Web Server Implementation

To facilitate users to identify BBPs, we established an online web service named BBPpredict that was implemented based on optimized features and the RF model. BBPpredict can be accessed at http://i.uestc.edu.cn/BBPpredict/cgi-bin/BBPpredict.pl, conveniently. The web service of BBPpredict was developed by using Perl and Html, *Python* and Matlab. Users can paste peptide sequences or upload a sequence file to predict BBPs, as illustrated in **Figure 3A**. Then click the “Predict” button to make predictions, and the predictive results are depicted in **Figure 3B**.

BBPpredict allows users to adjust the threshold of the probability value (tp) to distinguish between predicted positives and negatives, which can range from 0 to 1. As shown in **Table 6**, with the increase of tp, the value of SN decreases, and the SP increases. When tp is 0.5, ACC achieves the highest score of 77.27%, MCC reaches the highest value of 0.5455.

4 DISCUSSION

In the past 30 years, many studies have demonstrated that BBPs are promising for the treatment of CNS diseases. BBPs can pass through the BBB and enter brain parenchyma without destroying BBB. Them can be used as transport carriers of DNA, RNA and protein as well as drug-assisted treatment and diagnosis of CNS diseases. However, the discovery of BBPs is still a thorny problem. Only a few hundreds of peptides have been experimentally confirmed as BBPs so far, since BBPs were discovered in 1996 (Banks and Kastin, 1996). Therefore, to facilitate the treatment of CNS diseases, it is necessary to employ computational methods to rapidly discover and identify more novel BBPs.

At present, two BBPs predictors, BBPpred (Dai et al., 2021) and B3Pred (Kumar et al., 2021a), have been proposed. Compared with these two predictors, our developed BBPpredict tool was based on a larger training dataset (as shown in **Table 4**). Besides the difference of the training dataset, a nested cross-validation strategy was utilized in the construction of BBPpredict. For common cross-validation, the model parameters were determined manually, and the accuracy based on the cross-validation would be affected by the artificial selection of model parameters, which usually overestimate the accuracy based on the cross-validation. For nested cross-validation, the model parameters were determined automatically. We speculated that this might be a reason why the previous two predictors had better performance in the cross-validation but had poor performance in our independent testing dataset. BBPpredict showed a large improvement in performance with nearly 6% sensitivity, 12% specificity, 10% accuracy and 0.20 MCC increase, compared with BBPpred and B3Pred. The elevated performance can save cost for researchers to identify BBPs and speed up the discovery of BBPs.

The BBPpredict website allows users to set the tp value. We tested the performance of BBPpredict in the independent testing dataset and provided sensitivity and specificity values under different tp values, which can serve as reference for users and increases the confidence they can have about the positive predictions.

We also reconstructed the BBPs/non-BBPs classification models with different machine learning methods using the new feature vectors that were generated from 16 feature extraction methods, including AAC, DPC, CKSAAGP, PAAC, GAAC, Grouped Di-Peptide Composition (GDPC) (Chen et al., 2018; Chen et al., 2020), Dipeptide Deviation from Expected Mean (DDE) (Chen et al., 2020), Composition (CTDC) (Dubchak et al., 1995; Dubchak et al., 1999; Chen et al., 2020), Transition (CTDT) (Dubchak et al., 1995; Dubchak et al., 1999; Chen et al., 2020), Distribution (CTDD) (Chen et al., 2020), Amphiphilic Pseudo-Amino Acid Composition (APAAC) (Chou, 2005; Jiao and Du, 2016), Quasi-sequence-order (QSOrder) (Chen et al., 2020), Normalized Moreau-Broto Autocorrelation (NMBroto) (Chen et al., 2018), Geary correlation (Geary) (Chen et al., 2020), Moran correlation (Moran) (Feng and Zhang, 2000; Chen et al., 2020) and Sequence-Order-Coupling Number (SOCNumber) (Lim et al., 2015). The detailed description of the last 11 feature encoding approaches can be found in the **Supplementary Materials**. F-score was used for feature sorting, grid search with five-fold cross-validation was utilized to select the best feature parameters and the best classifier parameters for different classifiers. **Supplementary Tables S6,7** illustrated the detailed results of five-fold cross-validation and independent testing dataset of reconstructed classification models, respectively. However, the addition of feature encoding methods did not improve the classification performance of the model. We speculate that it is caused by the high correlation between the extracted features based on different feature extracting methods, which might induce highly correlated features in the final feature subset. As the feature number is limited, the highly correlated features might reduce useful information for model construction. Another possible reason might be the limited sample size, which might cause high false positive rate during the process of feature selection. The increase of feature size would lead to the increase of false positive features, which would affect the robustness of the predictive model.

BBPs pass through BBB via six penetration mechanisms, including diffusion transport, carrier-mediated transcytosis, efflux transporter, receptor-mediated transcytosis, adsorptive-mediated transcytosis and cell-mediated transcytosis (Zhou et al., 2021). The abilities of BBPs to penetrate BBB vary depending on their penetration mechanisms (Sánchez-Navarro et al., 2017). Therefore, we speculate the differences in their penetration mechanisms may affect the reliability of screening in the procession of model construction. However, BBPs of distinct penetration mechanisms were not further divided when constructing the positive sample of BBPpred, B3Pred and BBPpredict, because the number of BBPs for a specific transport mechanism is insufficient to construct a BBP predictor.

In the present work, we utilized RF algorithm to construct BBP predictor. The RF is an ensemble algorithm which is composed of several weak classifiers (decision trees). Our constructed model contains 63 decision trees. We speculate that these different decision trees might cover different penetration mechanisms and it might be the reason why the RF algorithm is superior to other machine learning algorithms. In the future, if the number of BBPs with a certain transport mechanism increase, it is possible and preferable to construct new BBP predictors using BBPs with the same penetrating mechanism.

5 CONCLUSION

In this study, we proposed an RF-based predictor for identifying BBPs, called BBPpredict, which is available for free at http://i.uestc.edu.cn/BBPpredict/cgi-bin/BBPpredict.pl. To find the optimal classifier, eight traditional machine learning algorithms and one deep learning algorithm were used for developing models. The RF algorithm was selected to construct BBPpredict after comparing the results of nine classifiers in the five-fold cross-validation and independent test. The RF-based model reached an AUC of 0.9030 with an accuracy of 81.90% and an AUC of 0.8332 with an accuracy of 77.27% in the nested five-fold cross-validation and independent testing dataset, respectively. We also compared BBPpredict with two existing BBPs predictors, BBPpred and B3Pred. The results showed that BBPpredict was remarkably higher in accuracy, MCC, sensitivity and specificity than these two predictors. BBPpredict is a promising classification model, and we expect it to play a positive role in the discovery of BBPs to facilitate the development of drugs for CNS diseases.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

XC, QZ, BL, CL, SY, JL, BH, HC, and JH developed the web interface of the predictor. XC conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft. BH, HC, and JH conceived and designed the experiments, authored or reviewed drafts of the paper. All authors approved the final draft.

FUNDING

This work was supported by the National Natural Science Foundation of China (grant numbers: 61901130, 61901129, and 62071099), Science and Technology Department of Guizhou Province (Grant Numbers: (2020)1Y407 and ZK [2022]-General-038) and Guizhou University (Grant Numbers: (2018)54, (2018)55 and (2020)5).

ACKNOWLEDGMENTS

The authors are grateful to the reviewers for their valuable suggestions and comments, which will lead to the improvement of this paper.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.845747/full#supplementary-material

REFERENCES

Banks, W. A. (2016). From Blood-Brain Barrier to Blood-Brain Interface: New Opportunities for CNS Drug Delivery. *Nat. Rev. Drug Discov.* 15 (4), 275–292. doi:10.1038/nrd.2015.21

Banks, W. A., and Kastin, A. J. (1996). Passage of Peptides across the Blood-Brain Barrier: Pathophysiological Perspectives. *Life Sci.* 59 (23), 1923–1943. doi:10.1016/s0024-3205(96)00380-3

Bhadra, P., Yan, J., Li, J., Fong, S., and Siu, S. W. I. (2018). AmPEP: Sequence-Based Prediction of Antimicrobial Peptides Using Distribution Patterns of Amino Acid Properties and Random forest. *Sci. Rep.* 8 (1), 1697. doi:10.1038/s41598-018-19752-w

Bhasin, M., and Raghava, G. P. S. (2004). Classification of Nuclear Receptors Based on Amino Acid Composition and Dipeptide Composition. *J. Biol. Chem.* 279 (22), 23262–23266. doi:10.1074/jbc.M401932200

Chang, C.-C., and Lin, C.-J. (2011). LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.* 2 (3), 1–27. doi:10.1145/1961189.1961199

Chen, K., Jiang, Y., Du, L., and Kurgan, L. (2009). Prediction of Integral Membrane Protein Type by Collocated Hydrophobic Amino Acid Pairs. *J. Comput. Chem.* 30 (1), 163–172. doi:10.1002/jcc.21053

Chen, K., Kurgan, L. A., and Ruan, J. (2007b). Prediction of Flexible/rigid Regions from Protein Sequences Using K-Spaced Amino Acid Pairs. *BMC Struct. Biol.* 7, 25. doi:10.1186/1472-6807-7-25

Chen, K., Kurgan, L. A., and Ruan, J. (2008). Prediction of Protein Structural Class Using Novel Evolutionary Collocation-Based Sequence Representation. *J. Comput. Chem.* 29 (10), 1596–1604. doi:10.1002/jcc.20918

Chen, K., Kurgan, L., and Rahbari, M. (2007a). Prediction of Protein Crystallization Using Collocation of Amino Acid Pairs. *Biochem. Biophysical Res. Commun.* 355 (3), 764–769. doi:10.1016/j.bbrc.2007.02.040

Chen, Z., Zhao, P., Li, F., Leier, A., Marquez-Lago, T. T., Wang, Y., et al. (2018). iFeature: a Python Package and Web Server for Features Extraction and Selection from Protein and Peptide Sequences. *Bioinformatics* 34 (14), 2499–2502. doi:10.1093/bioinformatics/bty140

Chen, Z., Zhao, P., Li, F., Marquez-Lago, T. T., Leier, A., Revote, J., et al. (2020). iLearn: an Integrated Platform and Meta-Learner for Feature Engineering, Machine-Learning Analysis and Modeling of DNA, RNA and Protein Sequence Data. *Brief Bioinform* 21 (3), 1047–1057. doi:10.1093/bib/bbz041

Chou, K.-C. (2001). Prediction of Protein Cellular Attributes Using Pseudo-amino Acid Composition. *Proteins* 43 (3), 246–255. doi:10.1002/prot.1035

Chou, K.-C. (2005). Using Amphiphilic Pseudo Amino Acid Composition to Predict Enzyme Subfamily Classes. *Bioinformatics* 21 (1), 10–19. doi:10.1093/bioinformatics/bth466

Dai, R., Zhang, W., Tang, W., Wynendaele, E., Zhu, Q., Bin, Y., et al. (2021). BBPpred: Sequence-Based Prediction of Blood-Brain Barrier Peptides with Feature Representation Learning and Logistic Regression. *J. Chem. Inf. Model.* 61 (1), 525–534. doi:10.1021/acs.jcim.0c01115

Drappatz, J., Brenner, A., Wong, E. T., Eichler, A., Schiff, D., Groves, M. D., et al. (2013). Phase I Study of GRN1005 in Recurrent Malignant Glioma. *Clin. Cancer Res.* 19 (6), 1567–1576. doi:10.1158/1078-0432.Ccr-12-2481

Dubchak, I., Muchnik, I., Holbrook, S. R., and Kim, S. H. (1995). Prediction of Protein Folding Class Using Global Description of Amino Acid Sequence. *Proc. Natl. Acad. Sci. U.S.A.* 92 (19), 8700–8704. doi:10.1073/pnas.92.19.8700

Dubchak, I., Muchnik, I., Mayor, C., Dralyuk, I., and Kim, S.-H. (1999). Recognition of a Protein Fold in the Context of the SCOP Classification. *Proteins* 35 (4), 401–407. doi:10.1002/(sici)1097-0134(19990601)35:4<401::aid-prot3>3.0.co;2-k

Feng, Z.-P., and Zhang, C.-T. (2000). Prediction of Membrane Protein Types Based on the Hydrophobic index of Amino Acids. *J. Protein Chem.* 19 (4), 269–275. doi:10.1023/a:1007091128394

He, B., Chen, H., and Huang, J. (2019). PhD7Faster 2.0: Predicting Clones Propagating Faster from the Ph.D.-7 Phage Display Library by Coupling PseAAC and Tripeptide Composition. *PeerJ* 7, e7131. doi:10.7717/peerj.7131

He, B., Kang, J., Ru, B., Ding, H., Zhou, P., and Huang, J. (2016). SABinder: A Web Service for Predicting Streptavidin-Binding Peptides. *Biomed. Res. Int.* 2016, 1–8. doi:10.1155/2016/9175143

Hochreiter, S., and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Comput.* 9 (8), 1735–1780. doi:10.1162/neco.1997.9.8.1735

Jiao, Y.-S., and Du, P.-F. (2016). Predicting Golgi-Resident Protein Types Using Pseudo Amino Acid Compositions: Approaches with Positional Specific Physicochemical Properties. *J. Theor. Biol.* 391, 35–42. doi:10.1016/j.jtbi.2015.11.009

Kumar, V., Agrawal, P., Kumar, R., Bhalla, S., Usmani, S. S., Varshney, G. C., et al. (2018). Prediction of Cell-Penetrating Potential of Modified Peptides Containing Natural and Chemically Modified Residues. *Front. Microbiol.* 9, 725. doi:10.3389/fmicb.2018.00725

Kumar, V., Patiyal, S., Dhall, A., Sharma, N., and Raghava, G. P. S. (2021a). B3Pred: A Random-Forest-Based Method for Predicting and Designing Blood-Brain Barrier Penetrating Peptides. *Pharmaceutics* 13 (8), 1237. doi:10.3390/pharmaceutics13081237

Kumar, V., Patiyal, S., Kumar, R., Sahai, S., Kaur, D., Lathwal, A., et al. (2021b). B3Pdb: an Archive of Blood-Brain Barrier-Penetrating Peptides. *Brain Struct. Funct.* 226 (8), 2489–2495. doi:10.1007/s00429-021-02341-5

Kurzrock, R., Gabrail, N., Chandhasin, C., Moulder, S., Smith, C., Brenner, A., et al. (2012). Safety, Pharmacokinetics, and Activity of GRN1005, a Novel Conjugate of Angiopep-2, a Peptide Facilitating Brain Penetration, and Paclitaxel, in Patients with Advanced Solid Tumors. *Mol. Cancer Ther.* 11 (2), 308–316. doi:10.1158/1535-7163.Mct-11-0566

Lee, T.-Y., Lin, Z.-Q., Hsieh, S.-J., Bretaña, N. A., and Lu, C.-T. (2011). Exploiting Maximal Dependence Decomposition to Identify Conserved Motifs from a Group of Aligned Signal Sequences. *Bioinformatics* 27 (13), 1780–1787. doi:10.1093/bioinformatics/btr291

Li, F.-M., and Wang, X.-Q. (2016). Identifying Anticancer Peptides by Using Improved Hybrid Compositions. *Sci. Rep.* 6, 33910. doi:10.1038/srep33910

Lim, S., Kim, W.-J., Kim, Y.-H., Lee, S., Koo, J.-H., Lee, J.-A., et al. (2015). dNP2 Is a Blood-Brain Barrier-Permeable Peptide Enabling ctCTLA-4 Protein Delivery to Ameliorate Experimental Autoimmune Encephalomyelitis. *Nat. Commun.* 6, 8244. doi:10.1038/ncomms9244

Muttenthaler, M., King, G. F., Adams, D. J., and Alewood, P. F. (2021). Trends in Peptide Drug Discovery. *Nat. Rev. Drug Discov.* 20 (4), 309–325. doi:10.1038/s41573-020-00135-8

Nance, E., Pun, S. H., Saigal, R., and Sellers, D. L. (2022). Drug Delivery to the central Nervous System. *Nat. Rev. Mater* 7 (4), 314–331. doi:10.1038/s41578-021-00394-w

Nonaka, M., Suzuki-Anekoji, M., Nakayama, J., Mabashi-Asazuma, H., Jarvis, D. L., Yeh, J.-C., et al. (2020). Overcoming the Blood-Brain Barrier by Annexin A1-Binding Peptide to Target Brain Tumours. *Br. J. Cancer* 123 (11), 1633–1643. doi:10.1038/s41416-020-01066-2

Oller-Salvia, B., Sánchez-Navarro, M., Giralt, E., and Teixidó, M. (2016). Blood-brain Barrier Shuttle Peptides: an Emerging Paradigm for Brain Delivery. *Chem. Soc. Rev.* 45 (17), 4690–4707. doi:10.1039/c6cs00076b

Sánchez-Navarro, M., Giralt, E., and Teixidó, M. (2017). Blood-brain Barrier Peptide Shuttles. *Curr. Opin. Chem. Biol.* 38, 134–140. doi:10.1016/j.cbpa.2017.04.019

Saravanan, V., and Gautham, N. (2015). Harnessing Computational Biology for Exact Linear B-Cell Epitope Prediction: A Novel Amino Acid Composition-Based Feature Descriptor. *OMICS: A J. Integr. Biol.* 19 (10), 648–658. doi:10.1089/omi.2015.0095

Terstappen, G. C., Meyer, A. H., Bell, R. D., and Zhang, W. (2021). Strategies for Delivering Therapeutics across the Blood-Brain Barrier. *Nat. Rev. Drug Discov.* 20 (5), 362–383. doi:10.1038/s41573-021-00139-y

Van Dorpe, S., Bronselaer, A., Nielandt, J., Stalmans, S., Wynendaele, E., Audenaert, K., et al. (2012). Brainpeps: the Blood-Brain Barrier Peptide Database. *Brain Struct. Funct.* 217 (3), 687–718. doi:10.1007/s00429-011-0375-0

Walzer, M., and Vizcaíno, J. A. (2020). Review of Issues and Solutions to Data Analysis Reproducibility and Data Quality in Clinical Proteomics. *Methods Mol. Biol.* 2051, 345–371. doi:10.1007/978-1-4939-9744-2_15

Wang, J., Li, J., Yang, B., Xie, R., Marquez-Lago, T. T., Leier, A., et al. (2019). Bastion3: a Two-Layer Ensemble Predictor of Type III Secreted Effectors. *Bioinformatics* 35 (12), 2017–2028. doi:10.1093/bioinformatics/bty914

Wei, L., Tang, J., and Zou, Q. (2017a). SkipCPP-Pred: an Improved and Promising Sequence-Based Predictor for Predicting Cell-Penetrating Peptides. *BMC Genomics* 18 (Suppl. 7), 742. doi:10.1186/s12864-017-4128-1

Wei, L., Xing, P., Su, R., Shi, G., Ma, Z. S., and Zou, Q. (2017b). CPPred-RF: A Sequence-Based Predictor for Identifying Cell-Penetrating Peptides and Their Uptake Efficiency. *J. Proteome Res.* 16 (5), 2044–2053. doi:10.1021/acs.jproteome.7b00019

Xie, R., Wu, Z., Zeng, F., Cai, H., Wang, D., Gu, L., et al. (2021). Retro-enantio Isomer of Angiopep-2 Assists Nanoprobes across the Blood-Brain Barrier for Targeted Magnetic Resonance/fluorescence Imaging of Glioblastoma. *Sig Transduct Target. Ther.* 6 (1), 309. doi:10.1038/s41392-021-00724-y

Zhao, Y.-W., Lai, H.-Y., Tang, H., Chen, W., and Lin, H. (2016). Prediction of Phosphothreonine Sites in Human Proteins by Fusing Different Features. *Sci. Rep.* 6, 34817. doi:10.1038/srep34817

Zhou, X., Smith, Q. R., and Liu, X. (2021). Brain Penetrating Peptides and Peptide-Drug Conjugates to Overcome the Blood-Brain Barrier and Target CNS Diseases. *WIREs Nanomed Nanobiotechnol* 13 (4), e1695. doi:10.1002/wnan.1695

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.