 doi:10.20944/preprints202106.0279.v1

In silico tool for predicting and designing Blood-Brain Barrier Penetrating Peptides

Vinod Kumar[1,2], Sumeet Patiyal[1], Anjali Dhall[1], Neelam Sharma[1], *Gajendra P.S. Raghava[1]

[1]Department of Computational Biology, Indraprastha Institute of Information Technology, Okhla, India

[2]Bioinformatics Centre, CSIR-Institute of Microbial Technology, Sector-39A, Chandigarh, India

* Corresponding author

Professor, Department of Computational Biology, Indraprastha Institute of Information

Technology, Okhla Industrial Estate, Phase III, New Delhi 110020.

India. Tel.: +91 011 26907444

E-mail address: raghava@iiitd.ac.in

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 June 2021 doi:10.20944/preprints202106.0279.v1

ABSTRACT

Blood-brain-barrier is a major obstacle in treating brain-related disorders as it does not allow to deliver drugs in the brain. In order to facilitate delivery of drugs in brain, we developed a method for predicting blood-brain-barrier penetrating peptides. These blood-brain barriers penetrating peptides (B3PPs) can act as therapeutic as well as drug delivery agents. We trained, tested, and evaluated our models on blood-brain-barrier peptides obtained from the B3Pdb database. First, we compute a wide range of peptide features then we select relevant peptide features. Finally, we developed numerous machine learning-based models for predicting blood-brain-barrier peptides using selected features. Our model based on random forest performed best on the top 80 selected features and achieved a maximum 85.08% accuracy with 0.93 AUROC. We also developed a web server, B3pred that implements our best models. It has three major modules that allow users to; i) predict B3PPs, ii) scanning B3PPs in a protein sequence, and iii) designing B3PPs using analogs. Our web server and standalone software is freely available at https://webs.iiitd.edu.in/raghava/b3pred/.

Keywords: Blood-Brain Barrier, Penetrating Peptides, Machine learning techniques, Drug delivery, Therapeutic peptides

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 June 2021 doi:10.20944/preprints202106.0279.v1

Introduction

The blood-brain barrier (BBB) is the primary barrier between the brain's interstitial fluid and the blood. It makes the connection between the central nervous system (CNS) and the peripheral nervous system (PNS) [1–4]. The neurovascular unit (NVU) is the structural and functional unit of the BBB, formed by the neurons, macrophages, endothelial cells, astrocytes, and pericytes [5] (**Fig.1**). NUV regulates the biochemical environment between the blood and the brain, which is essential for neural functions. The endothelial cells of the NUV allow the entry or exit of the molecules like glucose, amino-acids, proteins/peptides in the CNS [6–8]. In the last few decades, researchers have made many attempts to develop drug delivery systems that can deliver drugs in the brain. Despite advances made by the scientific community in developing drug delivery systems, it is still challenging to penetrate the BBB [9].

In the past, researchers have attempted to develop peptides/proteins-based drug delivery vehicles. In this approach, a major challenge is to identify peptides that can penetrate the BBB [10]. In addition, researchers are also exploring peptide-based therapeutics to treat CNS-associated diseases such as neurodegenerative disorders like Parkinson's disease, Alzheimer's disease [11,12], and glioblastoma [13]. It means peptides can be used as therapeutic agents as well as drug delivery vehicles. In recent studies, numerous peptides such as shuttle peptides [14], self-assembled peptides [15], and peptide-decorated nanoparticles [16] have been used for efficient drug delivery to the brain. Some neuropeptides are utilized as potential therapeutic targets against many neurological diseases such as epilepsy [17,18], depression [19,20], and neuroimmune disorders [21]. Due to the low toxicity of these peptides, they may act as potential peptide-based drugs candidates against neurological diseases. The major limitation of these peptide-based drugs is its low bioavailability, short half-life [22], and penetration of BBB [23]. For example, tumor homing peptides (THPs) [24] and cell-penetrating peptides (CPPs) [25] can be used as drug delivery vehicles [26,27]. The tumor homing peptides need a carrier to cross the BBB, while selected CPPs can directly pass through the BBB [28].

Insert Fig. 1 here

Fig. 1 Representation of the Blood-Brain Barrier and B3PPs to cross into CNS

CPPs are short peptides, act as molecular delivery vehicles, and are able to deliver various therapeutic molecules inside a cell [29][30]. There are CPPs that can even cross the blood-brain barrier are called blood-brain barrier penetrating peptides (B3PPs). These B3PPs can be used to deliver several cargo molecules (e.g., peptides/proteins, siRNA, plasmid DNA) in the brain [31–34]. Mainly these peptides are obtained from the naturally occurring

proteins/peptides like signal peptides, RNA/DNA-binding proteins, viral proteins, antimicrobial peptides, etc.[35]. Several studies have shown that B3PPs may be synthesized chemically or designed with rDNA technology [36–38]; to enhance the stability and half-life of the B3PPs [39]. In the past, several methods have been developed for predicting cell-penetrating peptides, such as cellPPD [40], SkipCPP-Pred [41], CPPred-RF [42], KELM-CPPpred [43], CellPPDMod [44], and CPPred-FL [45]. In addition, various methods have been developed for predicting chemical-based drug delivery vehicles to cross the blood-brain barrier [46–48]. In contrast, a limited attempt has been made to develop methods to predict B3PPs. Recently, Dai et al. developed an *in-silico* method BBPpred to identify B3PPs [49].

In this study, we have developed a computational tool named "B3Pred" for predicting B3PPs with high reliability and precision. This method is able to classify BBPs/non-BBPs and CPPs/BBPs and uses a large dataset for training and validation. We used three datasets, i.e., Dataset_1 (269 B3PPs and 269 CPPs), Dataset_2 (269 B3PPs, and 269 non-B3PPs), and Dataset_3 (269 B3PPs and 2690 non-B3PPs) for training and validation. We have used more than 9000 descriptors/features for the generation of prediction models using several machine learning techniques such as RF, DT, LR, XGB, SVM, and GBM. Further, in order to serve the scientific community working in this era, we have provided a web server and a standalone package, which is freely available at (https://webs.iiitd.edu.in/raghava/b3pred/)

Material and Methods

Dataset Collection

In this study, we have collected 465 blood-brain barrier penetrating peptides (B3PPs) from the B3Pdb database (https://webs.iiitd.edu.in/raghava/b3pdb/). We consider B3PPs having a length of more than five amino acid (AA) residues and less than or equal to 30 AA residues. For the positive dataset, we got 269 unique B3PPs. The major challenge of this type of study is to generate an authenticated negative dataset. We have used three negative datasets in this study. Firstly, we collected unique 269 cell-penetrating peptides (CPPs) [50] other than B3PPs and called them non-B3PPs or negative dataset1. In negative dataset-2, we randomly generated 269 non-B3PPs from the Swiss-Prot database [51]. Our third negative dataset is ten times the positive dataset, i.e., 2690 unique non-B3PPs randomly generated using the Swiss-Prot

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 June 2021 doi:10.20944/preprints202106.0279.v1

database. Finally, we got three datasets, i.e., Dataset_1 (269 B3PPs and 269 CPPs), Dataset_2 (269 B3PPs, and 269 non-B3PPs), and Dataset_3 (269 B3PPs and 2690 non-B3PPs).

Amino Acid Composition

Amino acid composition (AAC) analysis of peptides helps us find out whether there is any amino acid compositional similarity or any compositional differences in different datasets. We compared the amino acid composition of B3PPs, CPPs, and randomly generated peptides for the negative dataset. The following equation calculates AAC:

$$AAC_{(i)} = \frac{AAR_i}{TNR} \times 100 \quad [1]$$

Where AAC $_{(i)}$ is the percentage composition of the amino acid (i); AAR_i is the number of residues of type i, and TNR is the total number of residues in the peptide [52].

Two Sample Logo

Two sample logo (TSL) tool [53] used in this study to identify the amino-acid preference at a specific position in the peptide sequences. This tool needs an input amino-acid sequence vector of fixed length since the minimum size of peptides in all datasets is five residues; hence we select five residues from the N-terminal and five amino-acids from the C-terminal of the peptide sequences. To create a fixed input vector, the N-terminus side residues and C-terminus residues were grouped together to generate a sequence of 10 amino-acid residues. We used 10-residues sequences generated from our dataset peptides to develop TSL. To build two sample logos, we have used all B3PPs and non-B3PPs of three different negative datasets.

Generation of Peptide Features

In order to calculate a wide range of features from the protein or peptide sequences, we use the Pfeature package [54]. Pfeature is used to generate thousands of features/descriptors. Currently, we compute the composition-based module of Pfeature to calculate >9000 descriptors of peptide sequences for positive and negative datasets. This module calculates fifteen different features (AAC, DPC, RRI, DDOR, SE, SER, SEP, CTD, CeTD, PAAC, APAAC, QSO, TPC,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 June 2021 doi:10.20944/preprints202106.0279.v1

ABC, SOCN). The input vector of 9189 descriptors is used further for feature selection and machine learning purposes.

Feature Selection

In this study, we have used the SVC-L1 feature selection technique to extract an essential set of features from all the datasets. We choose the SVC-L1 method because it is much faster than other feature selection methods [55]. This method applies the L1 penalty to select a relevant set of features after selecting the non-zero coefficients. SVC-L1 mainly considers regularization and loss function. During the optimization process, the L1 regularization generates a sparse matrix by choosing some model features. The other important parameter used in this technique is the "C" parameter; its value is directly proportional to the selected features. Smaller the value of "C", the lesser number of features determined by the method. Currently, we choose the default value (i.e., 0.01) of "C" parameter [56]. Using SVC-L1, 73 important sets of features have been identified from 9189 features for Dataset_1 (B3PPs and CPPs peptides) and Dataset_2 (B3PPs and balanced non-B3PPs). Similarly, 145 features have been selected for the Dataset_3 (i.e., B3PPs and random non-B3PPs).

Feature Ranking

After the selection of an important set of features, we rank the features on the basis of their importance of classification using a feature selector method. The Feature-selector method is based on a decision tree-like algorithm and uses Light Gradient Boosting Machine (LightGBM) [57]. It computes the rank of each feature on the basis of the feature that is used to split the dataset across all the trees. Further, the top-most ranked features for each dataset were used in different machine learning techniques for the classification of B3PPs and non-B3PPs.

Machine Learning Techniques

In order to classify B3PPs and non-B3PPs, we have used several machine learning algorithms. In this study, we mainly implement Decision tree (DT), Random Forest (RF), Logistic Regression (LR), k-nearest neighbors (KNN), Gaussian Naive Bayes (GNB), XGBoost (XGB), and Support Vector Classifier (SVC) machine learning classifiers. The different classification methods were implemented with the help of a python-based library known as Scikit-learn [58]. DT algorithms work on the basis of non-parametric supervised learning models. The major

aim of the classifier is to identify the output instance by learning various decision rules provided in the form of input data [59]. The GNB method is a probabilistic classifier and develops on the Bayes theorem. It was based on the assumption that the consecutive variable of every group follows Gaussian distribution or normal distribution [60]. Random forest is an ensemble-based classifier; which predicts a single tree as a response variable by training the number of decision trees. It also controls the overfitting of the models [61]. LR technique is used to achieve the logistic/logit model, which gives the likelihood of an event to happen. It applies a logistic function to predict the response variable or occurrence of a class [62]. KNN method is an instance-based classifier. It usually collects the instances of the training dataset. Its prediction is based on the maximum number of votes given to a particular class closer to the nearest neighbor data point [63]. XGB classifier uses the scalable tree boosting algorithm, in which an iterative approach is used for the prediction of the final output [64]. SVC developed on the library of support vector machines (libsvm). It usually fits the data points provided as input features and provides the most suitable fit of a hyperplane that categorizes the data into two classes [65].

Cross-validation Techniques

In order to train, test, and evaluate our classification model, we used 5-fold cross-validation and external validation techniques. Several classification and prediction methods use 80:20 splitting of the complete dataset for training and validation purposes in the last few decades [66,67]. In the current study, we have implemented a similar strategy to evaluate our classification model. For each dataset, 80% of data is used for training, and the remaining 20% used for external validation. Further, we apply the 5-fold cross-validation techniques on the training dataset. The training data is equally divided into five sets/folds in which four folds were used for training, and the fifth fold is used for testing the model. This process is iteratively done five times, in which each set is used for testing the model. The final performance is computed by taking the average of each set.

Performance Evaluation Parameters

We used standard evaluation parameters to compute the performance of classification models. Threshold-dependent and independent parameters were used in this study. The performance of the models is calculated by threshold-dependent parameters, such as sensitivity (Sens),

accuracy (Acc), and specificity (Spec). Area Under the Receiver Operating Characteristic (AUROC) curve, the threshold-independent parameter is used to measure the models' performance. AUROC generates a curve by plotting sensitivity against (1-specificity) on various thresholds. Threshold-dependent parameters were computed by using the given equations:

$$Sensitivity = \frac{TP}{TP+FN} \times 100 \quad [2]$$

$$Specificity = \frac{TN}{TN+FP} \times 100 \quad [3]$$

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \times 100 \quad [4]$$

TP =True Positive, FP =False Positive, TN =True Negative, FN =False Negative.

Web server Implementation

We have developed a web server named "B3Pred" (https://webs.iiitd.edu.in/raghava/b3pred/) to identify blood-brain barrier penetrating peptides and non-B3PPs. We used HTML5, JAVA, CSS3, and PHP scripts to develop the front-end and back-end of the webserver. B3Pred server is compatible with all the latest devices like modern devices such as mobile, tablet, iMac, and desktop. It mainly incorporates predict, design, and protein scan modules.

Results

Amino Acid Composition Analysis

Amino Acid Composition analysis shows the percentage composition of specific peptides in the particular dataset as mentioned above. The compositional difference is clearly visible in the graph, which represents the respective dataset's percentage AAC (**Fig. 2**). Arginine is highest in CPPs and B3PPs, which shows that it plays a role in the penetration of peptides into the cells. Tyrosine, an aromatic amino acid, is high in B3PPs as compared to other datasets. Unique amino acids Proline and Glycine are prevalent in B3PPs as a contrast to the other datasets.

Insert Fig. 2 here

Fig. 2: Amino Acid Composition percentage of peptides in three datasets, i.e., B3PPs in blue, CPPs in orange, and random peptides in gray color

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 June 2021 doi:10.20944/preprints202106.0279.v1

Amino Acid Position Analysis

The preferential amino acid position is denoted in **Fig. 3**, which is generated with the help of Two sample logo software. The preferred position of amino acids can be seen in the figure, and it helps us in understanding and designing the B3PPs of research interest. Tyrosine, glycine, and arginine are the most highly preferred amino acids in the first three positions in B3PPs. Two sample logo suggest that Tyrosine, glycine, arginine, and lysine are more preferred throughout the B3PPs 10 amino acid length. Hence these amino acids play a crucial role in the composition and position of amino acids in B3PPs.

Insert Fig. 3 here

Fig. 3: Two-Sample Logo representation of all the three datasets (i.e., Dataset_1, Dataset_2, and Dataset_3), amino acids preferred positions can be seen in the TSL.

B3PPs Prediction Methods on different datasets

B3PPs prediction methods were prepared using various machine learning techniques such as Random Forest (RF), XG Boosting (XGB), Logistic Regression (LR), Support Vector Classifier (SVC), K-Nearest Neighbor (KNN), Gaussian Naive Bayes (GNB), and Decision Tree (DT) on various datasets. The best model was implemented in the webserver and standalone versions. As we created three different datasets for the prediction of B3PPs, we generated 9189 peptide features by using Pfeature. Nine thousand one hundred eighty-nine generated peptide features of each dataset were scrutinized and reduced by an SVC-L1 feature selection technique. The feature selection technique led us to 73 features of Dataset_1, 73 features of Dataset_2, and 145 features of Dataset_3 (Supplementary file S1).

After selecting features on all the datasets, we performed the machine learning techniques using different methods. We analysed, the output on all the datasets and interpreted the results. Dataset_1 consists of 269 B3PPs and 269CPPs as positive and negative datasets, respectively. We analysed that out of selected 73 features, and the RF performed best on 73 features. Performance obtained on the RF method was 85.12% accuracy, 0.92 AUROC on the training dataset, and 84.25% accuracy, 0.89 AUROC on the validation dataset. KNN performed worst on the Dataset_1 i.e., 65.58% accuracy, 0.74 AUROC on training and 50.92% accuracy, 0.64 AUROC on validation dataset (**Table 1)**.

Insert Table 1 here

Table 1. Various Machine learning method's results on Dataset_1

Dataset_2 consists of 269 B3PPs from B3Pdb, and 269 non-B3PPs randomly generated peptides. We performed the machine learning approaches on Dataset_2 and the above-mentioned seven different methods to predict B3PPs. We analysed that RF performed best on selected 73 features. The performance of the good performing methods on Dataset_2 is in Table 2. The RF method performed with 82.09 % accuracy, 0.90 AUROC on the training dataset, and 81.48% accuracy, 0.88 AUROC on the validation dataset (**Table 2**).

Insert Table 2 here

Table 2. Various Machine learning method's results on Dataset_2

The final Dataset, i.e., Dataset_3, consists of 269 B3PPs, and 2690 randomly generated non-B3PPs. We applied the machine learning methods on the Dataset_3 and analysed, the different methods performed on this dataset of selected top 80 features. We analysed that RF performed best among all the methods on the top 80 selected features. The performance of the RF model is 85.25% accuracy, 0.93 AUROC on the training dataset, and 82.93% accuracy, 0.90 AUROC on the validation dataset. It was the highest performing among all the methods on all the datasets, so we have incorporated this RF model in our webserver for the prediction of the B3PPs (**Table 3**).

Insert Table 3 here

Table 3. Various Machine learning method's results on Dataset_3

We also plotted the AUROC curves for the final dataset, i.e., Dataset_3. The best performing method among all the methods were selected for the demonstration of AUROC. We can clearly demark from the AUROC plot that all the methods performed well on the training dataset and validation dataset except GNB and DT **(Fig. 4)**.

Insert Fig. 4 here

Fig. 4: AUROC plot of various machine learning methods on top selected features of Dataset_3. A.) AUROC curve for the training dataset B.) AUROC curve for validation dataset

Webserver and Standalone Software

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 June 2021 doi:10.20944/preprints202106.0279.v1

One of the major objectives of this study is to facilitate the scientific community in discovering B3PPs based drug delivery vehicles that can deliver cargos in brain tissues. Thus, we developed a standalone software as well as a web-based service to assist the researcher in finding new B3PPs or designing efficient B3PPs. Our web server B3Pred has three major modules, namely Predict, Design, and Scan. Predict module of B3pred allow users to predict B3PPs in a set of protein sequences submitted by the user. It allows users to select models developed on any dataset used in this study. The design module of B3pred was developed to discover the most promiscuous B3PPs for a given peptide. This module first generates all possible analogs of a peptide then predicts the score for each analog. It also allows users to short analogs based on their score to select the best analog of a peptide. The protein scan module provides the facility to identity the B3PPs region in the query protein of the user. It allows the user to select the length of the peptide segment to be scanned in the protein sequence submitted by the user. In addition to web-based service, we also developed standalone software for searching B3PPs at a large scale, like searching B3Ps at the genome level.

Insert Fig. 5 here

Fig. 5: Descriptive representation of predict module in B3Pred webserver.

Comparison with the existing method

In order to understand the benefits and drawbacks of the new method, it is crucial to compare the new method with existing methods. However, many methods have been developed in the past to predict the BBB penetrating potential of chemical compounds. Best of our knowledge, recently, BBPpred has been developed to predict B3PPs. BBPpred is trained on 100 B3PPs and 100 non-B3PPs, and the model is tested on only 19 B3PPs and 19 non-B3PPs. On the other hand, B3Pred is trained and tested on three different datasets such as dataset_1 contains 269 B3P peptides and 269 CPPs, dataset_2 comprises 269 B3P peptides, and 269 non-B3P peptides randomly generated using the Swiss-Prot database, and dataset_3 accommodates 269 B3P peptides, and 2690 non-B3P peptides randomly generated using the Swiss-Prot database. In terms of performance, BBPpred achieved maximum AUROC 0.87, whereas B3Pred achieved AUROC 0.92, 0.90, and 0.93 on dataset_1, dataset_2, and dataset_3, respectively. BBPpred only provides the prediction facility; on the other hand, B3Pred provides a prediction, design, and scan facility. In addition, B3Pred is also available as standalone software so that users can run on their local machine on a large scale.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 June 2021 doi:10.20944/preprints202106.0279.v1

Discussion & Conclusion

The Blood-Brain Barrier (BBB) is the natural guard of the brain, which inhibits unwanted molecules to cross the brain tissues [68]. Unfortunately, brain-related or neurological diseases have increased tremendously in the last few decades. In order to treat neurological disorders, such as Alzheimer's disease, Parkinson's disease, neuroinflammation, there is a need for drugs that can be used to treat brain-associated diseases. Due to advancements in technology, researchers are able to discover drugs to treat these disorders in vitro. One of the major hurdles in treating brain-associated disease is delivering drugs in brain tissue, as the blood-brain barrier inhibits these drug molecules from reaching brain tissue [69]. The transportation or delivery of the therapeutic molecules penetrating the barriers of the brain is the bottleneck challenge in treating brain tumors and CNS diseases [70].

Several in silico methods have been developed to predict and improve the delivery of the therapeutic molecules circumventing BBB. Like other therapeutic molecules, blood-brain barrier penetrating peptides (B3PPs) have a significant role in neurological disorders. A study has shown that D-Ala-Peptide T-amide (DAPTA), or peptide T is an antiviral peptide that can cross the blood-brain barrier. Intranasal Peptide T obtained from the envelope protein of the human immunodeficiency virus (HIV). Peptide shows antiviral properties and usually inhibits the chemokine (CCR5) receptors and also acts as B3PPs [71,72]. Researchers explained, AH-D is an amphipathic α-helical BBB penetrating peptide that acts as a therapeutic agent for deadly viruses. It is used as a direct antiviral agent (DAA) to inhibit specific viral proteins. A recent study suggests that potential antiviral AH-D is a target against deadly viruses such as chikungunya virus, Zika, dengue, and yellow fever, with different inhibitory and cytotoxic concentrations [73–76]. These studies provide information that such peptides can be helpful in viral infections, along with if any neurological complications arise due to viruses. These peptides can be used as therapeutic substitutes for antiviral drugs which are unable to cross the brain. This may help in controlling the neurological complications that arose due to Covid-19 [77].

In the present scenario, there is the utmost need to develop an efficient prediction tool that can accurately predict the peptides that are having the property of penetrating through the blood-brain barrier. To facilitate the researchers working in this area, we proposed a method named

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 June 2021 doi:10.20944/preprints202106.0279.v1

B3pred for predicting B3PPs. We have used more than 9000 descriptors to build the prediction model. The RF-based model has achieved the maximum AUROC of 0.93 and 0.90 on training and validation datasets, respectively. We have also developed the free webserver name B3pred and have incorporated various modules such as prediction, design, and scan for B3PPs, to analyze and design the desired B3PPs. We believe that our method would help in the accurate prediction of B3PPs and aid the scientific community working in this area.

Figures and Tables

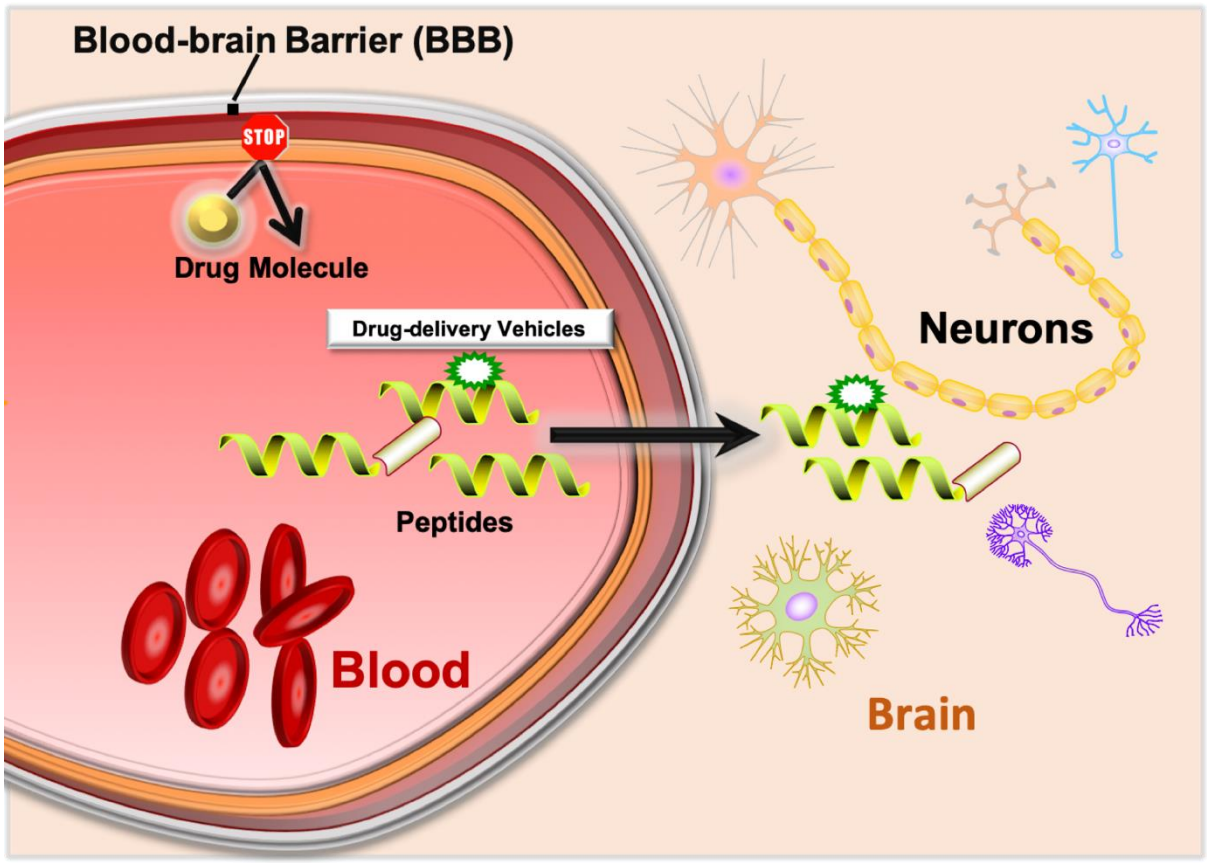
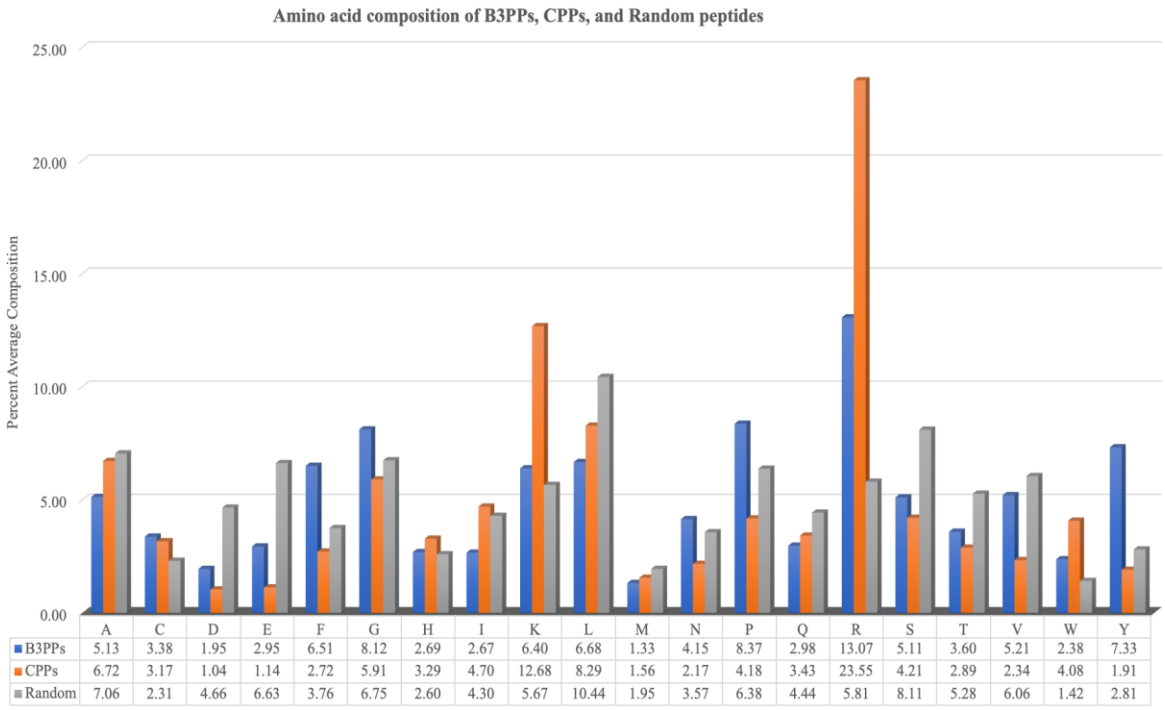


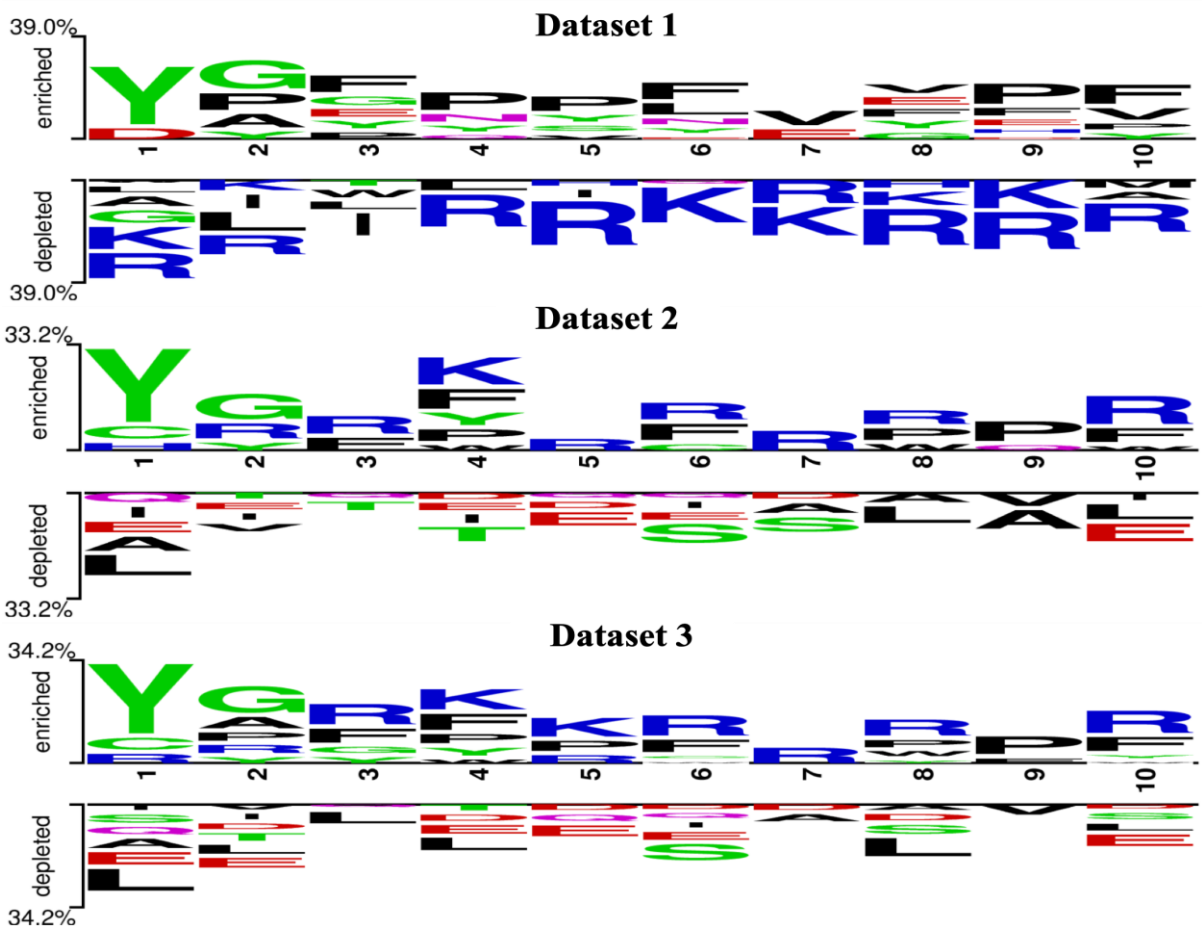
Fig. 1 Representation of the Blood-Brain Barrier and B3PPs to cross into CNS

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 June 2021 doi:10.20944/preprints202106.0279.v1



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
B3PPs	5.13	3.38	1.95	2.95	6.51	8.12	2.69	2.67	6.40	6.68	1.33	4.15	8.37	2.98	13.07	5.11	3.60	5.21	2.38	7.33
CPPs	6.72	3.17	1.04	1.14	2.72	5.91	3.29	4.70	12.68	8.29	1.56	2.17	4.18	3.43	23.55	4.21	2.89	2.34	4.08	1.91
Random	7.06	2.31	4.66	6.63	3.76	6.75	2.60	4.30	5.67	10.44	1.95	3.57	6.38	4.44	5.81	8.11	5.28	6.06	1.42	2.81

Fig. 2: Amino Acid Composition percentage of peptides in three datasets, i.e., B3PPs in blue, CPPs in orange, and random peptides in gray color



Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 June 2021 doi:10.20944/preprints202106.0279.v1

Fig. 3: Two-Sample Logo representation of all the three datasets (i.e., Dataset_1, Dataset_2, and Dataset_3), amino acids preferred positions can be seen in the TSL.

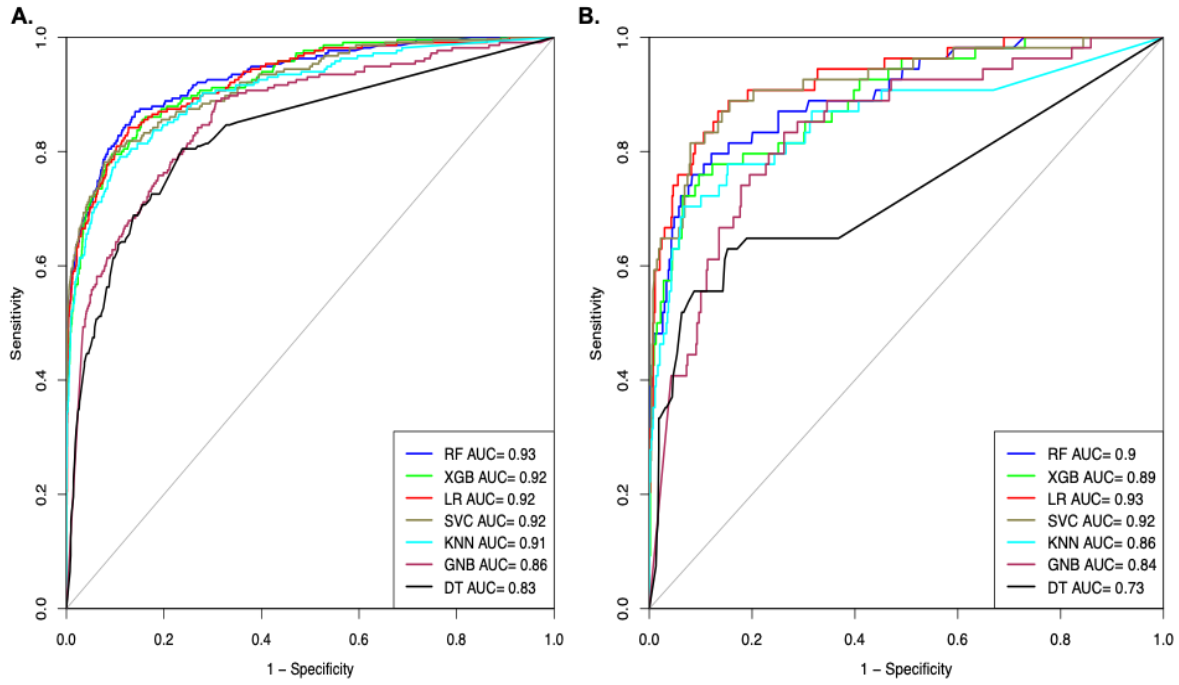


Fig. 4: AUROC plot of various machine learning methods on top selected features of Dataset_3. A.) AUROC curve for the training dataset B.) AUROC curve for validation dataset

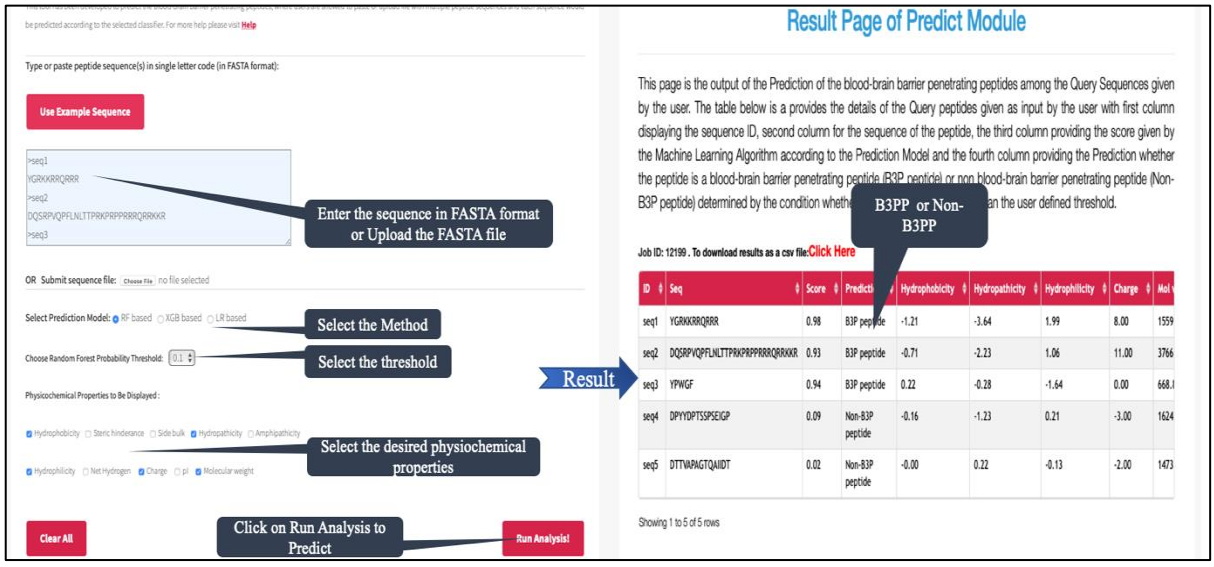


Fig. 5: Descriptive representation of predict module in B3Pred webserver.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 June 2021 doi:10.20944/preprints202106.0279.v1

Methods	Training Dataset_1					Validation Dataset_1				
	Sens	Spec	Acc	AUROC	MCC	Sens	Spec	Acc	AUROC	MCC
RF	86.04	84.18	85.12	0.92	0.70	79.63	88.88	84.25	0.89	0.68
XGB	81.39	82.32	81.86	0.89	0.63	68.51	90.74	79.63	0.87	0.60
LR	82.79	83.25	83.02	0.90	0.66	77.77	83.33	80.55	0.90	0.61
SVC	83.25	82.79	83.02	0.90	0.65	74.07	85.18	79.63	0.89	0.59
KNN	66.51	64.65	65.58	0.74	0.31	38.88	62.96	50.92	0.64	0.10
GNB	84.18	82.32	83.25	0.87	0.66	83.33	79.63	81.48	0.89	0.63
DT	78.14	75.34	76.74	0.82	0.53	62.96	77.77	70.37	0.78	0.41

Table 1. Various Machine learning method's results on Dataset_1

Methods	Training Dataset_2					Validation Dataset_2				
	Sens	Spec	Acc	AUROC	MCC	Sens	Spec	Acc	AUROC	MCC
RF	80.57	84.18	82.09	0.90	0.64	75.92	87.03	81.48	0.88	0.63
XGB	80.46	81.39	80.93	0.88	0.62	79.63	88.89	84.25	0.88	0.68
LR	80.46	86.52	83.48	0.90	0.67	81.48	87.03	84.26	0.91	0.69
SVC	79.07	84.65	81.86	0.88	0.64	74.07	92.59	83.33	0.91	0.67
KNN	50.23	80.93	65.58	0.74	0.32	48.18	77.77	62.93	0.72	0.27

GNB	72.55	87.44	80	0.86	0.61	53.70	94.44	74.07	0.86	0.52
DT	73.02	73.95	73.49	0.79	0.47	74.07	70.37	72.22	0.76	0.44

Table 2. Various Machine learning method's results on Dataset_2

Methods	**Training Dataset_3**					**Validation Dataset_3**				
	Sens	**Spec**	**Acc**	**AUROC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUROC**	**MCC**
RF	86.97	85.08	85.25	0.93	0.51	81.48	83.08	82.93	0.90	0.44
XGB	72.55	93.82	91.88	0.92	0.58	72.22	92.00	90.20	0.892	0.52
LR	80.93	89.73	88.93	0.92	0.54	83.33	89.40	88.85	0.93	0.55
SVC	80.00	84.75	84.32	0.90	0.45	85.18	82.15	82.43	0.90	0.45
KNN	83.72	80.76	81.03	0.88	0.43	79.63	78.44	78.54	0.84	0.37
GNB	80.46	75.74	76.20	0.84	0.35	83.33	72.67	73.65	0.86	0.34
DT	85.11	65.00	66.83	0.82	0.30	64.82	63.20	63.40	0.72	0.20

Table 3. Various Machine learning method's results on Dataset_3

Availability of Data and Materials

Datasets available at https://webs.iiitd.edu.in/raghava/b3pred/download.php.

Author's contribution

VK and SP collected and processed the datasets. VK and SP created the prediction models. VK, SP, AD, NS, and GPSR analysed and interpreted the results. VK and SP developed the webserver. VK, SP, AD, and GPSR prepared the manuscript. GPSR coordinated the project. All authors have read and approved the final manuscript.

Conflict of Interest

There is no Conflict of Interest.

References

1. Abbott NJ, Rönnbäck L, Hansson E. Astrocyte-endothelial interactions at the blood-brain barrier. Nat. Rev. Neurosci. 2006; 7:41–53
2. Kniesel U, Wolburg H. Tight junctions of the blood-brain barrier. Cell. Mol. Neurobiol. 2000; 20:57–76
3. FENSTERMACHER J, GROSS P, SPOSITO N, et al. Structural and Functional Variations in Capillary Systems within the brain. Ann. N. Y. Acad. Sci. 1988; 529:21–30
4. Rhea EM, Banks WA. Role of the Blood-Brain Barrier in Central Nervous System Insulin Resistance. Front. Neurosci. 2019; 13:
5. Muoio V, Persson PB, Sendeski MM. The neurovascular unit - concept review. Acta Physiol. 2014; 210:790–798
6. Oldendorf WH. Brain uptake of radiolabeled amino acids, amines, and hexoses after arterial injection. Am. J. Physiol. 1971; 221:1629–1639
7. Tietz S, Engelhardt B. Brain barriers: Crosstalk between complex tight junctions and adherens junctions. J. Cell Biol. 2015; 209:493–506
8. Pardridge WM. Blood-brain barrier delivery. Drug Discov. Today 2007; 12:54–61
9. Islam Y, Leach AG, Smith J, et al. Peptide based drug delivery systems to the brain. Nano Express 2020; 1:012002
10. Banks WA. Peptides and the blood-brain barrier. Peptides 2015; 72:16–19
11. Aileen Funke S, Willbold D. Peptides for Therapy and Diagnosis of Alzheimer's Disease. Curr. Pharm. Des. 2012; 18:755–767
12. Baig MH, Ahmad K, Saeed M, et al. Peptide based therapeutics and their use for the treatment of neurodegenerative and other diseases. Biomed. Pharmacother. 2018; 103:574–581
13. Raucher D. Tumor targeting peptides: novel therapeutic strategies in glioblastoma. Curr. Opin. Pharmacol. 2019; 47:14–19
14. Oller-Salvia B, Sá Nchez-Navarro M, Giralt E, et al. Chem Soc Rev Chemical Society Reviews Blood-brain barrier shuttle peptides: an emerging paradigm for brain delivery. This J. is Cite this Chem. Soc. Rev 2016; 45:4690
15. Wu LP, Ahmadvand D, Su J, et al. Crossing the blood-brain-barrier with nanoligand drug carriers self-assembled from a phage display peptide. Nat. Commun. 2019; 10:
16. Nosrati H, Tarantash M, Bochani S, et al. Glutathione (GSH) Peptide Conjugated Magnetic Nanoparticles As Blood-Brain Barrier Shuttle for MRI-Monitored Brain Delivery of Paclitaxel. ACS Biomater. Sci. Eng. 2019; 5:1677–1685
17. Solbrig M V., Koob GF. Epilepsy, CNS viral injury and dynorphin. Trends Pharmacol. Sci. 2004; 25:98–104
18. Balasubramaniam A. Clinical potentials of neuropeptide Y family of hormones. Am. J. Surg. 2002; 183:430–434
19. Claes SJ. Corticotropin-releasing hormone (CRH) in psychiatry: From stress to psychopathology. Ann. Med. 2004; 36:50–61
20. Ströhle A, Holsboer F. Stress Responsive Neurohormones in Depression and Anxiety. Pharmacopsychiatry 2003; 36:
21. Li C, Wu X, Liu S, et al. Roles of Neuropeptide Y in Neurodegenerative and Neuroimmune Diseases. Front. Neurosci. 2019; 13:
22. Dwibhashyam VSNM, Nagappa A. Strategies for enhanced drug delivery to the central nervous system. Indian J. Pharm. Sci. 2008; 70:145–153
23. Reese TS, Karnovsky MJ. Fine structural localization of a blood-brain barrier to exogenous peroxidase. J. Cell Biol. 1967; 34:207–217

24. Kapoor P, Singh H, Gautam A, et al. TumorHoPe: a database of tumor homing peptides. PLoS One 2012; 7:e35187
25. Gautam A, Singh H, Tyagi A, et al. CPPsite: a curated database of cell penetrating peptides. Database (Oxford). 2012; 2012:bas015
26. Sharma A, Kapoor P, Gautam A, et al. Computational approach for designing tumor homing peptides. Sci. Rep. 2013; 3:1607
27. Gautam A, Sharma M, Vir P, et al. Identification and characterization of novel protein-derived arginine-rich cell-penetrating peptides. Eur. J. Pharm. Biopharm. 2015; 89:93–106
28. Shergalis A, Bankhead A, Luesakul U, et al. Current challenges and opportunities in treating glioblastomas. Pharmacol. Rev. 2018; 70:412–445
29. Stalmans S, Wynendaele E, Bracke N, et al. Chemical-Functional Diversity in Cell-Penetrating Peptides. PLoS One 2013; 8:
30. Stalmans S, Bracke N, Wynendaele E, et al. Cell-penetrating peptides selectively cross the blood-brain barrier in vivo. PLoS One 2015; 10:
31. Yamano S, Dai J, Hanatani S, et al. Long-term efficient gene delivery using polyethylenimine with modified Tat peptide. Biomaterials 2014; 35:1705–1715
32. Huwyler J, Wu D, Pardridge WM. Brain drug delivery of small molecules using immunoliposomes. Proc. Natl. Acad. Sci. U. S. A. 1996; 93:14164–14169
33. Knight A, Carvajal J, Schneider H, et al. Non-viral neuronal gene delivery mediated by the H(C) fragment of tetanus toxin. Eur. J. Biochem. 1999; 259:762–769
34. El-Andaloussi S, Holm T, Langel U. Cell-Penetrating Peptides: Mechanisms and Applications. Curr. Pharm. Des. 2005; 11:3597–3611
35. Milletti F. Cell-penetrating peptides: Classes, origin, and current landscape. Drug Discov. Today 2012; 17:850–860
36. Stewart KM, Horton KL, Kelley SO. Cell-penetrating peptides as delivery vehicles for biology and medicine. Org. Biomol. Chem. 2008; 6:2242–2255
37. Mueller J, Kretzschmar I, Volkmer R, et al. Comparison of cellular uptake using 22 CPPs in 4 different cell lines. Bioconjug. Chem. 2008; 19:2363–2374
38. Meade AJ, Meloni BP, Mastaglia FL, et al. The application of cell penetrating peptides for the delivery of neuroprotective peptides/proteins in experimental cerebral ischaemia studies. J. Exp. Stroke Transl. Med. 2009; 02:22–40
39. Mathur D, Prakash S, Anand P, et al. PEPlife: A Repository of the Half-life of Peptides. Sci. Rep. 2016; 6:36617
40. Gautam A, Chaudhary K, Kumar R, et al. In silico approaches for designing highly effective cell penetrating peptides. J. Transl. Med. 2013; 11:74
41. Wei L, Tang J, Zou Q. SkipCPP-Pred: An improved and promising sequence-based predictor for predicting cell-penetrating peptides. BMC Genomics 2017; 18:
42. Wei L, Xing P, Su R, et al. CPPred-RF: A Sequence-based Predictor for Identifying Cell-Penetrating Peptides and Their Uptake Efficiency. J. Proteome Res. 2017; 16:2044–2053
43. Pandey P, Patel V, George N V, et al. KELM-CPPpred: Kernel Extreme Learning Machine Based Prediction Model for Cell-Penetrating Peptides. J. Proteome Res. 2018; 17:3214–3222
44. Kumar V, Agrawal P, Kumar R, et al. Prediction of cell-penetrating potential of modified peptides containing natural and chemically modified residues. Front. Microbiol. 2018;
45. Qiang X, Zhou C, Ye X, et al. CPPred-FL: a sequence-based predictor for large-scale identification of cell-penetrating peptides by feature representation learning. Brief. Bioinform. 2018; 21:11–23
46. Shaker B, Yu M-S, Song JS, et al. LightBBB: computational prediction model of blood–brain-barrier penetration based on LightGBM. Bioinformatics 2021; 37:1135–1139
47. Carpenter TS, Kirshner DA, Lau EY, et al. A Method to Predict Blood-Brain Barrier

Permeability of Drug-Like Compounds Using Molecular Dynamics Simulations. Biophys. J. 2014; 107:630–641
48. Mensch J, Oyarzabal J, Mackie C, et al. In vivo, in vitro and in silico methods for small molecule transfer across the BBB. J. Pharm. Sci. 2009; 98:4429–4468
49. Dai R, Zhang W, Tang W, et al. BBPpred: Sequence-Based Prediction of Blood-Brain Barrier Peptides with Feature Representation Learning and Logistic Regression. J. Chem. Inf. Model. 2021; 61:525–534
50. Agrawal P, Bhalla S, Usmani SS, et al. CPPsite 2.0: a repository of experimentally validated cell-penetrating peptides. Nucleic Acids Res. 2016; 44:D1098-103
51. Boutet E, Lieberherr D, Tognolli M, et al. Uniprotkb/swiss-prot, the manually annotated section of the uniprot knowledgebase: How to use the entry view. Methods Mol. Biol. 2016; 1374:23–54
52. Agrawal P, Raghava GPS. Prediction of Antimicrobial Potential of a Chemically Modified Peptide From Its Tertiary Structure. Front. Microbiol. 2018; 9:2551
53. Vacic V, Iakoucheva LM, Radivojac P. Two Sample Logo: A graphical representation of the differences between two sets of sequence alignments. Bioinformatics 2006; 22:1536–1537
54. Pande A, Patiyal S, Lathwal A, et al. Computing wide range of protein/peptide features from their sequence and structure. bioRxiv 2019; 599126
55. Tang J, Alelyani S, Liu H. Feature selection for classification: A review. Data Classif. Algorithms Appl. 2014; 37–64
56. Chang K-W, Hsieh C-J, Lin C-J. LIBLINEAR: A Library for Large Linear Classification Rong-En Fan Xiang-Rui Wang. J. Mach. Learn. Res. 2008; 9:
57. Ke G, Meng Q, Finley T, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree.
58. Pedregosa FABIANPEDREGOSA F, Michel V, Grisel OLIVIERGRISEL O, et al. Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot. J. Mach. Learn. Res. 2011; 12:
59. Webb GI. Decision Tree Grafting From the All-Tests-But-One Partition. 2000;
60. Zhang H. Exploring conditions for the optimality of naïve bayes. Int. J. Pattern Recognit. Artif. Intell. 2005; 19:183–198
61. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Mach. Learn. 2006; 63:3–42
62. Tolles J, Meurer WJ. Logistic regression: Relating patient characteristics to outcomes. JAMA - J. Am. Med. Assoc. 2016; 316:533–534
63. Mucherino A, Papajorgji PJ, Pardalos PM. Data Mining in Agriculture. 2009; 34:
64. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.
65. Chang C-C, Lin C-J. LIBSVM: A Library for Support Vector Machines. 2001;
66. Kumar V, Kumar R, Agrawal P, et al. A Method for Predicting Hemolytic Potency of Chemically Modified Peptides From Its Structure. Front. Pharmacol. 2020; 11:
67. Dhall A, Patiyal S, Sharma N, et al. Computer-aided prediction and design of IL-6 inducing peptides: IL-6 plays a crucial role in COVID-19. Brief. Bioinform. 2020;
68. Thangudu S, Cheng F-Y, Su C-H. Advancements in the Blood-Brain Barrier Penetrating Nanoplatforms for Brain Related Disease Diagnostics and Therapeutic Applications. Polymers (Basel). 2020; 12:1–23
69. He Q, Liu J, Liang J, et al. Towards Improvements for Penetrating the Blood-Brain Barrier-Recent Progress from a Material and Pharmaceutical Perspective. Cells 2018; 7:24
70. Banks WA. From blood-brain barrier to blood-brain interface: New opportunities for

 doi:10.20944/preprints202106.0279.v1

CNS drug delivery. Nat. Rev. Drug Discov. 2016; 15:275–292
71. Polianova MT, Ruscetti FW, Pert CB, et al. Antiviral and immunological benefits in HIV patients receiving intranasal peptide T (DAPTA). Peptides 2003; 24:1093–1098
72. Barrera CM, Kastin AJ, Banks WA. D-[Ala1]-peptide T-Amide is transported from blood to brain by a saturable system. Brain Res. Bull. 1987; 19:629–633
73. Jackman JA, Costa V V., Park S, et al. Therapeutic treatment of Zika virus infection using a brain-penetrating antiviral peptide. Nat. Mater. 2018; 17:971–977
74. Cho NJ, Dvory-Sobol H, Xiong A, et al. Mechanism of an amphipathic α-helical peptide's antiviral activity involves size-dependent virus particle lysis. ACS Chem. Biol. 2009; 4:1061–1067
75. Zou J, Shi PY. Targeting vesicle size. Nat. Mater. 2018; 17:955–956
76. Boldescu V, Behnam MAM, Vasilakis N, et al. Broad-spectrum agents for flaviviral infections: Dengue, Zika and beyond. Nat. Rev. Drug Discov. 2017; 16:565–586
77. Mao XY, Jin WL. The COVID-19 Pandemic: Consideration for Brain Infection. Neuroscience 2020; 437:130–131