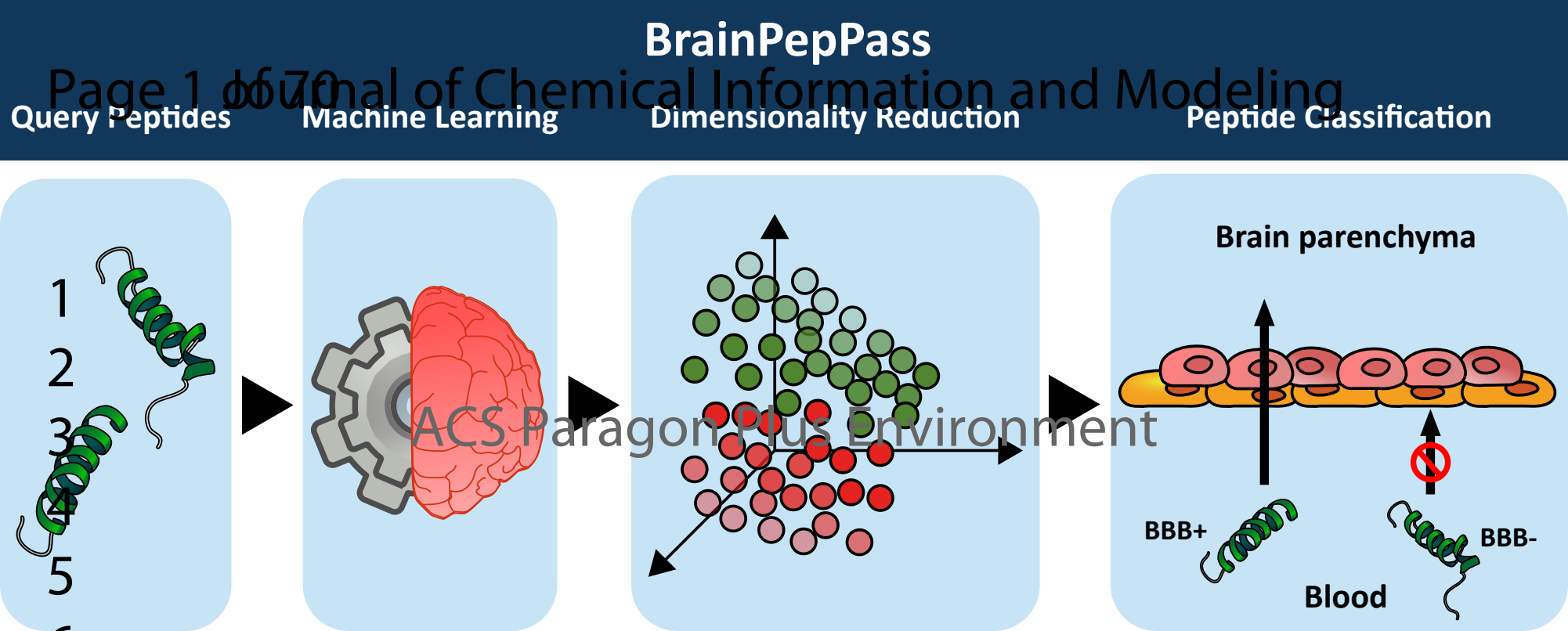
Journal of Chemical Information and Modeling



BrainPepPass: A Framework Based on Supervised Dimensionality Reduction for Predicting Blood-Brain Barrier-Penetrating Peptides

Ewerton Cristhian Lima de Oliveira[1,5], Hannah Hirmz[2*], Evelien Wynendaele[2], Juliana Auzier Seixas Feio[1], Igor Matheus Souza Moreira[1], Kauê Santana da Costa[3, ‡], Anderson H. Lima[4], Bart De Spiegeleer[2, ‡], Claudomiro de Souza de Sales Júnior[1, ‡]*

[1]Laboratório de Inteligência Computacional e Pesquisa Operacional, Campos Belém, Instituto de Tecnologia, Universidade Federal do Pará, 66075-110, Belém, Pará, Brasil

[2] Drug Quality and Registration (DruQuaR) Group, Faculty of Pharmaceutical Sciences, Ghent University, Ottergemsesteenweg 460, B-9000 Ghent, Belgium.

[3]Laboratório de Simulação Computacional, Campos Marechal Rondon, Instituto de Biodiversidade, Universidade Federal do Oeste do Pará, 68040-255, Santarém, Pará, Brasil.

[4]Laboratório de Planejamento e Desenvolvimento de Fármacos, Instituto de Ciências Exatas e Naturais, Universidade Federal do Pará, 66075-110, Belém, Pará, Brasil.

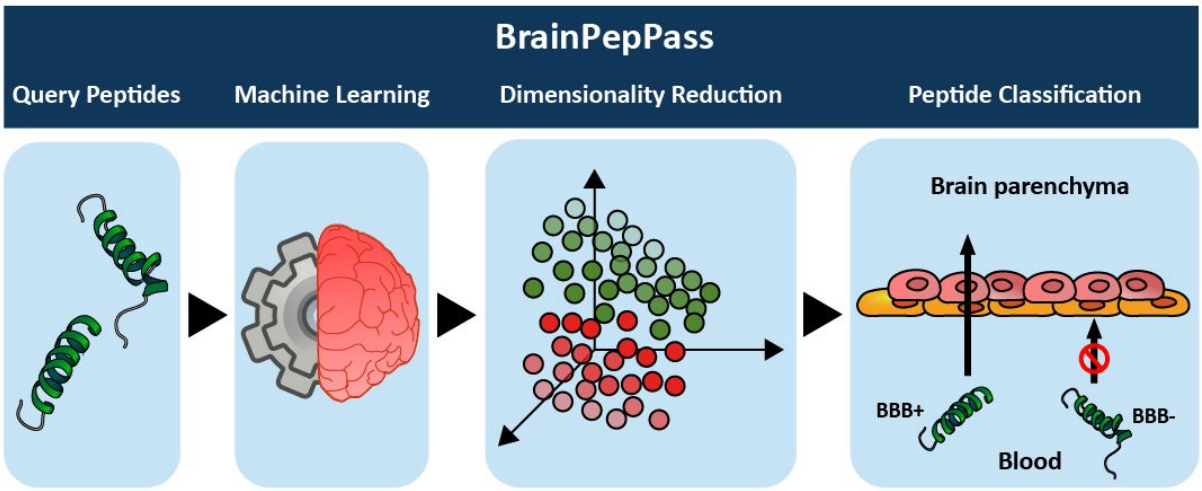
[5]Instituto Tecnológico Vale, 66055-090, Belém, Pará, Brasil.

* Authors contributed equally to this work.

‡ Corresponding authors

This paper is dedicated to the memory of a great biomedical scientist, Professor Abba Jeremiah Kastin, who passed away last year.

GRAPHICAL ABSTRACT



ABSTRACT

Peptides that pass through the blood-brain barrier (BBB) are not only implicated in brain-related pathologies but are also promising therapeutic tools for treating brain diseases, e.g., as shuttles carrying active medicines across the BBB. Computational prediction of BBB-penetrating peptides (B3PPs) has emerged as an interesting approach because of its ability to screen large peptide libraries in a cost-effective manner. In this study, we present BrainPepPass, a machine learning (ML) framework that utilizes supervised manifold dimensionality reduction and extreme gradient boosting (XGB) algorithms to predict natural and chemically modified B3PPs. The results indicate that the proposed tool outperforms other classifiers, with average accuracies exceeding 94% and 98% in 10-fold cross-validation and

leave-one-out cross-validation (LOOCV), respectively. In addition, accuracy values ranging from 45% to 97.05% were achieved in the independent tests. The BrainPepPass tool is available in a public repository for academic use (https://github.com/ewerton-cristhian/BrainPepPass).

INTRODUCTION

Blood brain-penetrating peptides are oligopeptide chains that can naturally traverse the blood-brain barrier (BBB); thus, for example facilitating the enhanced uptake of molecular cargoes in a non-selective way. Hence, they are also called BBB shuttle peptides.[1–4] Until the 1970s, peptides were believed not to cross the BBB. The late Abba J Kastin († in 2022) was the first researcher who experimentally tried to refute this assumption. After injecting radiolabeled peptides such as ^{125}I-Met-enkephalin and ^{3}H-α-melanocyte-stimulating hormone into the carotid artery of mice, Kastin and colleagues observed radioactivity in different brain regions, providing the first indications that certain endogenous peptides cross the BBB.[5–7] William Banks continued and expanded this research, becoming a protagonist in the field of BBB permeability of peptides. Their research shed light on the function of these endogenous peptides as they showed that in crossing the BBB, peptides act as informational molecules that

inform the brain of peripheral events. Conversely, peptides crossing from the brain to the blood can deliver information in the brain-to-blood direction.[8] Not only physiological functions but also pathologies are attributed to the BBB passage of certain peptides. For instance, BBB dysfunction results in amyloid-β disposition in the brain by preventing its normal transport through the BBB. Amyloid plaques formed by amyloid-β aggregation are considered pathological triggers of Alzheimer's disease.[9] Another example is the transport of insulin through the BBB, which is decreased in obese people[10,11] but seems to be increased in people with diabetes mellitus.[12,13] BBB-penetrating peptides (B3PPs) are being explored in drug development as potential shuttle molecules capable of transporting bio-active drugs across the BBB. In addition, some B3PPs may serve as cell-penetrating peptides.[14] Peptides, including the B3PPs, show low immunogenicity and toxicity, and are amenable to chemical synthesis, offering a plethora of possibilities for functional modifications and improvements. Therefore, B3PPs have opened up new therapeutic and diagnostic horizons.[2,4,15]

Determining whether and to what extent peptides can cross the BBB is a challenge that requires the development of appropriate *in vitro* and *in vivo* techniques to address the technical difficulties in studying these molecules. Various experimental methods have been utilized to determine the permeability of peptides across the BBB, including static *in vitro* models

encompassing transwell monoculture models, co-culture models, and triple-cell co-culture models. These are straightforward and inexpensive methods that do not capture the physiological complexity involved in BBB permeation. More advanced *in vitro* models such as the blood-brain barrier specific parallel artificial membrane permeability (BBB-PAMPA), bovine brain microvessel endothelial cells (BBMEC), dynamic *in vitro* models, microfluidic models, and induced pluripotent stem cells (iPSC)-based models, have been developed to more closely mimic the *in vivo* situation.[16–19] However, these models are expensive and involve complex and rigid procedures, some of which are not well established.[18,19] Finally, i*n vivo* experimental methods include the brain uptake index (BUI), multiple time regression (MTR) analysis or Gjedde-Patlak plot, in situ brain perfusion, brain microdialysis, and quantitative radiography. These experimental methods involve more complex, cost- and time-consuming, and labor-intensive techniques when compared to computational tools, but provide the most complete and detailed quantitative information.[18,20–22]

With the development of artificial intelligence technology, machine learning (ML) models have been applied in many biochemistry research fields, including protein and protein-like molecule analysis[23–27]. For instance, Bao et al. developed several tools covering a wide range of applications, such as Golgi_DF, which classifies Golgi proteins using deep forest

algorithms,[28] and Phage_UniR_LGBM, which classifies phage virion proteins using UniRep features and the LightGBM algorithm.[24]

Information on BBB permeability is also often difficult to interpret, because of the multitude of research methods used. Their corresponding output responses ensure that BBB permeability information is not always straightforward to compare, especially in the absence of generally agreed controls such as BSA as a negative control and dermorphin as a positive control. To circumvent this problem and allow a direct comparison of BBB influx results, Stalmans et al. introduced a classification method and unified the response of BBB influx transport. The results of different BBB influx response types, which quantitatively express brain influx, were classified into five classes of BBB influx magnitude based on the distribution of the results for individual response types. This classification can be immediately applied to new BBB influx results of peptides and allows the direct comparison and ranking of peptides independent of the response type.[29]

Owing to expensive, time-consuming, and labor-intensive experimental methods, there is an imminent need for efficient *in silico* methods to estimate the BBB permeability of peptides. Several computational methods for estimating the BBB permeability of small molecules (excluding peptides) have already been developed.[30–37] However, for estimating the BBB

permeability of peptides, *in silico* methods have only been sparsely investigated.[37] Dai et al. presented a sequence-based prediction approach to identify whether a peptide can penetrate the BBB. Using a benchmark dataset, a feature representation learning strategy was designed to characterize sequence-based features from a wide variety of feature descriptors[38]. Subsequently, a three-step feature-selection method was adopted to filter irrelevant and redundant features, resulting in seven optimal features. Based on the optimal features, a predictive model was developed using logistic regression (LR).[37] Zou employed physicochemical properties of amino acids to represent peptide sequences, and the maximal information coefficient (MIC) and Pearson's correlation coefficient (PCC) were used to extract useful information from them. A similarity network fusion algorithm was utilized to integrate these two different types of features, followed by the Fisher algorithm to select the discriminative features. Finally, these selected features were input into support vector machine (SVM) to distinguish B3PPs from non-B3PPs.[39] However, despite the valuable information provided by these *in silico* methods for predicting peptide penetration of the BBB, the lack of computational tools to predict this pharmacokinetic property for both natural and chemically modified peptides hampered the efficient exploration of biotechnological and pharmaceutical applications of these molecules against several brain diseases.

In this article, we describe BrainPepPass, a novel ML-based framework dedicated to predicting not only whether natural peptides can cross or not the BBB, but also whether chemically modified peptides have this property. To the best of our knowledge, BrainPepPass is the first tool in this research field that employs a supervised manifold dimensionality reduction algorithm in the preprocessing stage, in combination with extreme gradient boosting (XGB) models. The recently extended Brainpeps database was used as the most up-to-date and complete data source for this study. Moreover, we investigated how distinct groups of molecular descriptors, including physicochemical and structural properties, correlate with the BBB permeability of peptides. In addition, we have provided a repository with the tool developed in this study to predict the BBB permeability of peptides.

MATERIAL AND METHODS

Datasets

Brainpeps consolidates extensive information related to peptides that interact with and penetrate the BBB. The database contains pertinent details of peptides, such as their nomenclature, primary structure, bibliographical references, as well as pharmacokinetic

indicators. The molecules included in this database were subjected to experimental evaluation, where their BBB penetration characteristics were assessed using a diverse range of methods.[40]

A total of 328 peptides from Brainpeps were extracted in the MOL format, which was essential for our study because this file format encodes the chemical modifications in the peptide's structure, in addition to encoding the cyclic peptides present in this database. The peptides were classified into two categories: blood-brain barrier permeable (BBB+) or nonpermeable (BBB-). This designation was based on experimental indicators used to evaluate the brain penetration levels of the peptides. The six indicators employed in this study included the unidirectional influx constant (K_{in}), measured by multiple time regression (MTR) as well as by *in situ* brain perfusion methods; BBB permeability (P), both *in vitro* and *in vivo*; endothelial permeability (P_e) (measured using the parallel artificial membrane permeability assay [PAMPA]); and apparent permeability coefficient (P_{app}).

The parameter K_{in} or unidirectional influx constant is an indicator that characterizes the steady-state unidirectional influx transfer of peptides from the bloodstream to the brain after a single passage, and is measured in mL/(g.min). This indicator is obtained by performing either multiple time regression (MTR) or an *in situ* brain perfusion experiment. After intravenously injecting with a radiolabeled compound, the brain and plasma or perfusate concentrations is

measured at several time points, which allows the construction of a concentration-time profile, where the slope of this linear regression measures K_{in}.[41] This *in vivo* method has been utilized to classify the penetration of specific peptides such as conotoxin cVc1.1,[42] or somatropin-derived or modified peptides,[43] and quorum-sensing peptides such as PapRIV.[44]

Conversely, the permeability indicator P, which can be measured using both *in vitro* and *in vivo* techniques, expresses the rate at which a peptide moves from the blood to the brain in units of distance per time, usually cm/s.[29] $P_{in\ vitro}$ data is acquired using the brain microvessel endothelial cell (BMEC) culture model. In this technique, bovine, porcine, mouse, rat, or human BMEC form a monolayer on a rat-tail collagen-coated filter or a microporous membrane placed in a diffusion apparatus consisting of a donor and receptor chamber. The test peptide is placed in the donor compartment and the amount of peptide is measured in samples periodically taken from the acceptor compartment. The amount of test peptide in the acceptor compartment can then be plotted as a function of time to calculate $P_{in\ vitro,}$ which is dependent on the initial concentration of the test peptide in the donor compartment as well as the membrane surface area. $P_{in\ vivo}$ is measured by performing *in situ* brain perfusion experiments in animal models, such as rats.[43,45]

Two other permeability indicators, related to P, determined using slightly different methods and calculations are P_e and P_{app}. P_e evaluates the *in vitro* ability of a molecule to penetrate the endothelial cell layer, which represents the primary barrier to the entry of substances into the brain, and is measured in 10^{-6} cm/s. Di et al. proposed a classification system for molecules based on their P_e values, where values greater than 4.10^{-6} cm/s indicate high penetration, values less than 2.10^{-6} cm/s indicate low penetration, and values between 2.10^{-6} and 4.10^{-6} cm/s indicate uncertain permeation.[46] These ranges were derived empirically using parallel artificial membrane permeation assays (PAMPA). This technique utilizes a porcine polar brain lipid artificial membrane between the donor and acceptor compartments to predict the blood-brain barrier permeation of molecular compounds, including peptides. This technique, in its different variants, has been used to classify the permeability of various compounds in the brain, including 3-hydroxy-2-pyridinealdoxime,[47] furosemide, ranitidine, donepezil, tacrine,[48] platyphyllenone, alnusone,[49] gingerol, and shogaol derivatives.[50]

The apparent permeability coefficient (P_{app}) is a similar indicator used to evaluate the *in-vitro* ability of a molecule to traverse a cell-barrier, such as the blood-brain barrier, and is expressed in units of cm/s. *In vitro* models of the BBB, such as monolayers of endothelial cells, are commonly employed to evaluate the permeability of compounds using P_{app}.[51] In a similar

manner to P_e, a classification system for molecules based on their P_{app} values was proposed by Yoon et al., where values greater than 20.10^{-6} cm/s are considered indicative of high permeability, while values lower than 2.10^{-6} cm/s suggest low permeability.[52] Each of these BBB influx response describes the BBB influx from a different viewpoint using different techniques, thereby providing different information.

As previously discussed, studies investigating the entry of peptides and other compounds into the BBB have employed established limits on physicochemical indicators to determine their permeability. Specifically, we propose to categorize peptides as belonging to either the BBB+ or BBB- class based on the limits described for each indicator. For P_e and P_{app}, high penetration was designated as belonging to the BBB+ class, whereas low penetration was classified as BBB-. Peptides with penetration rates between low and high were classified according to the proximity of their respective values to one of the two thresholds (high or low). For example, a peptide with a penetration level numerically closer to a low threshold was classified as BBB-. In terms of K_{in} and P, we dichotomously classified peptides based on the five groups proposed by Stalmans et al. to evaluate the influx of peptides across the BBB.[29] Peptides belonging to the very low and low influx categories were classified as BBB- class, while those in the medium, high, and very high influx categories were classified as BBB+. The

proposed classification system is expected to facilitate a more direct understanding of BBB permeability and contribute to the development of more effective permeability prediction strategies.

The implementation of the proposed criteria for classifying peptides based on their permeability in the BBB using physicochemical indicators in our study yielded a database containing 231 BBB+ and 97 BBB- peptides. However, the database was unbalanced in terms of the number of peptides belonging to each class, necessitating the division of the data into three balanced datasets to avoid issues associated with overfitting. Each dataset comprised the same 97 BBB- peptides and 97 randomly sampled BBB+ peptides.

Of the 231 peptides classified as BBB+ in the complete database, 10 were randomly selected to constitute the test sample of permeable molecules (TSP), whereas the remaining 221 peptides were used to compose the training samples of permeable molecules (TRP). The BBB+ peptides in Dataset 1 were composed of 87 molecules randomly sampled from TRP (TRP-1) and 10 peptides from TSP, totaling 97 BBB+ peptides. Similarly, the BBB+ peptides in Dataset 2 were a combination of distinct 87 molecules randomly sampled from the TRP (TRP-2) and 10 peptides from the TSP. The training samples of BBB+ peptides in Dataset 3 (TRP-3) were composed of the remaining 57 samples from TRP with 15 randomly selected peptides from

TRP-1 and 15 peptides randomly selected from TRP-2, while the test samples include the same molecules from TSP, totaling 97 BBB+ peptides (see Figure 1).

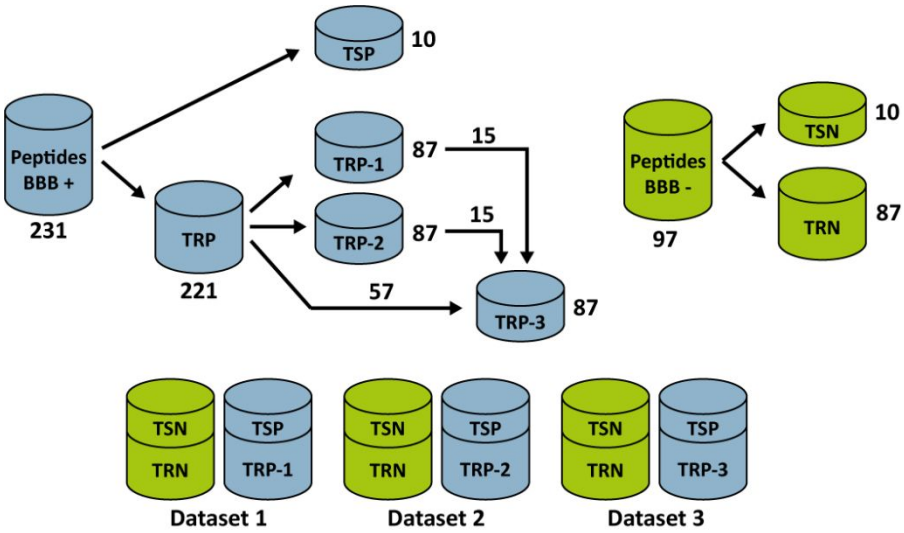


Figure 1. Construction of the three balanced datasets. TRP-[1,2,3]: training samples of BBB+ peptides, TSP: test samples of BBB+ peptides, TRN: training samples of BBB- peptides, TSN: test samples of BBB- peptides.

Overall, each of the three datasets comprised a balanced number of samples from both classes, with 97 BBB+ and 97 BBB- peptides. In each dataset, 174 peptides (87 BBB+ and 87 BBB-) were dedicated to training, while 20 peptides (10 BBB+ and 10 BBB-) were used for testing. This approach ensured that each dataset could be utilized for both training and testing

and that the models developed using these datasets were adequately validated. Supplementary Table S1 provides information regarding the peptides used in each dataset.

Molecular Properties

In this study, we investigated the permeability of peptides across the blood-brain barrier by analyzing a set of molecular properties. The molecular properties were grouped into four distinct feature compositions (FCs). The first feature composition (FC-1) comprised several key descriptors including molecular weight (MW), calculated water-octanol partition coefficient (LogP), calculated octanol-water distribution coefficient (LogD) at pH 7.4, topological polar surface area (TPSA), number of hydrogen bond acceptors (HBA), donors (HBD), nitrogen count (nN), oxygen count (nO), and nitrogen plus oxygen count (nN+nO). Previous studies have highlighted the importance of these descriptors in filtering molecules that are likely to reach the central nervous system (CNS).[1,53–55] Furthermore, some of these descriptors have also been linked to the oral bioavailability of compounds, as proposed by Lipinski's rule of five and other related studies on bioavailability and biomembrane permeability.[56–59]

The second feature composition (FC-2) comprised Mordred's molecular descriptors, which consist of a combination of structural and physicochemical descriptors. Mordred is a Python library for molecular descriptor calculations that encompasses 2D, 3D, constitutional, and electronic descriptors.[60] The 749 descriptors in this FC were extracted from the 231 molecules using this package, after filtering out molecular properties with missing, non-numeric, or non-Boolean values.

The third feature composition (FC-3) was constructed by selecting the ten best-correlated molecular descriptors from FC-2 using Kendall's correlation coefficient. The fourth feature composition (FC-4) was obtained by combining FC-1 and FC-3. Supplementary Tables S2, S3, and S4 provide information on the molecular descriptors of FC-1, FC-2, and FC-3.

To calculate the FC-1 descriptors, we utilized the RDKit package in Python to extract the properties from the peptides, except for LogD, which was determined for each molecule using the Instant JChem software. The other descriptors were calculated using the Mordred package.

Proposed Machine Learning Framework

The ML-based framework proposed in this study for predicting B3PPs, BrainPepPass, is a generic architecture, comprising two stages: dimensionality reduction pattern learning (DRPL) and classification. The DRPL step involves projecting a high-dimensional dataset of molecular descriptors onto a three-dimensional (3D) space, with the dual objective of facilitating low-dimensional visualization of peptides and enabling the clustering of molecules based on their BBB+ or BBB- class labels. To this end, we employed a supervised Laplacian eigenmaps (sLE) algorithm, which has been demonstrated to be effective in reducing high-dimensional data with class labels.[61] However, the original sLE algorithm, like t-SNE, lacks the capacity for independent dataset dimensionality reduction, which renders it unsuitable for BrainPepPass prediction of the permeability of new peptides. To overcome this limitation, we propose using an XGB regression (XGBr) algorithm.

As shown in Figure 2a, the DRPL stage entails the dimensionality reduction of a high-dimensional dataset consisting of molecular descriptors by FC to a 3D representation. Specifically, the same original n-dimensional data was used as input, and their 3D projection was used as the target to train an XGBr algorithm to learn the DR pattern performed by sLE. To select the optimal hyperparameters for the XGBr, we performed a grid search with a predetermined range of values and utilized a 10-fold cross-validation technique to compute the

average accuracy metric. Supplementary Table S5 presents the search range and best hyperparameters obtained through this process.

The subsequent stage of our ML-based framework involved the classification of molecules using an XGB classifier (XGBc). As illustrated in Figure 2b, the training of the XGBc model leveraged the 3D data generated by the DRPL stage as input, with the target being the class labels of the peptides (BBB+ or BBB-). To optimize the performance of XGBc, we performed a grid search for the optimal set of hyperparameters, similar to the approach used for XGBr. Supplementary Table S6 presents the search range and best hyperparameters obtained through this process.

To summarize, the ML-based framework proposed in this study for predicting peptide penetration across the BBB consists of a pipeline comprising the XGBr and XGBc algorithms. The former algorithm was trained to learn the DR pattern produced by the sLE algorithm, whereas the latter was responsible for predicting whether a given peptide could penetrate the BBB, as shown in Figure 2c. Importantly, BrainPepPass also facilitates 3D visualization of new data, which is a key feature for analyzing the extent of separation of a given peptide from its original cluster.

The methodology proposed in this study is summarized in the flowchart shown in Figure 3. Figure 3a illustrates the data extraction step from Brainpeps, where peptides are classified into BBB+ and BBB- based on physicochemical indicators, the three datasets are structured, and the molecular descriptors are calculated. Figure 3b shows the preprocessing step in which the FCs are constructed for each dataset. Figure 3c shows the final step, in which the data are partitioned into training and testing sets, and each of the algorithms analyzed in this study is trained and tested for their respective comparisons using various metrics.

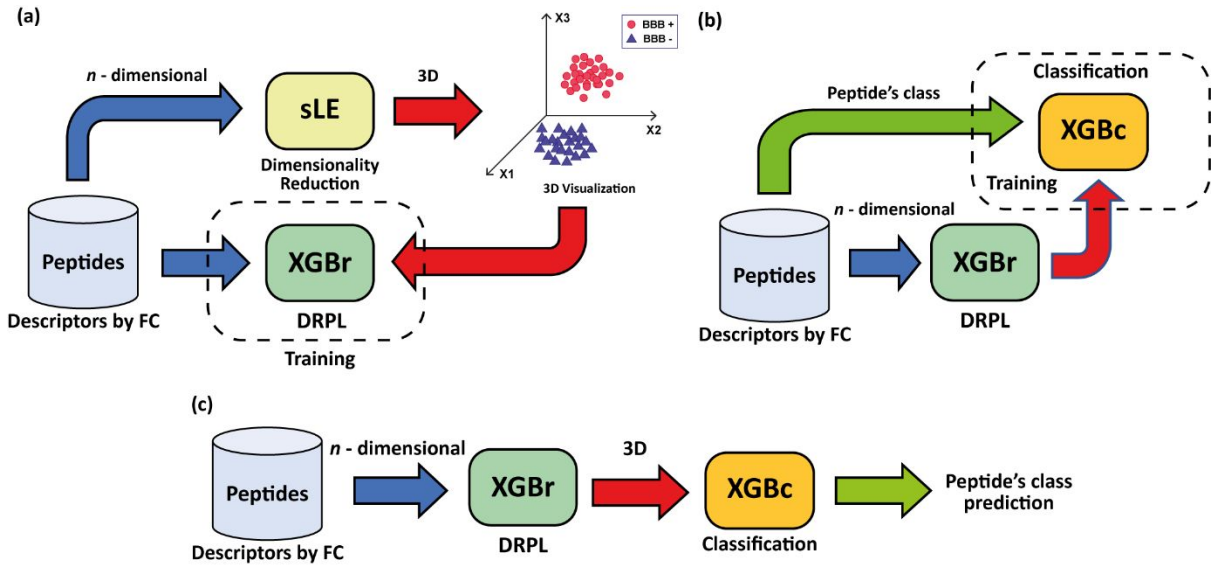


Figure 2. Stages of the BrainPepPass. (a) DRPL stage. (b) Training of the XGBc for predicting peptides permeability. (c) Final architecture of the BrainPepPass.

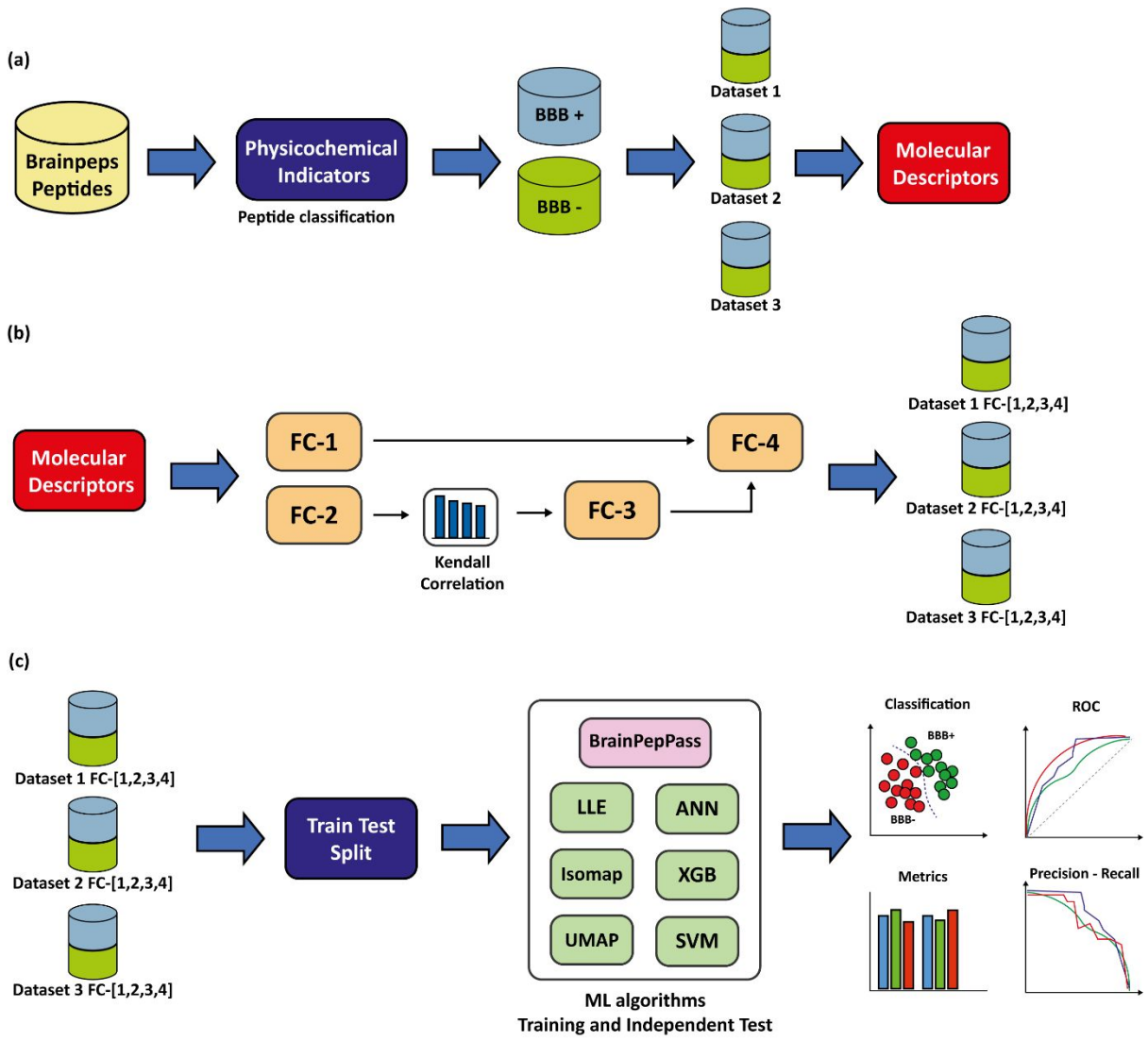


Figure 3. Flowchart of the proposed method to predict B3PPs. (a) Data extraction and molecular properties calculation. (b) Data preprocessing to generate the FCs. (c) B3PPs prediction with BrainPepPass and comparison with other algorithms.

RESULTS AND DISCUSSION

We conducted a thorough analysis using BrainPepPass to predict the permeability of peptides across the BBB by considering multiple factors. First, we performed a Kendall correlation analysis to examine the relationship between the selected structural and physicochemical properties of the peptides and their corresponding permeability classes for each FC. This analysis provided insights into the behavior of previously studied permeability properties in our dataset and helped identify the most relevant descriptors to compose FC-3 and FC-4. Thus, the present study evaluated the performance of the proposed ML-based framework trained with three datasets for each FC using 10-fold cross-validation and independent testing. Additionally, leave-one-out cross-validation (LOOCV) was used to investigate the predictive generalization of BrainPepPass. Moreover, the proposed method was compared with state-of-the-art classifiers previously used in the same research field, such as artificial neural network (ANN), SVM, and XGB. We also compared BrainPepPass using the same ML architecture as shown in Figure 2 but with other manifold DR algorithms such as locally linear embedding (LLE), isometric mapping (Isomap), and uniform manifold approximation and projection (UMAP). These comparisons assessed the predictive power and information content of the molecular descriptors in discriminating permeability classes of peptides.

Correlation Analysis and Feature Selection

We employed the Kendall correlation to evaluate the association between the molecular descriptors investigated in FC-1 (see Figure 4a) and all 749 descriptors calculated in FC-2 and their corresponding permeability classes. Based on these results, we identified the top ten descriptors with the highest correlation values to form FC-3 (see Figure 4b).

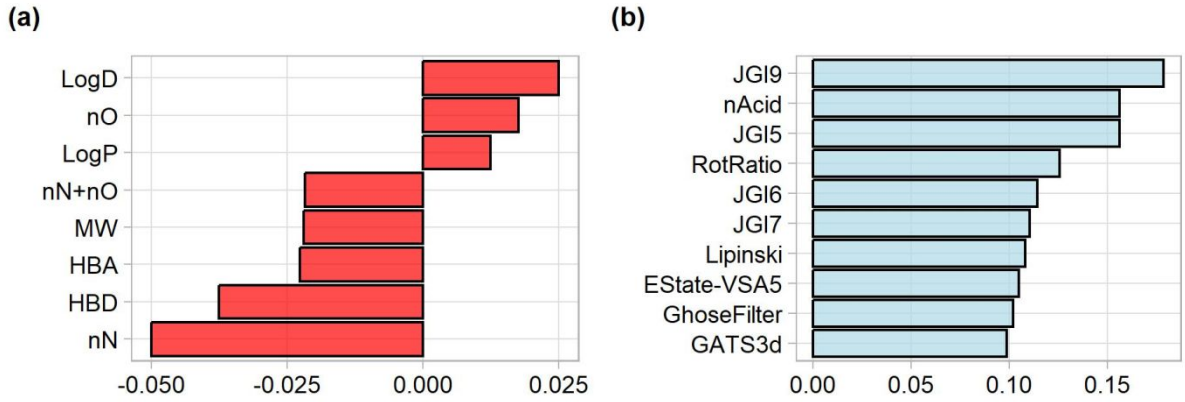


Figure 4. Kendall correlation analysis on molecular descriptors regarding to permeability across the BBB. (a) Molecular descriptors previously reported as associated with the permeability of small molecules across the BBB (FC-1). (b) 10 most correlated Mordred's descriptors (FC-3).

The findings depicted in Figure 4 highlight some aspects of the relationship between the molecular properties and permeability of peptides across the BBB. Specifically, the properties

included in FC-1, which have been previously reported to exhibit the highest correlation with the permeability of this biological barrier for small molecules, displayed comparatively lower correlation values in the Kendall correlation analysis than other structural and physicochemical properties derived from Mordred.

Figure 4b shows some properties related to the electrical properties of the investigated molecules. For example, the charge index (JGIx) is a topological descriptor that characterizes the molecular charge distribution on the *x*-th order.[62] Notably, values of 0.156, 0.114, 0.11, and 0.178 were obtained for *JGI5*, *JGI6*, *JGI7*, and *JGI9*, respectively. In addition, Estate-VSA calculates the sum of the van der Waals surface area contributions to the electron topological states within a specific range.[63] A value of 0.104 was observed for *Estate-VSA5*. The Geary coefficient (GATSyd) is a general index of 2D-autocorrelation with lag *y* applied to a molecular graph.[64] It describes the topology of a molecule associated with atomic masses, polarizabilities, and Sanderson electronegativities, weighted by sigma electrons. The descriptor *GATS3d* achieved a correlation value of 0.098.

Although no study has focused on the relationship between JGIx and BBB penetration, studies based on lipid bilayer membrane models to evaluate this pharmacokinetic property of compounds have revealed that charged molecules can modify the dipole potential of the

membrane through electrostatic interactions and interact with the BBB through attraction and repulsion.[55] Similarly, estate-VSA descriptors have not been investigated in depth regarding their association with the penetration of this barrier. However, Liu et al. concluded that high van der Waals surface area values are associated with good permeability of molecules in the BBB because molecules with high values tend to protonate and carry positive charges in molecules.[65] Similarly, the correlation between the Geary coefficient and BBB penetration has not yet been investigated in previous studies.

Additionally, the correlations of other selected features related to the structural properties of the molecules were investigated. For instance, *RotRatio*, which represents the ratio of the number of rotatable bonds to the total number of bonds in the molecule, showed a correlation value of 0.125. Similarly, the *nAcid* property, representing the number of acidic groups, showed a correlation value of 0.156. Some studies have indicated that the smaller the number of rotational bonds, that is, five or fewer bonds, the greater the permeability of the molecule in the CNS.[54] However, the number of acidic groups was not directly evaluated as a parameter to filter molecules that could penetrate the BBB, although Dichiara et al. observed in their studies that acidic compounds are among the least permeable across the barrier.[53]

Several of the descriptors selected for this analysis were linked to the drug-likeness of the molecules and other molecular properties. For example, the descriptor *Lipinski* represents a logical feature based on Lipinski's rule of five, which determines whether a molecule can be considered an orally available drug by satisfying specific numerical criteria for MW, LogP, HBA, and HBD.[57] Another cheminformatics filter, *GhoseFilter*, defines drug-likeness constraints for molecules based on their LogP and MW values, total number of atoms, and molar refractivity.[66] The correlation values for *Lipinski* and *GhoseFilter* were found to be 0.108 and 0.101, respectively.

The molecular descriptors present in FC-1 have a long history of investigation regarding their correlation with the pharmacokinetic property of small molecules in crossing the BBB. However, our results indicated that the correlation values for these properties were approximately two to four times lower than those for *GATS3d*, which had the lowest correlation value in FC-3. This finding suggests that the descriptive properties of charge distribution in peptides are closely related to their ability to penetrate the BBB. Notably, our analysis also demonstrated that *Lipinski* and *GhoseFilter* descriptors are also significantly associated with peptide penetration of the BBB.

The results of this investigation also contribute to refining the selection of descriptors for FC-4, which combines the properties of FC-1 and FC-3. FC-4 is essential for evaluating whether combining these descriptors will provide information gain to correctly classify molecules that cross the BBB.

Cross-validation and Independent Test Analysis

We evaluated the predictive capacity of BrainPepPass based on the accuracy of 10-fold cross-validation using the training portion of each dataset. This metric was applied to three dataset samples and four FCs, and 72 simulations were performed for different values of the sLE gamma parameter: 0.01, 0.02, 0.05, 0.1, 0.2, and 0.5. The best models were selected based on the highest accuracy values in the cross-validation for a fixed gamma, which was determined by filtering among all simulations (Figure 5a). The results demonstrate the contribution of each group of descriptors in predicting B3PPs using the proposed ML-based framework. BrainPepPass achieved values greater than 93% of average accuracy for all FCs. FC-1 exhibited the worst performance, with average accuracy values between 93.6% and 96%, whereas FC-2, which comprised the largest number of features, achieved an accuracy of 99.4% for the three datasets. FC-3 obtained values between 97.68% and 98.86%, whereas the FC-4

merged both FC-1 and FC-3 descriptors, obtaining accuracy values ranging from 98.8% to 100%.

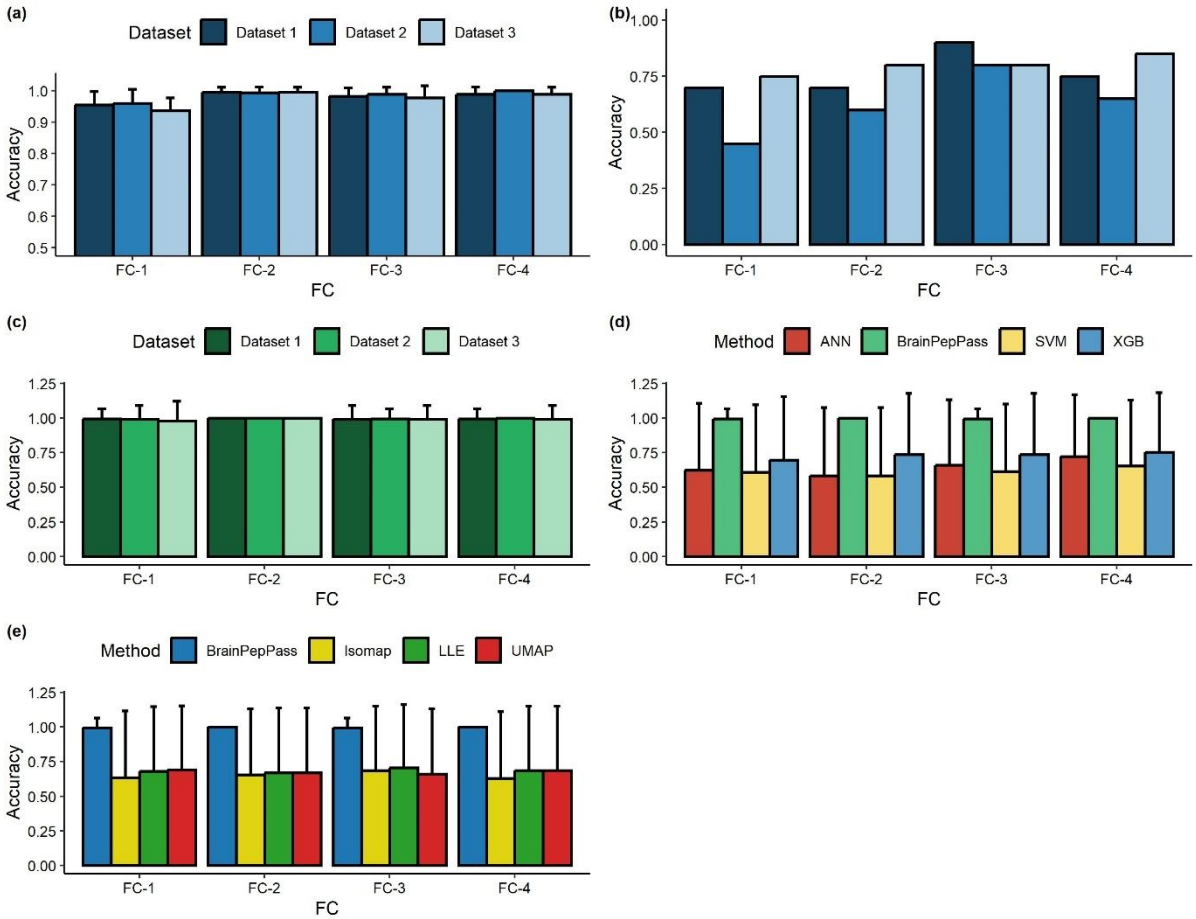


Figure 5. Accuracy achieved by BrainPepPass in each FC. (a) 10-fold cross-validation. (b) Independent test. (c) LOOCV. (d) LOOCV comparison with ANN, SVM, and XGB models. (e) LOOCV comparison with frameworks based on LLE, Isomap, and UMAP manifold algorithms.

The ANOVA test was applied to the accuracy values obtained for each fold of the 10-fold cross-validation performed on the three datasets, 1, 2, and 3, FC-2 and FC-4. The ANOVA test showed no statistically significant difference between the three datasets, yielding *p*-values of 0.526, 0.331, and 0.541 for datasets 1, 2, and 3, respectively. However, from a computational perspective, a significant difference exists between the models, as BrainPepPass trained with FC-2 requires the calculation of 749 descriptors, whereas the framework based on FC-4 requires only 19.

We also examined the predictive performance of our ML-based framework by applying external validation on peptides that were not part of the cross-validation analysis (Figure 5b). The accuracy outcomes obtained by the proposed tool for each feature composition indicate that the feature distribution between the training and test data in each of the three datasets may have been different. This is particularly evident when the performances of FC-2 and FC-4 are compared with the performance of FC-3. The ten descriptors selected from Mordred demonstrated superior predictive performance, achieving values ranging from 80% to 90% in predicting which peptides can penetrate the BBB. The FC-4 model achieved an accuracy of 85% for one of the datasets. Other performance metrics for the best BrainPepPass models by FC were also calculated, as shown in Table 1. The FC-3 model also yielded high F1-score and

Matthew's correlation coefficient (MCC) values, along with the maximum recall value for one of the datasets. The area under the receiver operating characteristic curve (ROC-AUC) values between 0.74 and 0.84 also indicate that BrainPepPass has a good ability to distinguish between the two classes (BBB+ and BBB-). These results indicate that framework can accurately predict which peptides can penetrate the BBB among all relevant instances using as much the selected molecular descriptors grouped in FC-3, as the properties included in FC-4, which maintained similar performance. Supplementary Table S7 provides the metric values obtained by BrainPepPass and their respective gamma values using the three datasets in the 10-fold cross-validation and independent tests, respectively.

Table 1. Independent test analysis of the best BrainPepPass models by FC.

FC	Dataset	Accuracy	F1 Score	MCC	Precision	Recall	ROC-AUC
FC-1	Dataset 3	0.75	0.71	0.52	0.86	0.60	0.74
FC-2	Dataset 3	0.80	0.82	0.61	0.75	0.90	0.75
FC-3	Dataset 1	0.90	0.91	0.82	0.83	1.00	0.74
FC-4	Dataset 3	0.85	0.84	0.70	0.89	0.80	0.84

The findings from the two analyses demonstrate the efficacy of the proposed ML-based framework in accurately predicting B3PPs, with cross-validation accuracy values exceeding 90% and values between 75% and 90% for the external validation set. Furthermore, this study highlights the contribution of the descriptors evaluated in terms of their association with BBB permeability and their comparison with descriptors associated with the charge distribution of the molecules. However, the independent test step involved a limited number of samples, with each erroneous prediction causing 5% reduction in the accuracy of each model. Consequently, determining the optimal BrainPepPass configuration was challenging. Therefore, this study also employs the leave-one-out cross-validation (LOOCV) metric as a complementary analysis to evaluate the proposed framework.

LOOCV Analysis

LOOCV is a model evaluation method similar to k-fold cross-validation, in which the testing set contains only one sample (k =1), and the remaining samples are used for training.[67] We conducted an LOOCV evaluation using three complete datasets (consisting of the training and testing subsets) for each FC. The results demonstrated that FC-2 enabled the BrainPepPass to

attain a mean accuracy with a peak value in all datasets, whereas FC-4 displayed comparable efficacy only for Dataset 2. Datasets 1 and 3 achieved accuracy scores of 99% and 98%, respectively. Supplementary Table S8 lists metric values obtained by BrainPepPass using the three datasets in LOOCV.

A comparison of the results of LOOCV with those obtained in the 10-fold cross-validation shows that the feature compositions that provided more information for predicting B3PPs were FC-2 and FC-4, highlighting the importance of the molecular descriptors of both FCs in differentiating the two classes of peptides. The outcomes from the LOOCV of the datasets belonging to these two FCs were analyzed through a pairwise comparison using an ANOVA test. The results indicate no statistically significant difference between the means of Datasets 1 and 3, with p-values of 0.318 and 0.157, respectively. Upon comparing the performance of the most effective models in LOOCV with that achieved in an independent test, FC-3 and FC-4 descriptors ranked among the highest in their ability to predict B3PPs. Specifically, BrainPepPass, based on FC-3, exhibited only a single misclassification in LOOCV and two misclassifications in the independent test. In contrast, the model based on FC-4 achieved satisfactory classification in LOOCV but failed to correctly classify the three molecules in the external validation. Although FC-2 obtained the third-highest accuracy value in the

independent test, it outperformed FC-3 by achieving the maximum classification value in the LOOCV experiment.

According to the three evaluation methods employed in the proposed ML-based framework, FC-4 predicted B3PPs with the highest accuracy. This descriptor group employed a less complex model consisting of 19 descriptors in contrast to FC-2, which also displayed high accuracy values. The success of FC-4 can be attributed to the efficacy sLE algorithm in reducing the dimensionality of the molecular descriptors. Figure 6 shows the projection of the molecular descriptors belonging to this feature composition in a 3D space after dimensionality reduction was performed during the pattern learning phase of the proposed framework. Our observations revealed that Dataset 1 exhibits an overlap between two peptides belonging to different classes (see the blue arrow in Figure 6a), whereas Dataset 3 displayed an overlap between at least three peptides from distinct classes (see the blue arrow in Figure 6c). This pattern is consistent with the results shown in Figure 5c. Additionally, the 3D projections of FC-4 reveal the potential for differentiation of BBB+ and BBB- peptides, besides clustering both classes, through the integration of molecular descriptors investigated in FC-1 and those selected from Mordred.

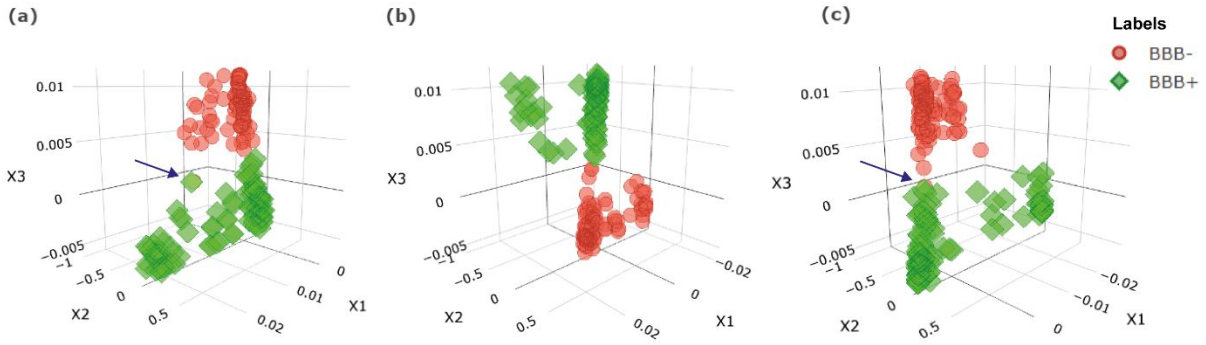


Figure 6. Dimensionality reduction result of BrainPepPass in pattern learning stage for FC-4. (a) Dataset 1. (b) Dataset 2. (c) Dataset 3.

Performance Comparison with other ML Algorithms

We compared the performance of BrainPepPass with SVM, ANN, and XGB models using the same three evaluation methods as used previously. We also compared the proposed ML-based framework with the same architecture shown in Figure 2 using the LLE, Isomap, and UMAP algorithms. The results of the optimal models for 10-fold cross-validation were computed for all combinations of FCs and datasets. The results of the best models are presented in Table 2. Supplementary Table S9 presents the search range and best hyperparameters achieved by LLE, Isomap, and UMAP in this analysis.

Table 2. Cross-validation analysis for the best BrainPepPass, ANN, SVM, XGB, LLE, Isomap, and UMAP models by each FC.

Method	FC-1	FC-2	FC-3	FC-4
BrainPepPass	0.93 ± 0.04	0.99 ± 0.01	0.98 ± 0.02	0.98 ± 0.02
ANN	0.58 ± 0.14	0.57 ± 0.12	0.60 ± 0.14	0.61 ± 0.15
SVM	0.55 ± 0.12	0.59 ± 0.16	0.58 ± 0.09	0.55 ± 0.12
XGB	0.55 ± 0.11	0.60 ± 0.10	0.61 ± 0.11	0.59 ± 0.05
LLE	0.55 ± 0.11	0.61 ± 0.13	0.46 ± 0.14	0.51 ± 0.10
Isomap	0.45 ± 0.09	0.44 ± 0.08	0.48 ± 0.14	0.53 ± 0.10
UMAP	0.48 ± 0.16	0.52 ± 0.09	0.46 ± 0.13	0.54 ± 0.13

The results of the cross-validation analysis revealed that BrainPepPass surpassed the average accuracy of the other ML algorithms for all feature compositions. These classifiers could not achieve an accuracy higher than 61% even for FCs with a reduced number of molecular descriptors. The frameworks based on LLE, Isomap, and UMAP achieved significantly lower average accuracy results, with values between 0.44 and 0.61, when compared to BrainPepPass. This discrepancy in performance can be attributed to the capacity of each technique to capture the nonlinear correlation between the descriptors and permeability classes. Our observations indicate that the sLE algorithm incorporated in the BrainPepPass effectively discriminates the

peptides that can cross the BBB from those that cannot, learning the nonlinear correlations between molecular descriptors and peptide labels, even better than the other three manifold DR algorithms employed in this study, contributing to the overall high performance of BrainPepPass.

We also applied an independent test to the other algorithms and compared them with the BrainPepPass (see Table 3). The results demonstrated that the proposed tool outperformed the ML models in terms of accuracy and other evaluation metrics for almost all FCs in the B3PP prediction. Among the state-of-the-art algorithms and other frameworks, the ANN model achieved the highest accuracy for FC-3 (80%); however, it did not surpass the performance of the proposed ML-based framework for the same FC (90%). Furthermore, the F1-score, MCC, precision, recall and ROC-AUC metrics indicated the exceptional performance of BrainPepPass, achieving higher values than the other techniques in most scenarios. A comparison of the results obtained by the proposed framework with those of other models using LLE, Isomap, and UMAP shows that in no scenario did these models surpass the performance of BrainPepPass for any of the metrics, thereby corroborating the results achieved in the cross-validation analysis. The performance metric values obtained by ANN, SVM, XGB, LLE,

Isomap, and UMAP using the three datasets in the 10-fold cross-validation and independent tests are provided in Supplementary Tables S10, S11, S12, S13, S14, and S15.

Similarly, we evaluated the predictive capacity of all ML models using LOOCV and compared their performance with that of BrainPepPass (see Figures 5d and 5e). Our findings indicate that the classifiers failed to surpass the predictive capacity of the proposed ML-based framework for all the FCs. Among the ML models, XGB achieved the highest average accuracy value of 75.26% for FC-4, which was lower than the value achieved by BrainPepPass with the same feature composition. The average accuracy values for the ANN and SVM ranged between 58.25% and 72.16% across different FCs. The results achieved by the frameworks using manifold DR algorithms were also unable to surpass the predictive capacity of BrainPepPass. The framework using LLE showed the best performance with an average accuracy of 70.62% for FC-3, whereas Isomap and UMAP achieved values between 62.89% and 69.07% for all FCs. The performance values obtained by ANN, SVM, XGB, LLE, Isomap, and UMAP models using the three datasets in LOOCV analysis are provided in Supplementary Tables S16, S17, S18, S19, S20, and S21.

The findings indicated in Table 3 were corroborated by evaluating the independent test based on the results of the ROC curve shown in Figure 7. The achieved ROC-AUC value for

BrainPepPass for FC-1, FC-2, and FC-3 remained within a narrow range between 0.74 and 0.75, whereas for FC-4, it was 0.84, surpassing the other models. The XGB algorithm demonstrated the second-best performance in this analysis, achieving an ROC-AUC value of 0.83 for FC-3. The ROC curves for all models are provided in Supplementary Figure S1.

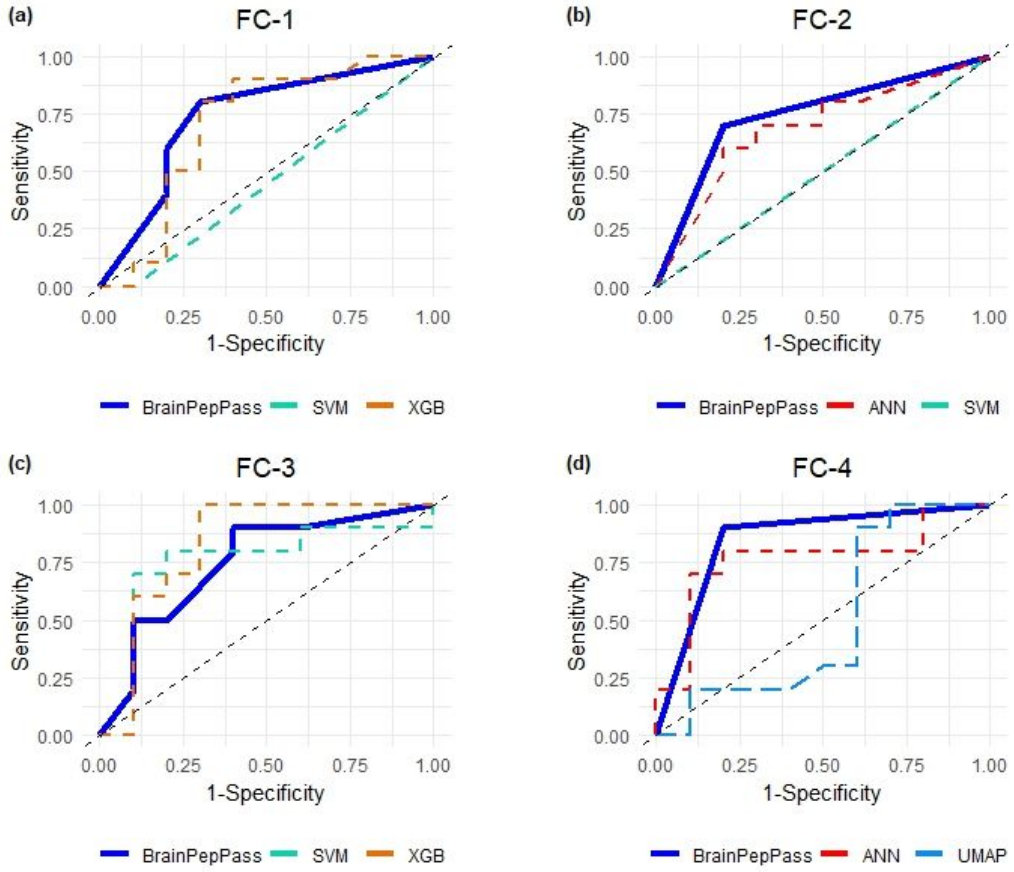


Figure 7. ROC curves comparison among the best BrainPepPass and the best and worst model by each FC in the independent test. (a) FC-1. (b) FC-2. (c) FC-3. (d) FC-4.

The difference in the performance of the proposed BrainPepPass for different FCs was assessed from the corresponding precision-recall curve (see Figure 8). The proposed framework achieved the highest average precision (AP) score of 0.66 for FC-1. The AP score for FC-2, FC-3, and FC-4 were 0.69, 0.70, and 0.78, respectively. The AP values for all FCs when the ANN, SVM, and XGB algorithms were used ranged from 0.5 to 0.8, whereas they were between 0.51 and 0.82 for the manifold-based algorithms. These results indicate that, despite the strong performance of BrainPepPass, the model using the UMAP algorithm demonstrates a greater balance between precision and recall when employing FC-3. The precision-recall curves for all models are provided in Supplementary Figure S2.

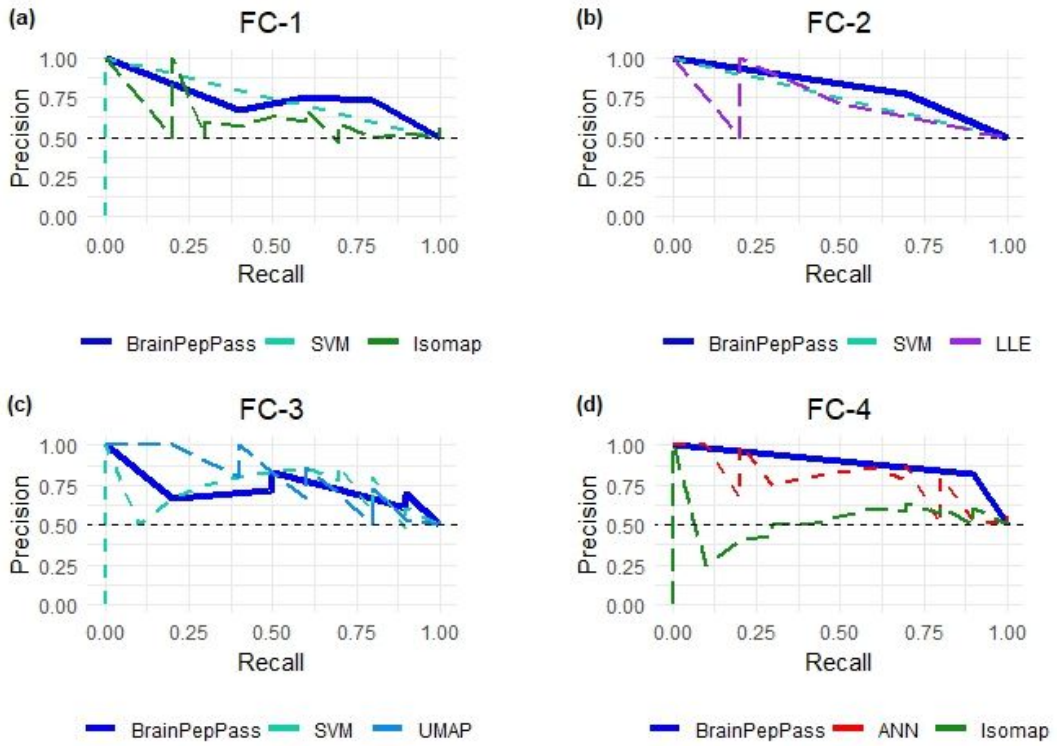


Figure 8. Precision-recall curves comparison among the best BrainPepPass and the best and worst model by each FC in the independent test. (a) FC-1. (b) FC-2. (c) FC-3. (d) FC-4.

Table 3. Independent test analysis for the best BrainPepPass, ANN, SVM, XGB, LLE, Isomap, and UMAP models by each FC.

FC-1	Method	Accuracy	F1-score	MCC	Precision	Recall	ROC-AUC
	BrainPepPass	0.75	0.70	0.52	0.85	0.75	0.74
	ANN	0.55	0.52	0.10	0.55	0.55	0.55
	SVM	0.65	0.58	0.31	0.71	0.65	0.45
	XGB	0.60	0.50	0.21	0.66	0.60	0.70

	LLE	0.65	0.63	0.30	0.66	0.60	0.57
	Isomap	0.60	0.60	0.20	0.60	0.60	0.60
	UMAP	0.75	0.70	0.52	0.85	0.60	0.45
FC-2							
	BrainPepPass	0.80	0.81	0.61	0.75	0.80	0.75
	ANN	0.65	0.58	0.31	0.71	0.65	0.69
	SVM	0.55	0.52	0.10	0.55	0.55	0.50
	XGB	0.70	0.70	0.40	0.70	0.70	0.63
	LLE	0.65	0.58	0.31	0.71	0.50	0.64
	Isomap	0.60	0.42	0.25	0.75	0.30	0.58
	UMAP	0.70	0.70	0.40	0.70	0.70	0.44
FC-3							
	BrainPepPass	0.90	0.90	0.81	0.83	0.90	0.75
	ANN	0.80	0.80	0.60	0.80	0.80	0.77
	SVM	0.70	0.66	0.40	0.75	0.70	0.75
	XGB	0.70	0.70	0.40	0.70	0.70	0.83
	LLE	0.75	0.73	0.50	0.77	0.70	0.71
	Isomap	0.70	0.72	0.40	0.66	0.80	0.76
	UMAP	0.75	0.73	0.50	0.77	0.70	0.76
FC-4							
	BrainPepPass	0.85	0.84	0.70	0.88	0.85	0.84
	ANN	0.65	0.69	0.31	0.61	0.65	0.77
	SVM	0.65	0.58	0.31	0.71	0.65	0.55
	XGB	0.70	0.75	0.43	0.64	0.70	0.72
	LLE	0.60	0.60	0.20	0.60	0.60	0.53
	Isomap	0.60	0.63	0.20	0.58	0.70	0.55

UMAP	0.60	0.55	0.20	0.62	0.50	0.50

It is noteworthy to compare the performance of the BrainPepPass with previously developed techniques for predicting B3PPs. While some ML-based tools, such as BBPpred[68], B3Pred,[69] BBPpredict,[37] and SCMB3PP[70] have been developed to predict the BBB permeability of peptides using ML algorithms trained with properties extracted from the primary structure of natural peptides encoded in FASTA format, the proposed ML-based framework presented herein employs a distinct approach by incorporating the 3D structure of these molecules encoded in MOL format. Additionally, most peptides used for training and testing BrainPepPass contain chemical modifications, which further distinguishes our tool from those that focus on natural peptides.

We conducted a comparative analysis between BrainPepPass and the BBBPpred, BBBPpredict, and SCMB3PP algorithms, which are available for public and free use. To assess the performance of the proposed model against other tools in an independent test, we selected the version based on FC-4 and trained with Dataset 2, which achieved the best performance in LOOCV analyses. We used 17 natural BBB+ peptides extracted from Brainpeps to compare the ML models performance, none of these molecules were utilized in any of the previously described training or independent testing steps for the selected BrainPepPass version. This

dataset was balanced with 17 natural BBB- peptides randomly extracted from the test dataset of the SCMB3PP tool, resulting in 34 structures for this analysis. We also developed a BrainPepPass model with FC-4, named BrainPepPass-N, which was exclusively trained using natural peptides collected from the same dataset that was used to train SCMB3PP model. Table 4 presents the values achieved by all the algorithms based on the key metrics. The peptide sequences used in this analysis are listed in Supplementary Table S22.

Table 4. Analysis of independent test comparing BrainPepPass and BrainPepPass-N with BBPpred, BBPpredict, and SCMB3PP algorithms using natural peptides.

Algorithm	Accuracy	F1-score	MCC	Precision	Recall
BrainPepPass	0.52	0.55	0.06	0.55	0.55
BrainPepPass-N	0.97	1.0	0.94	1.0	1.0
BBPpred	0.64	0.71	0.33	0.60	0.88
BBPpredict	0.55	0.66	0.15	0.53	0.88
SCMB3PP	0.91	0.90	0.82	0.93	0.88

According to the results presented in Table 4, BrainPepPass-N achieved the best outcomes, attaining an accuracy of approximately 97%, along with values exceeding 94% for the other metrics. This indicates that the proposed method, trained only on natural peptides and utilizing molecular descriptors from FC-4, can predict the permeability of natural peptides across the BBB with greater accuracy than that of the other tools. It is also important to highlight that the BrainPepPass model that was not exclusively trained on natural peptides failed to outperform the other tools. This could be attributed to the underfitting of this model with respect to natural peptide data, considering that it was predominantly trained on structures featuring chemical modifications.

Therefore, based on the results presented in this study, BrainPepPass exhibits exceptional performance in predicting peptide penetration across the BBB, surpassing existing ML classifiers in this area of research. BrainPepPass achieved average accuracy values exceeding 93% in the 10-fold cross-validation and between 75% and 90% in the independent test, with average accuracy values ranging between 99.48% and 100% according to LOOCV. For the FC-4 model, which exhibited a positive relationship between efficiency and complexity, average accuracy values of 99.21%, 75%, and 99.48% were achieved in cross-validation, independent testing, and LOOCV, respectively, across all three datasets. Although

BrainPepPass based on FC-4 did not achieve the best performance in predicting some natural peptides, BrainPepPass-N version attained an accuracy of 97% in the same test. These outcomes demonstrate that the proposed tool has impressive predictive capabilities in determining whether natural or chemically modified peptides can penetrate the BBB, based on the molecular descriptors examined.

CONCLUSION

Predicting the ability of natural and chemically modified peptides to penetrate the BBB is a significant challenge in computational and medicinal chemistry. The development of an efficient computational tool to perform this task requires solving the following problems: (1) obtaining experimentally validated data that include natural and chemically modified peptides; (2) performing an exploratory analysis of the correlation between several molecular descriptors related to BBB permeability; and (3) training a robust and powerful ML-based model to learn the non-linear pattern between peptide descriptors and permeability classes. However, public information on the pharmacokinetic properties of peptides that could be used as a reference for

experimental studies is scarce, making generating large datasets that facilitates the training of the algorithms difficult.

Our predictive model relies on an algorithmic architecture that reduces the various dimensions of molecular descriptors derived from peptides to three. By leveraging the ability of the sLE technique to cluster and segregate samples into their respective classes, our model incorporates a robust preprocessing step, thereby streamlining the prediction process of B3PPs. Moreover, our ML-based framework offers the additional advantage of processing information extracted from peptide structures in the MOL format, which encodes chemical modifications and cyclic chains in the molecular structure, thereby achieving a novel breakthrough in this field of research and improving the exploration of increasingly complex structures within this molecular class. Additionally, our investigation highlights the correlation between several molecular descriptors and BBB permeability, specifically emphasizing the role of charge distribution properties in the ability of peptides to permeate through the BBB.

Another advantage our of study was the improved performance of BrainPepPass relative to other machine learning models in predicting peptide penetration. The predictive capability of our tool in all applied tests surpassed that of other machine-learning classifiers as well as that of the same framework but with other manifold DR algorithms. Furthermore, the proposed

framework demonstrated good performance compared with other available tools for predicting natural B3PPs. This proves the ability of our tool in assisting the virtual screening of new peptides that penetrate the BBB, thereby contributing to the discovery and development of new bioactive molecules capable of reaching the CNS.

ASSOCIATED CONTENT

Data Availability Statement

The BrainPepPass tool is available in a GitHub repository, which can be accessed at: https://github.com/ewerton-cristhian/BrainPepPass. This repository contains information about the online versions of the BrainPepPass available for users, a user manual with instructions on how to use the tools, and the ML models used in the framework.

The source code of the BrainPepPass in Python language to execute the best model can be accessed at https://figshare.com/s/18d704599c397f54b3ac. The dataset of peptide structures can be accessed at https://figshare.com/s/f8ae1e2f6e4b2170807f. The scripts used to generate and evaluate the BrainPepPass and other ML models in the present work is available at https://figshare.com/s/8bc7ab7b424e04f680e0.

Supporting Information

The Supporting Information is available free of charge at JCIM web site.

List of peptides used; Table S1, list of molecular descriptors in FC-1; Table S2, list of molecular descriptors in FC-2; Table S3, list of molecular descriptors in FC-3; Table S4, hyperparameters employed in grid search for training XGBr in BrainPepPass; Table S5, hyperparameters employed in grid search for training XGBc in BrainPepPass; Table S6, results reached in cross-validation and independent test by BrainPepPass for all FCs and datasets; Table S7, results reached in LOOCV BrainPepPass for all FCs and datasets; Table S8, hyperparameters employed in grid search for training LLE, Isomap, and UMAP; Table S9, results reached in cross-validation and independent test by the ANN for all FCs and datasets; Table S10, results reached in cross-validation and independent test by the SVM for all FCs and datasets; Table S11, results reached in cross-validation and independent test by the XGB for all FCs and datasets; Table S12, results reached in cross-validation and independent test by the framework based on LLE for all FCs and datasets; Table S13, results reached in cross-validation and independent test by the framework based on Isomap for all FCs and datasets; Table S14, results reached in cross-validation and independent test by the framework based on

UMAP for all FCs and datasets; Table S15, results reached in LOOCV by the ANN for all FCs and datasets; Table S16, results reached in LOOCV by the SVM for all FCs and datasets; Table S17, results reached in LOOCV by the XGB for all FCs and datasets; Table S18, results reached in LOOCV by the framework based on LLE for all FCs and datasets; Table S19, results reached in LOOCV by the framework based on Isomap for all FCs and datasets; Table S20, results reached in LOOCV by the framework based on UMAP for all FCs and datasets; Table S21, list of natural peptide sequences used to compare BrainPepPass with other online tools; Table S22, ROC curves obtained by the best BrainPepPass, ANN, SVM, XGB, LLE, Isomap, and UMAP models by each FC in independent test; Figure S1, Precision-recall curves obtained by the best BrainPepPass, ANN, SVM, XGB, LLE, Isomap, and UMAP models by each FC in independent test; Figure S2.

AUTHOR INFORMATION

Corresponding Authors

Bart De Spiegeleer − Drug Quality and Registration (DruQuaR) Group,

Faculty of Pharmaceutical Sciences, Ghent University, B-9000 Ghent, Belgium;

orcid.org/0000-0001-6794-3108; Phone: +32 9 2648100; Email: Bart.DeSpiegeleer@ugent.be

Kauê Santana da Costa- Laboratório de Simulação Computacional, Campos Marechal Rondon, Instituto de Biodiversidade, Universidade Federal do Oeste do Pará, 68040-255, Santarém, Pará, Brasil. **E-mail:** kaue.costa@ufopa.ed.br.

Claudomiro de Souza de Sales Júnior - Laboratório de Inteligência Computacional e Pesquisa Operacional, Campos Belém, Instituto de Tecnologia, Universidade Federal do Pará, 66075-110, Belém, Pará, Brasil. **E-mail:** cssj@ufpa.br.

Authors

Ewerton Cristhian Lima de Oliveira - Laboratório de Inteligência Computacional e Pesquisa Operacional, Campos Belém, Instituto de Tecnologia, Universidade Federal do Pará, 66075-110, Belém, Pará, Brasil; Instituto Tecnológico Vale, 66055-090, Belém, Pará, Brasil; orcid.org/0000-0002-2338-7178.

Hannah Hirmz - Drug Quality and Registration (DruQuaR) Group, Faculty of Pharmaceutical Sciences, Ghent University, B-9000 Ghent, Belgium; orcid.org/0000-0002-7446-0267.

Evelien Wynendaele - Drug Quality and Registration (DruQuaR) Group, Faculty of Pharmaceutical Sciences, Ghent University, B-9000 Ghent, Belgium; orcid.org/0000-0003-0969-7580.

Juliana Auzier Seixas Feio - Laboratório de Inteligência Computacional e Pesquisa Operacional, Campos Belém, Instituto de Tecnologia, Universidade Federal do Pará, 66075-110, Belém, Pará, Brasil; orcid.org/0009-0009-9781-5911

Igor Matheus Souza Moreira - Laboratório de Inteligência Computacional e Pesquisa Operacional, Campos Belém, Instituto de Tecnologia, Universidade Federal do Pará, 66075-110, Belém, Pará, Brasil; orcid.org/0000-0002-8145-9395.

Anderson H. Lima - Laboratório de Planejamento e Desenvolvimento de Fármacos, Instituto de Ciências Exatas e Naturais, Universidade Federal do Pará, 66075-110, Belém, Pará, Brasil; orcid.org/0000-0002-8451-9912.

Kauê Santana - Laboratório de Simulação Computacional, Campos Marechal Rondon, Instituto de Biodiversidade, Universidade Federal do Oeste do Pará, 68040-255, Santarém, Pará, Brasil. orcid.org/0000-0002-2735-8016.

Bart De Spiegeleer – Drug Quality and Registration (DruQuaR) Group,

Faculty of Pharmaceutical Sciences, Ghent University, B-9000 Ghent, Belgium; orcid.org/0000-0001-6794-3108.

Claudomiro de Souza de Sales Junior - Laboratório de Inteligência Computacional e Pesquisa Operacional, Campos Belém, Instituto de Tecnologia, Universidade Federal do Pará, 66075-110, Belém, Pará, Brasil; orcid.org/0000-0002-2735-1383.

Notes

The authors declare no competing financial interest.

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

ACKNOWLEDGEMENTS

The authors are grateful to Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Pró-reitoria de Pesquisa da UFPA (PROPESP/UFPA), and Instituto Tecnológico Vale (ITV) for their financial and technical support.

REFERENCES

(1) Geldenhuys, W. J.; Mohammad, A. S.; Adkins, C. E.; Lockman, P. R. Molecular Determinants of Blood-Brain Barrier Permeation. *Ther Deliv* **2015**, *6* (8), 961–971. https://doi.org/10.4155/tde.15.32.

(2) Oller-Salvia, B.; Sánchez-Navarro, M.; Giralt, E.; Teixidó, M. Blood–Brain Barrier Shuttle Peptides: An Emerging Paradigm for Brain Delivery. *Chem Soc Rev* **2016**, *45* (17), 4690–4707. https://doi.org/10.1039/C6CS00076B.

(3) Díaz-Perlas, C.; Oller-Salvia, B.; Sánchez-Navarro, M.; Teixidó, M.; Giralt, E. Branched BBB-Shuttle Peptides: Chemoselective Modification of Proteins to Enhance Blood–Brain Barrier Transport. *Chem Sci* **2018**, *9* (44), 8409–8415. https://doi.org/10.1039/C8SC02415D.

(4) de Oliveira, E. C. L.; da Costa, K. S.; Taube, P. S.; Lima, A. H.; Junior, C. de S. de S. Biological Membrane-Penetrating Peptides: Computational Prediction and Applications. *Front Cell Infect Microbiol* **2022**, *12*. https://doi.org/10.3389/fcimb.2022.838259.

(5) Kastin, A. J.; Nissen, C.; Schally, A. V.; Coy, D. H. Blood-Brain Barrier, Half-Time Disappearance, and Brain Distribution for Labeled Enkephalin and a Potent Analog. *Brain Res Bull* **1976**, *1* (6), 583–589. https://doi.org/10.1016/0361-9230(76)90085-X.

(6) Pelletier, G.; Labrie, F.; Kastin, A. J.; Schally, A. V. Radioautographic Localization of Radioactivity in Rat Brain after Intracarotid Injection of 125I-α-Melanocyte-Stimulating Hormone. *Pharmacol Biochem Behav* **1975**, *3* (4), 671–674. https://doi.org/10.1016/0091-3057(75)90190-2.

(7) Kastin, A. J.; Nissen, C.; Nikolics, K.; Medzihradszky, K.; Coy, D. H.; Teplan, I.; Schally, A. V. Distribution of 3H-α-MSH in Rat Brain. *Brain Res Bull* **1976**, *1* (1), 19–26. https://doi.org/10.1016/0361-9230(76)90045-9.

(8) Banks, W. A. Peptides and the Blood–Brain Barrier. *Peptides (N.Y.)* **2015**, *72*, 16–19. https://doi.org/10.1016/j.peptides.2015.03.010.

(9) Wang, D.; Chen, F.; Han, Z.; Yin, Z.; Ge, X.; Lei, P. Relationship Between Amyloid-β Deposition and Blood–Brain Barrier Dysfunction in Alzheimer's Disease. *Front Cell Neurosci* **2021**, *15*. https://doi.org/10.3389/fncel.2021.695479.

(10) Baskin, D. G.; Stein, L. J.; Ikeda, H.; Woods, S. C.; Figlewicz, D. P.; Porte, D.; Greenwood, M. R. C.; Dorsa, D. M. Genetically Obese Zucker Rats Have Abnormally Low Brain Insulin Content. *Life Sci* **1985**, *36* (7), 627–633. https://doi.org/10.1016/0024-3205(85)90166-3.

(11) Kaiyala, K. J.; Prigeon, R. L.; Kahn, S. E.; Woods, S. C.; Schwartz, M. W. Obesity Induced by a High-Fat Diet Is Associated with Reduced Brain Insulin Transport in Dogs. *Diabetes* **2000**, *49* (9), 1525–1533. https://doi.org/10.2337/diabetes.49.9.1525.

(12) Urayama, A.; Banks, W. A. Starvation and Triglycerides Reverse the Obesity-Induced Impairment of Insulin Transport at the Blood-Brain Barrier. *Endocrinology* **2008**, *149* (7), 3592–3597. https://doi.org/10.1210/en.2008-0008.

(13) Banks, W. A.; Jaspan, J. B.; Kastin, A. J. Effect of Diabetes Mellitus on the Permeability of the Blood-Brain Barrier to Insulin. *Peptides (N.Y.)* **1997**, *18* (10), 1577–1584. https://doi.org/10.1016/S0196-9781(97)00238-6.

(14) Stalmans, S.; Bracke, N.; Wynendaele, E.; Gevaert, B.; Peremans, K.; Burvenich, C.; Polis, I.; De Spiegeleer, B. Cell-Penetrating Peptides Selectively Cross the Blood-Brain Barrier in Vivo. *PLoS One* **2015**, *10* (10), e0139652. https://doi.org/10.1371/journal.pone.0139652.

(15) Zhou, X.; Smith, Q. R.; Liu, X. Brain Penetrating Peptides and Peptide–Drug Conjugates to Overcome the Blood–Brain Barrier and Target CNS Diseases. *Wiley Interdiscip Rev Nanomed Nanobiotechnol* **2021**, *13* (4). https://doi.org/10.1002/wnan.1695.

(16) He, S.; Zhiti, A.; Barba-Bon, A.; Hennig, A.; Nau, W. M. Real-Time Parallel Artificial Membrane Permeability Assay Based on Supramolecular Fluorescent Artificial Receptors. *Front Chem* **2020**, *8*. https://doi.org/10.3389/fchem.2020.597927.

(17) Radan, M.; Djikic, T.; Obradovic, D.; Nikolic, K. Application of in Vitro PAMPA Technique and in Silico Computational Methods for Blood-Brain Barrier Permeability Prediction of Novel CNS Drug Candidates. *European Journal of Pharmaceutical Sciences* **2022**, *168*, 106056. https://doi.org/10.1016/j.ejps.2021.106056.

(18) Bagchi, S.; Chhibber, T.; Lahooti, B.; Verma, A.; Borse, V.; Jayant, R. D. In-Vitro Blood-Brain Barrier Models for Drug Screening and Permeation Studies: An Overview. *Drug Des Devel Ther* **2019**, *13*, 3591–3605. https://doi.org/10.2147/DDDT.S218708.

(19) Williams-Medina, A.; Deblock, M.; Janigro, D. In Vitro Models of the Blood–Brain Barrier: Tools in Translational Medicine. *Front Med Technol* **2021**, *2*. https://doi.org/10.3389/fmedt.2020.623950.

(20) Santana, K.; do Nascimento, L. D.; Lima e Lima, A.; Damasceno, V.; Nahum, C.; Braga, R. C.; Lameira, J. Applications of Virtual Screening in Bioprospecting: Facts, Shifts, and Perspectives to Explore the Chemo-Structural Diversity of Natural Products. *Front Chem* **2021**, *9*. https://doi.org/10.3389/fchem.2021.662688.

(21) Vastag, M.; Keseru, G. M. Current in Vitro and in Silico Models of Blood-Brain Barrier Penetration: A Practical View. *Curr Opin Drug Discov Devel* **2009**, *12* (1), 115–124.

(22) Martin, I. Prediction of Blood-Brain Barrier Penetration: Are We Missing the Point? *Drug Discov Today* **2004**, *9* (4), 161–162. https://doi.org/10.1016/S1359-6446(03)02961-1.

(23) Bao, W.; Yang, B. Protein Acetylation Sites with Complex-Valued Polynomial Model. *Frontiers (Boulder)* **2024**, *18* (3).

(24) Bao, W.; Cui, Q.; Chen, B.; Yang, B. Phage_UniR_LGBM: Phage Virion Proteins Classification with UniRep Features and LightGBM Model. *Comput Math Methods Med* **2022**, *2022*, 1–8. https://doi.org/10.1155/2022/9470683.

(25) de Oliveira, E. C. L.; da Costa, K. S.; Taube, P. S.; Lima, A. H.; Junior, C. de S. de S. Biological Membrane-Penetrating Peptides: Computational Prediction and Applications. *Front Cell Infect Microbiol* **2022**, *12*. https://doi.org/10.3389/fcimb.2022.838259.

(26) Shoombuatong, W.; Schaduangrat, N.; Pratiwi, R.; Nantasenamat, C. THPep: A Machine Learning-Based Approach for Predicting Tumor Homing Peptides. *Comput Biol Chem* **2019**, *80*, 441–451. https://doi.org/10.1016/j.compbiolchem.2019.05.008.

(27) Mishra, A.; Kabir, M. W. U.; Hoque, M. T. DiSBPred: A Machine Learning Based Approach for Disulfide Bond Prediction. *Comput Biol Chem* **2021**, *91*, 107436. https://doi.org/10.1016/j.compbiolchem.2021.107436.

(28) Bao, W.; Gu, Y.; Chen, B.; Yu, H. Golgi_DF: Golgi Proteins Classification with Deep Forest. *Front Neurosci* **2023**, *17*. https://doi.org/10.3389/fnins.2023.1197824.

(29) Stalmans, S.; Gevaert, B.; Wynendaele, E.; Nielandt, J.; De Tre, G.; Peremans, K.; Burvenich, C.; De Spiegeleer, B. Classification of Peptides According to Their Blood-Brain Barrier Influx. *Protein Pept Lett* **2015**, *22* (9), 768–775. https://doi.org/10.2174/0929866522666150622101223.

(30) Ecker, G.; Noe, C. In Silico Prediction Models for Blood-Brain Barrier Permeation. *Curr Med Chem* **2012**, *11* (12), 1617–1528. https://doi.org/10.2174/0929867043365071.

(31) Abbott, N. J.; Patabendige, A. A. K.; Dolman, D. E. M.; Yusof, S. R.; Begley, D. J. Structure and Function of the Blood–Brain Barrier. *Neurobiol Dis* **2010**, *37* (1), 13–25. https://doi.org/10.1016/j.nbd.2009.07.030.

(32) Mensch, J.; Oyarzabal, J.; Mackie, C.; Augustijns, P. In Vivo, in Vitro and in Silico Methods for Small Molecule Transfer across the BBB. *J Pharm Sci* **2009**, *98* (12), 4429–4468. https://doi.org/10.1002/jps.21745.

(33) Konovalov, D. A.; Coomans, D.; Deconinck, E.; Heyden, Y. Vander. Benchmarking of QSAR Models for Blood-Brain Barrier Permeation. *J Chem Inf Model* **2007**, *47* (4), 1648–1656. https://doi.org/10.1021/ci700100f.

(34) Bendels, S.; Kansy, M.; Wagner, B.; Huwyler, J. In Silico Prediction of Brain and CSF Permeation of Small Molecules Using PLS Regression Models. *Eur J Med Chem* **2008**, *43* (8), 1581–1592. https://doi.org/10.1016/j.ejmech.2007.11.011.

(35) Faramarzi, S.; Kim, M. T.; Volpe, D. A.; Cross, K. P.; Chakravarti, S.; Stavitskaya, L. Development of QSAR Models to Predict Blood-Brain Barrier Permeability. *Front Pharmacol* **2022**, *13*. https://doi.org/10.3389/fphar.2022.1040838.

(36) Liu, X.; Tu, M.; Kelly, R. S.; Chen, C.; Smith, B. J. Development of a Computational Approach to Predict Blood-Brain Barrier Permeability. *Drug Metabolism and Disposition* **2004**, *32* (1), 132–139. https://doi.org/10.1124/dmd.32.1.132.

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

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

(39) Zou, H. Identifying Blood-Brain Barrier Peptides by Using Amino Acids Physicochemical Properties and Features Fusion Method. *Peptide Science* **2022**, *114* (2). https://doi.org/10.1002/pep2.24247.

(40) Van Dorpe, S.; Bronselaer, A.; Nielandt, J.; Stalmans, S.; Wynendaele, E.; Audenaert, K.; Van De Wiele, C.; Burvenich, C.; Peremans, K.; Hsuchou, H.; De Tré, G.; De Spiegeleer, B. Brainpeps: The Blood-Brain Barrier Peptide Database. *Brain Struct Funct* **2012**, *217* (3), 687–718. https://doi.org/10.1007/s00429-011-0375-0.

(41) Gjedde, A. High- and Low-Affinity Transport of D-Glucose from Blood to Brain. *J Neurochem* **1981**, *36* (4), 1463–1471. https://doi.org/10.1111/j.1471-4159.1981.tb00587.x.

(42) Poth, A. G.; Chiu, F. C. K.; Stalmans, S.; Hamilton, B. R.; Huang, Y. H.; Shackleford, D. M.; Patil, R.; Le, T. T.; Kan, M. W.; Durek, T.; Wynendaele, E.; De Spiegeleer, B.; Powell, A. K.; Venter, D. J.; Clark, R. J.; Charman, S. A.; Craik, D. J. Effects of Backbone Cyclization on the Pharmacokinetics and Drug Efficiency of the Orally Active Analgesic Conotoxin CVc1.1. *Med Drug Discov* **2021**, *10*, 100087. https://doi.org/10.1016/j.medidd.2021.100087.

(43) Van Dorpe, S.; Bronselaer, A.; Nielandt, J.; Stalmans, S.; Wynendaele, E.; Audenaert, K.; Van De Wiele, C.; Burvenich, C.; Peremans, K.; Hsuchou, H.; De Tré, G.; De Spiegeleer, B. Brainpeps: The Blood-Brain Barrier Peptide Database. *Brain Struct Funct* **2012**, *217* (3), 687–718. https://doi.org/10.1007/s00429-011-0375-0.

(44) Janssens, Y.; Debunne, N.; De Spiegeleer, A.; Wynendaele, E.; Planas, M.; Feliu, L.; Quarta, A.; Claes, C.; Van Dam, D.; De Deyn, P. P.; Ponsaerts, P.; Blurton-Jones, M.; De Spiegeleer, B. PapRIV, a BV-2 Microglial Cell Activating Quorum Sensing Peptide. *Sci Rep* **2021**, *11* (1), 10723. https://doi.org/10.1038/s41598-021-90030-y.

(45) Abbruscato, T. J.; Williams, S. A.; Misicka, A.; Lipkowski, A. W.; Hruby, V. J.; Davis, T. P. Blood-to-Central Nervous System Entry and Stability of Biphalin, a Unique Double-Enkephalin Analog, and Its Halogenated Derivatives. *Journal of Pharmacology and Experimental Therapeutics* **1996**, *276* (3), 1049–1057.

(46) Di, L.; Kerns, E. H.; Fan, K.; McConnell, O. J.; Carter, G. T. High Throughput Artificial Membrane Permeability Assay for Blood-Brain Barrier. *Eur J Med Chem* **2003**, *38* (3), 223–232. https://doi.org/10.1016/S0223-5234(03)00012-6.

(47) Zorbaz, T.; Braïki, A.; Maraković, N.; Renou, J.; de la Mora, E.; Maček Hrvat, N.; Katalinić, M.; Silman, I.; Sussman, J. L.; Mercey, G.; Gomez, C.; Mougeot, R.; Pérez, B.; Baati, R.; Nachon, F.; Weik, M.; Jean, L.; Kovarik, Z.; Renard, P. Y. Potent 3-Hydroxy-2-Pyridine Aldoxime Reactivators of Organophosphate-Inhibited Cholinesterases with Predicted Blood–Brain Barrier Penetration. *Chemistry - A European Journal* **2018**, *24* (38), 9675–9691. https://doi.org/10.1002/chem.201801394.

(48) Rossi, M.; Freschi, M.; De Camargo Nascente, L.; Salerno, A.; De Melo Viana Teixeira, S.; Nachon, F.; Chantegreil, F.; Soukup, O.; Prchal, L.; Malaguti, M.; Bergamini, C.; Bartolini, M.; Angeloni, C.; Hrelia, S.; Soares Romeiro, L. A.; Bolognesi, M. L. Sustainable Drug Discovery of Multi-Target-Directed Ligands for Alzheimer's Disease. *J Med Chem* **2021**, *64* (8), 4972–4990. https://doi.org/10.1021/acs.jmedchem.1c00048.

(49) Felegyi-Tóth, C. A.; Tóth, Z.; Garádi, Z.; Boldizsár, I.; Nedves, A. N.; Simon, A.; Felegyi, K.; Alberti, Á.; Riethmüller, E. Membrane Permeability and Aqueous Stability Study of Linear and Cyclic Diarylheptanoids from Corylus Maxima. *Pharmaceutics* **2022**, *14* (6), 1250. https://doi.org/10.3390/pharmaceutics14061250.

(50) Simon, A.; Darcsi, A.; Kéry, Á.; Riethmüller, E. Blood-Brain Barrier Permeability Study of Ginger Constituents. *J Pharm Biomed Anal* **2020**, *177*, 112820. https://doi.org/10.1016/j.jpba.2019.112820.

(51) Garberg, P.; Ball, M.; Borg, N.; Cecchelli, R.; Fenart, L.; Hurst, R. D.; Lindmark, T.; Mabondzo, A.; Nilsson, J. E.; Raub, T. J.; Stanimirovic, D.; Terasaki, T.; Öberg, J. O.; Österberg, T. In Vitro Models for the Blood-Brain Barrier. *Toxicology in Vitro* **2005**, *19* (3), 299–334. https://doi.org/10.1016/j.tiv.2004.06.011.

(52) Yoon, C. H.; Kim, S. J.; Shin, B. S.; Lee, K. C.; Yoo, S. D. Rapid Screening of Blood-Brain Barrier Penetration of Drugs Using the Immobilized Artificial Membrane Phosphatidylcholine Column Chromatography. *J Biomol Screen* **2006**, *11* (1), 13–20. https://doi.org/10.1177/1087057105281656.

(53) Dichiara, M.; Amata, B.; Turnaturi, R.; Marrazzo, A.; Amata, E. Tuning Properties for Blood-Brain Barrier Permeation: A Statistics-Based Analysis. *ACS Chem Neurosci* **2020**, *11* (1), 34–44. https://doi.org/10.1021/acschemneuro.9b00541.

(54) Mikitsh, J. L.; Chacko, A. M. Pathways for Small Molecule Delivery to the Central Nervous System across the Blood-Brain Barrier. *Perspect Medicin Chem* **2014**, *6* (6), 11–24. https://doi.org/10.4137/PMc.s13384.

(55) Fong, C. W. Permeability of the Blood–Brain Barrier: Molecular Mechanism of Transport of Drugs and Physiologically Important Compounds. *Journal of Membrane Biology* **2015**, *248* (4), 651–669. https://doi.org/10.1007/s00232-015-9778-9.

(56) Lovering, F.; Bikker, J.; Humblet, C. Escape from Flatland: Increasing Saturation as an Approach to Improving Clinical Success. *J Med Chem* **2009**, *52* (21), 6752–6756. https://doi.org/10.1021/jm901241e.

(57) Lipinski, C. A. C. A. C. A.; Lombardo, F.; Dominy, B. W. B. W. B. W. B. W. B. W.; Feeney, P. J. P. J.; Lombardo, F.; Dominy, B. W. B. W. B. W. B. W. B. W.; Feeney, P. J. P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv Drug Deliv Rev* **1997**, *23* (1–3), 3–25. https://doi.org/10.1016/S0169-409X(96)00423-1.

(58) Matsson, P.; Doak, B. C.; Over, B.; Kihlberg, J. Cell Permeability beyond the Rule of 5. *Adv Drug Deliv Rev* **2016**, *101*, 42–61. https://doi.org/10.1016/j.addr.2016.03.013.

(59) Doak, B. C.; Over, B.; Giordanetto, F.; Kihlberg, J. Oral Druggable Space beyond the Rule of 5: Insights from Drugs and Clinical Candidates. *Chem Biol* **2014**, *21* (9), 1115–1142. https://doi.org/10.1016/j.chembiol.2014.08.013.

(60) Moriwaki, H.; Tian, Y. S.; Kawashita, N.; Takagi, T. Mordred: A Molecular Descriptor Calculator. *J Cheminform* **2018**, *10* (1), 4. https://doi.org/10.1186/s13321-018-0258-y.

(61) Raducanu, B.; Dornaika, F. A Supervised Non-Linear Dimensionality Reduction Approach for Manifold Learning. *Pattern Recognit* **2012**, *45* (6), 2432–2444. https://doi.org/10.1016/j.patcog.2011.12.006.

(62) Gálvez, J.; Garcia, R.; Salabert, M. T.; Soler, R. Charge Indexes. New Topological Descriptors. *J Chem Inf Comput Sci* **1994**, *34* (3), 520–525. https://doi.org/10.1021/ci00019a008.

(63) Reddy, A. S.; Kumar, S.; Garg, R. Hybrid-Genetic Algorithm Based Descriptor Optimization and QSAR Models for Predicting the Biological Activity of Tipranavir Analogs for HIV Protease Inhibition. *J Mol Graph Model* **2010**, *28* (8), 852–862. https://doi.org/10.1016/j.jmgm.2010.03.005.

(64) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Methods and Principles in Medicinal Chemistry; Wiley, 2000; Vol. 11. https://doi.org/10.1002/9783527613106.

(65) Bayat, Z.; Movaffagh, J.; Noruzi, S. Development of a Computational Approach to Predict Blood-Brain Permeability on Anti-Viral Nucleoside Analogues. *Russian Journal of Physical Chemistry A* **2011**, *85* (11), 1923–1930. https://doi.org/10.1134/S0036024411110021.

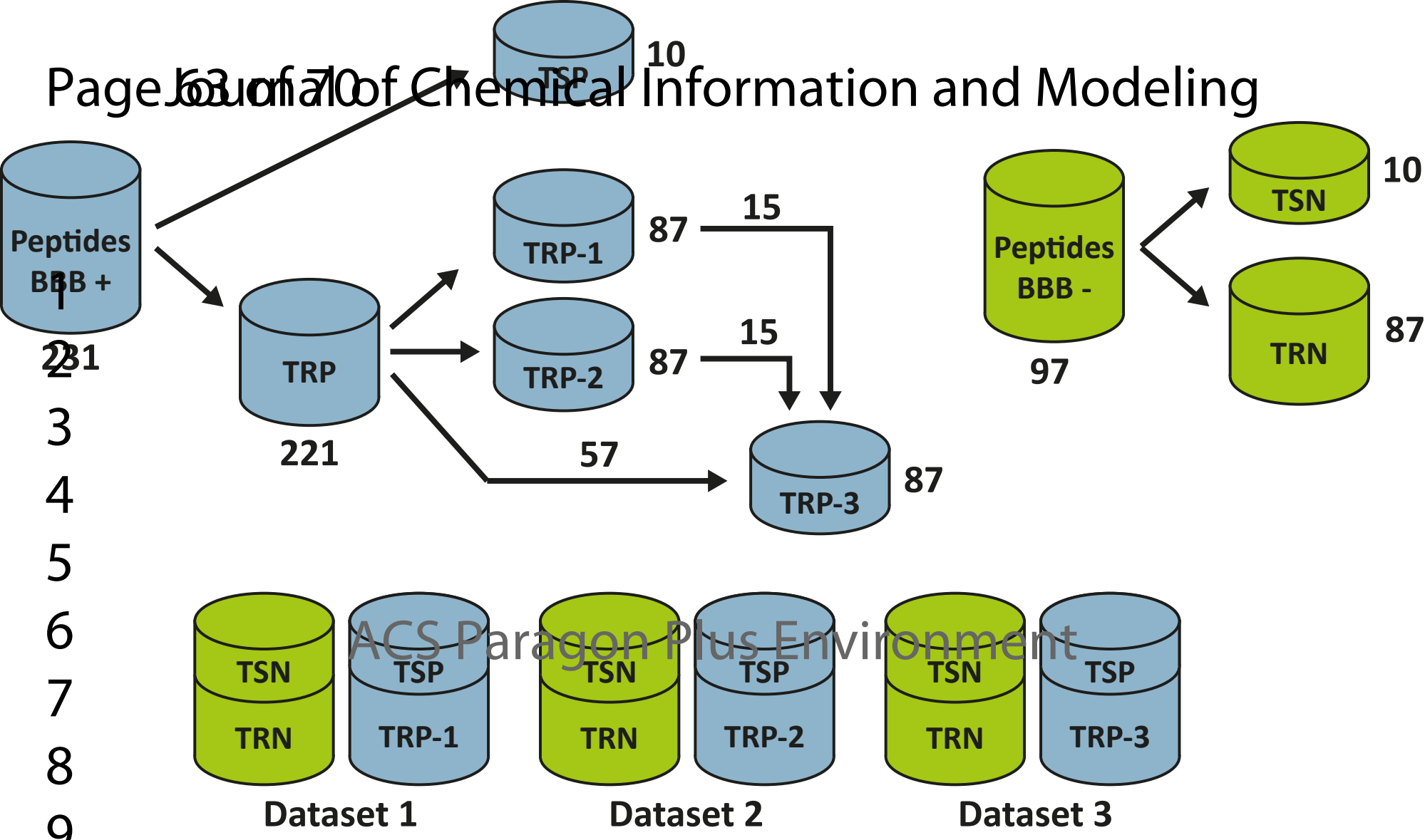
(66) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. A Knowledge-Based Approach in Designing Combinatorial or Medicinal Chemistry Libraries for Drug Discovery. 1. A Qualitative and Quantitative Characterization of Known Drug Databases. *J Comb Chem* **1999**, *1* (1), 55–68. https://doi.org/10.1021/cc9800071.

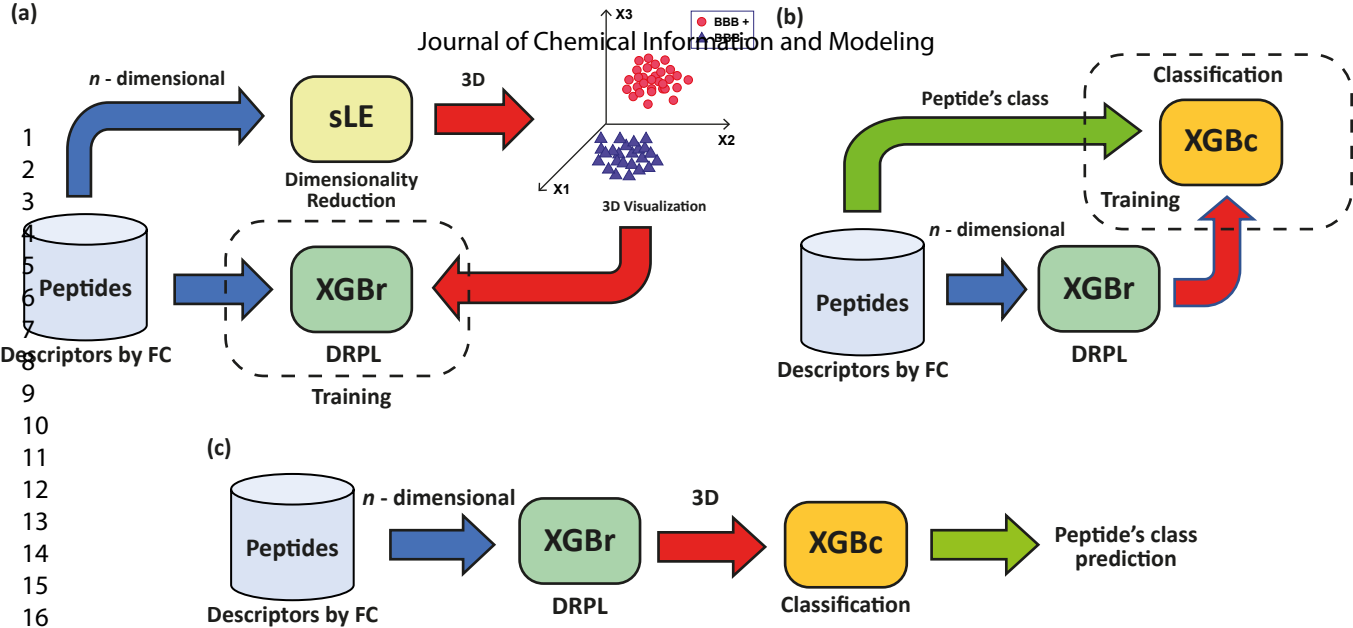
(67) Sohil, F.; Sohali, M. U.; Shabbir, J. *An Introduction to Statistical Learning with Applications in R*; 2022; Vol. 6. https://doi.org/10.1080/24754269.2021.1980261.

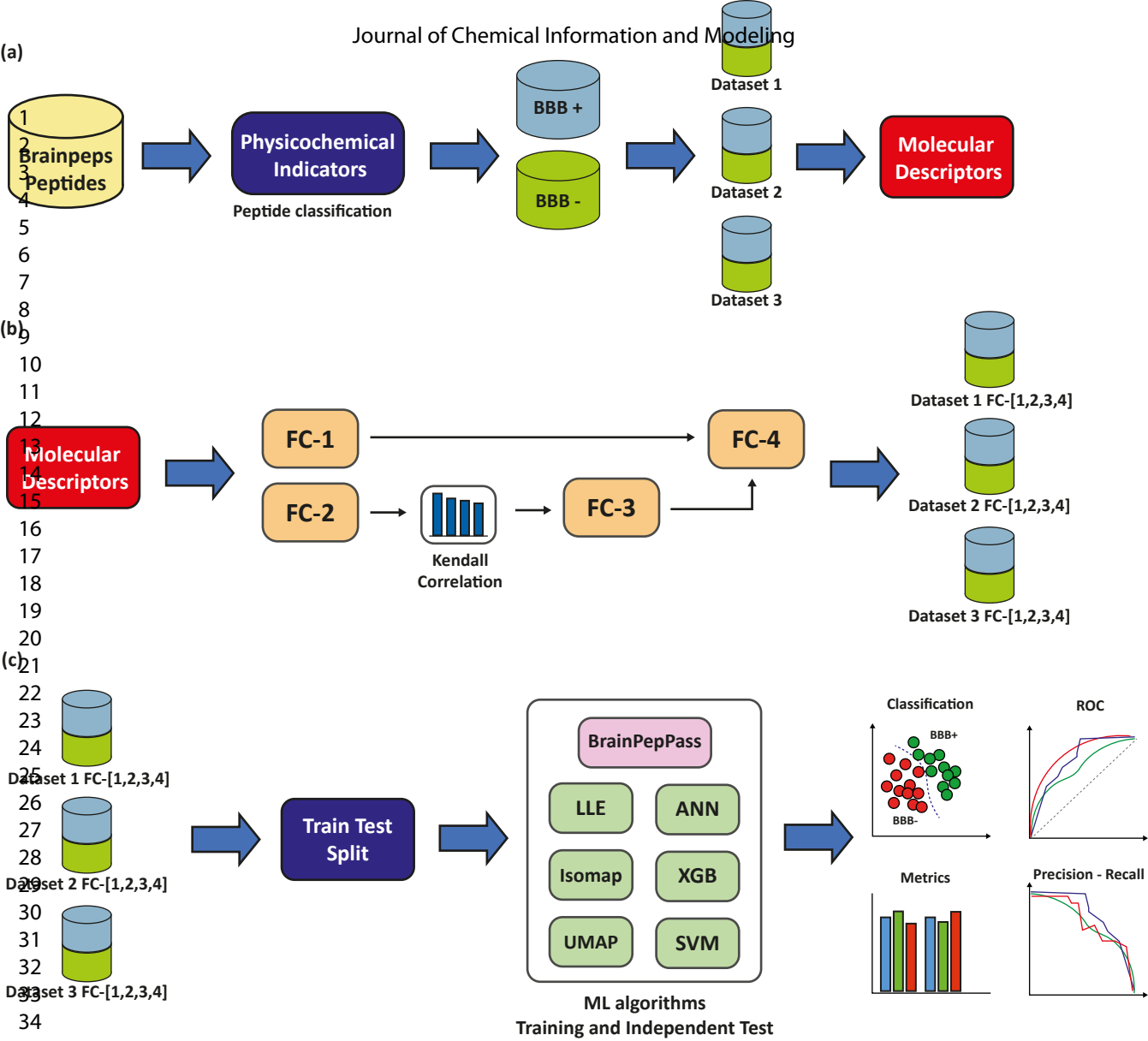
(68) Dai, R.; Zhang, W.; Tang, W.; Wynendaele, E.; Zhu, Q.; Bin, Y.; De Spiegeleer, B.; Xia, J. BBPpred: Sequence-Based Prediction of Blood-Brain Barrier Peptides with Feature Representation Learning and Logistic Regression. *J Chem Inf Model* **2021**, *61* (1), 525–534. https://doi.org/10.1021/acs.jcim.0c01115.

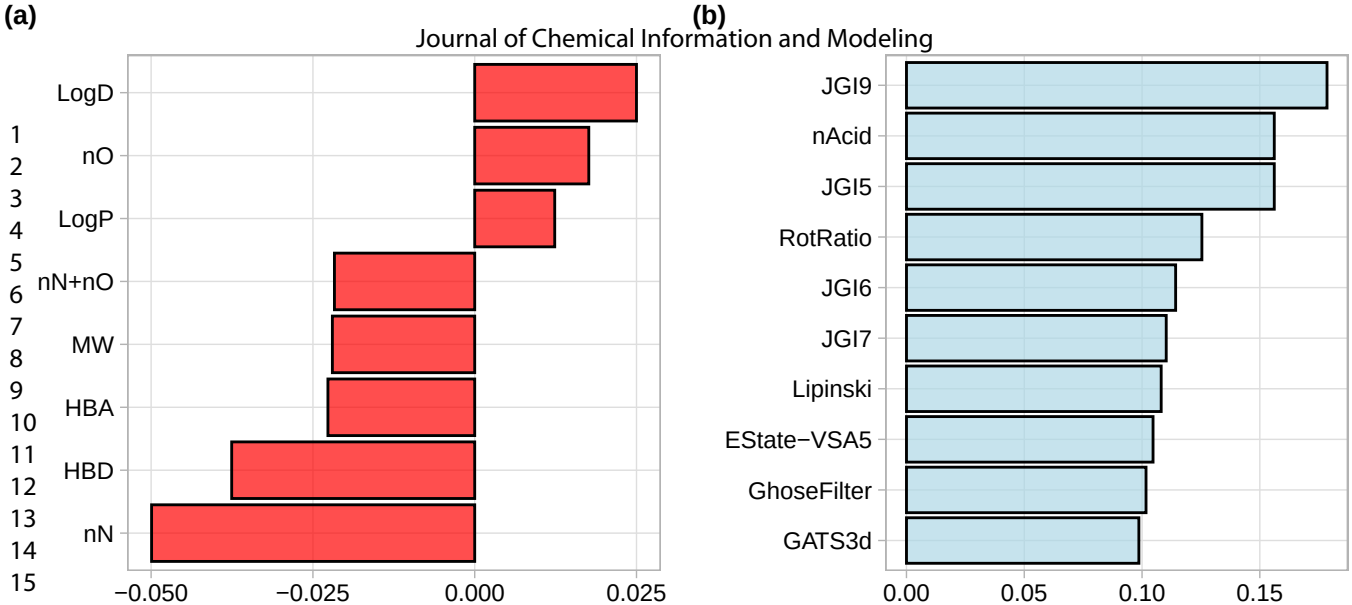
(69) Kumar, V.; Patiyal, S.; Dhall, A.; Sharma, N.; Raghava, G. P. S. B3pred: A Random-Forest-Based Method for Predicting and Designing Blood–Brain Barrier Penetrating Peptides. *Pharmaceutics* **2021**, *13* (8), 1237. https://doi.org/10.3390/pharmaceutics13081237.

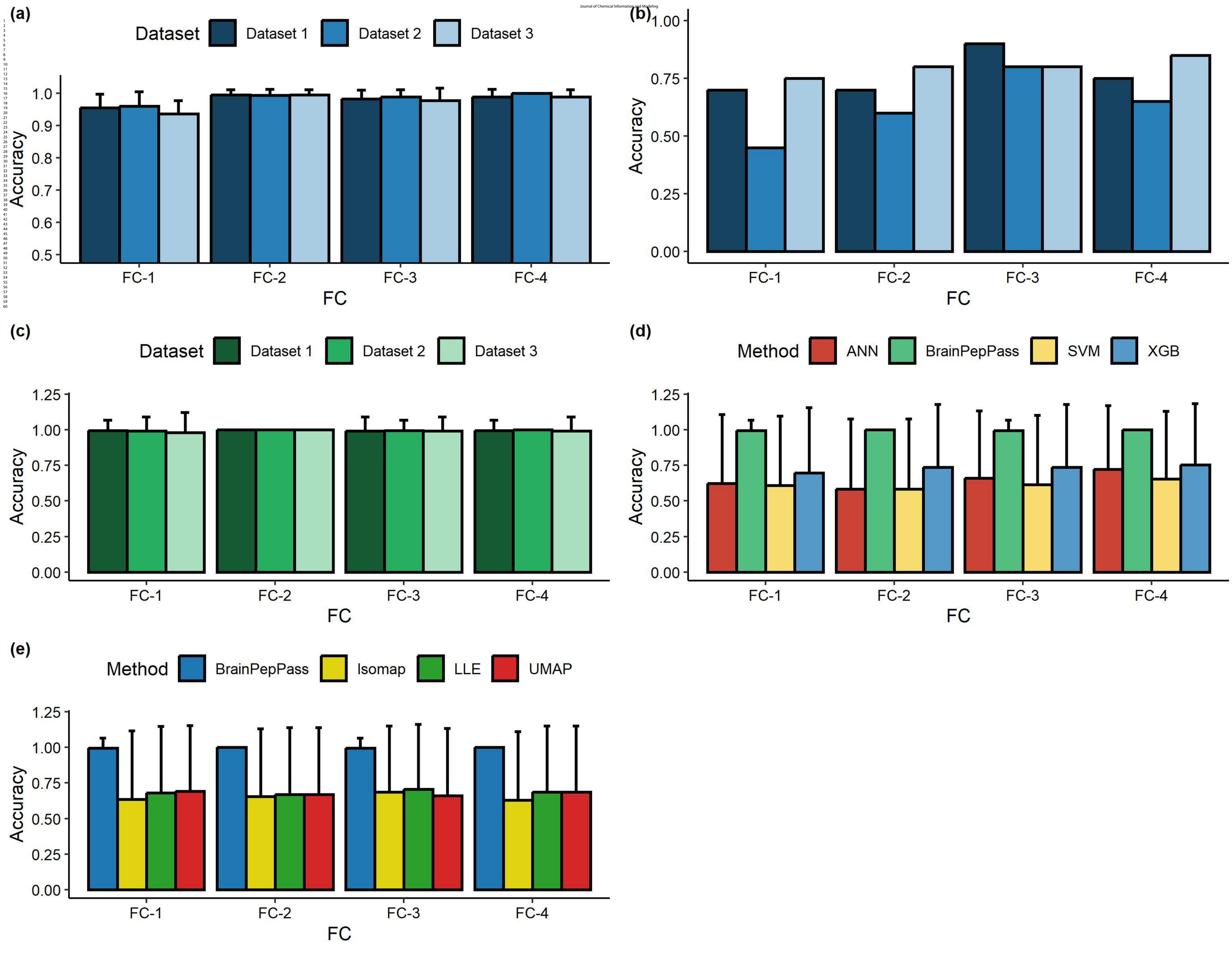
(70) Charoenkwan, P.; Chumnanpuen, P.; Schaduangrat, N.; Lio', P.; Moni, M. A.; Shoombuatong, W. Improved Prediction and Characterization of Blood-Brain Barrier Penetrating Peptides Using Estimated Propensity Scores of Dipeptides. *J Comput Aided Mol Des* **2022**, *36* (11), 781–796. https://doi.org/10.1007/s10822-022-00476-z.

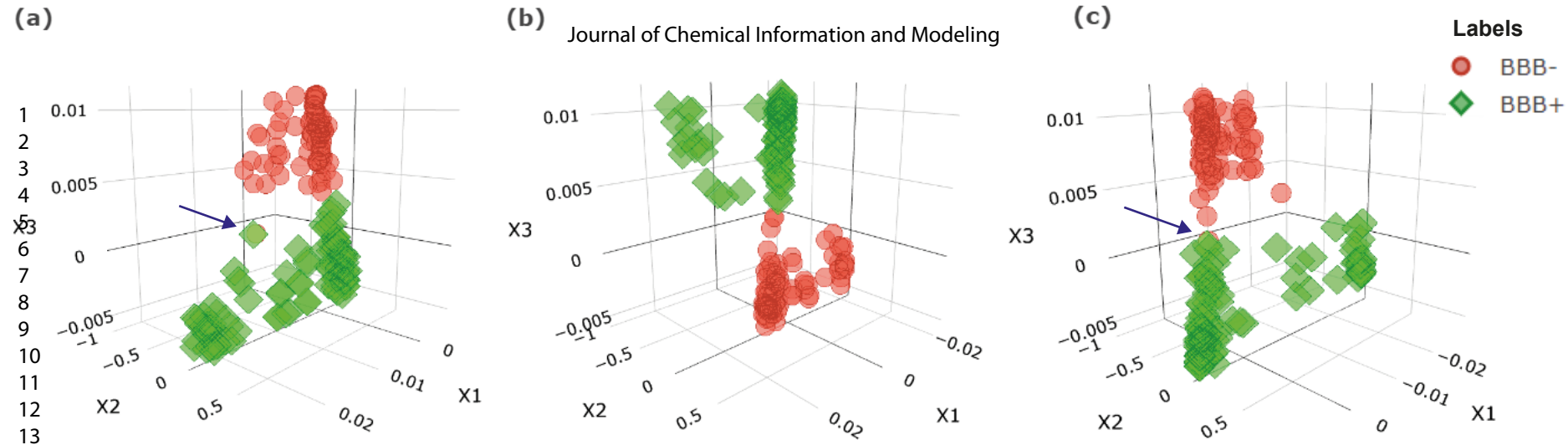


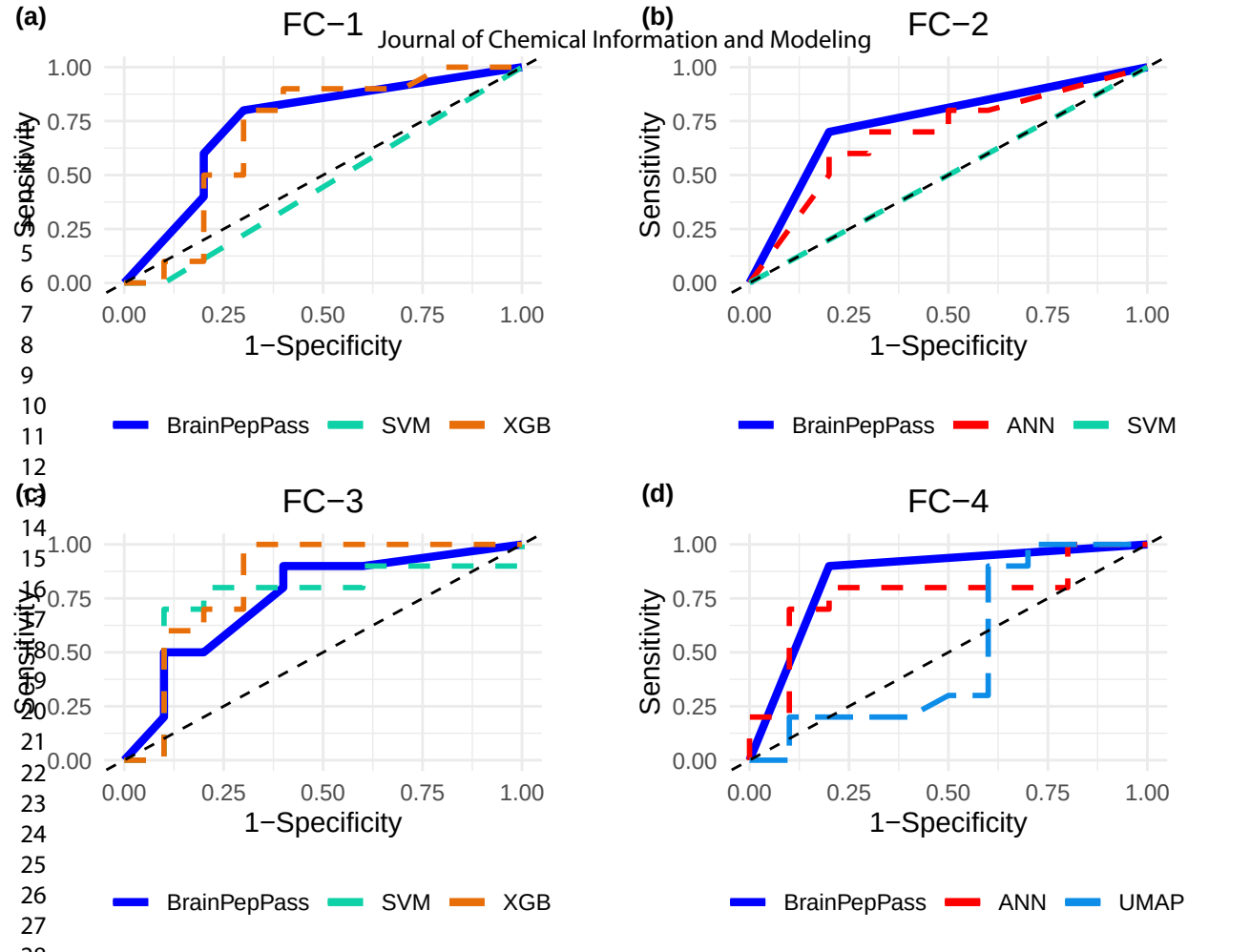

(a)
n - dimensional
sLE
Dimensionality Reduction
3D
X3
X2
X1
BBB +
BBB-
3D Visualization
Peptides
Descriptors by FC
XGBr
DRPL
Training
(b)
Peptide's class
Classification
XGBc
Training
n - dimensional
Peptides
Descriptors by FC
XGBr
DRPL
(c)
Peptides
Descriptors by FC
n - dimensional
XGBr
DRPL
3D
XGBc
Classification
Peptide's class prediction

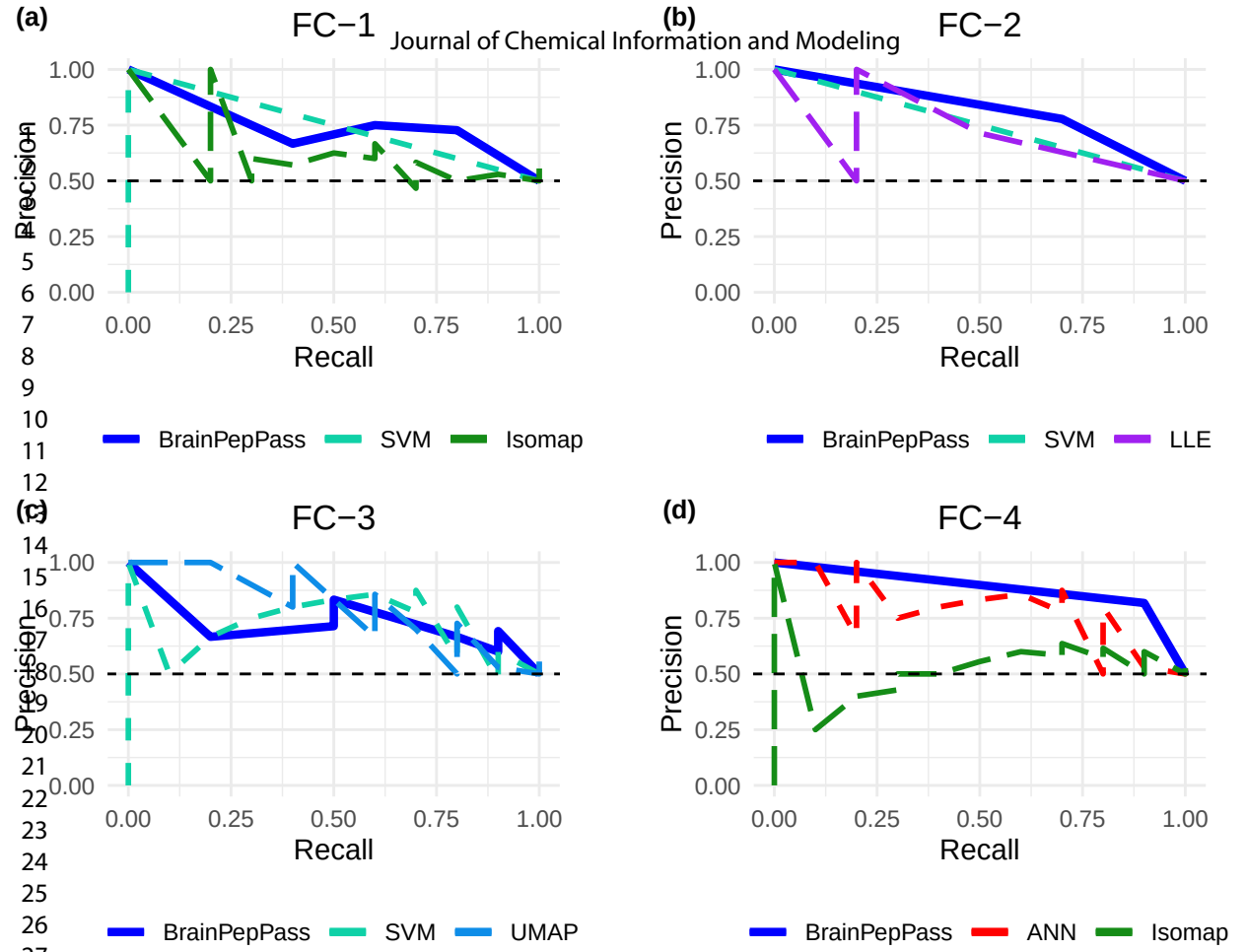

(a)
Brainpeps Peptides
Physicochemical Indicators
Peptide classification
BBB +
BBB -
Dataset 1
Dataset 2
Dataset 3
Molecular Descriptors
(b)
Molecular Descriptors
FC-1
FC-2
FC-3
FC-4
Kendall Correlation
Dataset 1 FC-[1,2,3,4]
Dataset 2 FC-[1,2,3,4]
Dataset 3 FC-[1,2,3,4]
(c)
Dataset 1 FC-[1,2,3,4]
Dataset 2 FC-[1,2,3,4]
Dataset 3 FC-[1,2,3,4]
Train Test Split
BrainPepPass
LLE
ANN
Isomap
XGB
UMAP
SVM
ML algorithms
Training and Independent Test
Classification
BBB+
BBB-
ROC
Metrics
Precision - Recall


(a)
LogD
nO
LogP
nN+nO
MW
HBA
HBD
nN
−0.050
−0.025
0.000
0.025
(b)
JGI9
nAcid
JGI5
RotRatio
JGI6
JGI7
Lipinski
EState−VSA5
GhoseFilter
GATS3d
0.00
0.05
0.10
0.15


(a)
Dataset
Dataset 1
Dataset 2
Dataset 3
Accuracy
1.0
0.9
0.8
0.7
0.6
0.5
FC-1
FC-2
FC-3
FC-4
FC
(b)
Accuracy
1.00
0.75
0.50
0.25
0.00
FC-1
FC-2
FC-3
FC-4
FC
(c)
Dataset
Dataset 1
Dataset 2
Dataset 3
Accuracy
1.25
1.00
0.75
0.50
0.25
0.00
FC-1
FC-2
FC-3
FC-4
FC
(d)
Method
ANN
BrainPepPass
SVM
XGB
Accuracy
1.25
1.00
0.75
0.50
0.25
0.00
FC-1
FC-2
FC-3
FC-4
FC
(e)
Method
BrainPepPass
Isomap
LLE
UMAP
Accuracy
1.25
1.00
0.75
0.50
0.25
0.00
FC-1
FC-2
FC-3
FC-4
FC


(a)
(b)
(c)
Labels
BBB-
BBB+
X1
X2
X3
0.01
0.005
0
−0.005
−1
−0.5
0
0.5
0.02
0.01
0
−0.01
−0.02


(a)
FC−1
Sensitivity
1−Specificity
0.00
0.25
0.50
0.75
1.00
BrainPepPass
SVM
XGB
(b)
FC−2
Sensitivity
1−Specificity
BrainPepPass
ANN
SVM
(c)
FC−3
Sensitivity
1−Specificity
BrainPepPass
SVM
XGB
(d)
FC−4
Sensitivity
1−Specificity
BrainPepPass
ANN
UMAP


(a)
FC−1
Precision
Recall
1.00
0.75
0.50
0.25
0.00
BrainPepPass
SVM
Isomap
(b)
FC−2
Precision
Recall
BrainPepPass
SVM
LLE
(c)
FC−3
Precision
Recall
BrainPepPass
SVM
UMAP
(d)
FC−4
Precision
Recall
BrainPepPass
ANN
Isomap