

FEDERAL UNIVERSITY OF PARÁ
INSTITUTE OF TECHNOLOGY
GRADUATE PROGRAM IN ELECTRICAL ENGINEERING

**DEVELOPMENT OF MACHINE LEARNING-BASED FRAMEWORKS TO
PREDICT PERMEABILITY OF PEPTIDES THROUGH CELL MEMBRANE AND
BLOOD-BRAIN BARRIER**

EWERTON CRISTHIAN LIMA DE OLIVEIRA

TD: 05/2024

UFPA / ITEC / PPGEE
Guamá University Campus
Belém - Pará - Brasil
2024

FEDERAL UNIVERSITY OF PARÁ
INSTITUTE OF TECHNOLOGY
GRADUATE PROGRAM IN ELECTRICAL ENGINEERING

EWERTON CRISTHIAN LIMA DE OLIVEIRA

**DEVELOPMENT OF MACHINE LEARNING-BASED FRAMEWORKS TO
PREDICT PERMEABILITY OF PEPTIDES THROUGH CELL MEMBRANE AND
BLOOD-BRAIN BARRIER**

Ph.D. Thesis submitted to the Examining Board
of the Graduate Program in Electrical Engineer-
ing from the Federal University of Pará as par-
tial requirement to obtain the Ph.D. Degree in
Electrical Engineering, Area of Concentration
in Applied Computing

UFPA / ITEC / PPGEE
Guamá University Campus
Belém - Pará - Brasil
2024

Dados Internacionais de Catalogação na Publicação (CIP) de acordo com ISBD
Sistema de Bibliotecas da Universidade Federal do Pará
Gerada automaticamente pelo módulo Ficat, mediante os dados fornecidos pelo(a)
autor(a)

- O48d Oliveira, Ewerton Cristhian Lima de.
DEVELOPMENT OF MACHINE LEARNING-BASED
FRAMEWORKS TO PREDICT PERMEABILITY OF
PEPTIDES THROUGH CELL MEMBRANE AND BLOOD-
BRAIN BARRIER / Ewerton Cristhian Lima de Oliveira. —
2024.
xvii, 125 f. : il. color.
- Orientador(a): Prof. Dr. Claudomiro de Souza de Sales
Junior
Coorientador(a): Prof. Dr. Anderson Henrique Lima E
Lima
Tese (Doutorado) - Universidade Federal do Pará,
Instituto de Tecnologia, Programa de Pós-Graduação em
Engenharia Elétrica, Belém, 2024.
1. Peptídeos. 2. Biomembranas. 3. B3PPs. 4.
Framework. 5. Aprendizagem de Máquina. I. Título.


**“DEVELOPMENT OF MACHINE LEARNING-BASED FRAMEWORKS TO
PREDICT PERMEABILITY OF PEPTIDES THROUGH CELL MEMBRANE AND
BLOOD-BRAIN BARRIER ”**

AUTOR: EWERTON CRISTHIAN LIMA DE OLIVEIRA


TESE DE DOUTORADO SUBMETIDA À BANCA EXAMINADORA APROVADA PELO
COLEGIADO DO PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA, SENDO
JULGADA ADEQUADA PARA A OBTENÇÃO DO GRAU DE DOUTOR EM ENGENHARIA
ELÉTRICA NA ÁREA DE COMPUTAÇÃO APLICADA.

APROVADA EM: 27/03/2024


BANCA EXAMINADORA:

Documento assinado digitalmente
 CLAUDOMIRO DE SOUZA DE SALES JÚNIOR
Data: 29/04/2024 13:10:32-0300
Verifique em <https://validar.iti.gov.br>


Prof. Dr. Claudomiro de Souza de Sales Júnior
(Orientador – PPGEE/UFPA)

Documento assinado digitalmente
 ANDERSON HENRIQUE LIMA E LIMA
Data: 29/04/2024 09:46:29-0300
Verifique em <https://validar.iti.gov.br>


Prof. Dr. Anderson Henrique Lima e Lima
(Coorientador - PPGQ/UFPA)

Documento assinado digitalmente
 ADRIANA ROSA GARCEZ CASTRO
Data: 01/05/2024 07:18:08-0300
Verifique em <https://validar.iti.gov.br>


Prof. Dr. Adriana Rosa Garcez Castro
(Avaliadora Interna - PPGEE/UFPA)

Documento assinado digitalmente
 ALDEBARO BARRETO DA ROCHA KLAUTAU JÚNIOR
Data: 20/05/2024 13:27:23-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. Aldebaro Barreto da Rocha Klautau Júnior
(Avaliador Interno - PPGEE/UFPA)

Documento assinado digitalmente
 JERONIMO LAMEIRA SILVA
Data: 30/04/2024 13:49:05-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. Jerônimo Lameira Silva
(Avaliador Externo ao Programa - PPGQ/UFPA)

Documento assinado digitalmente
 EDUARDO KREMPSEK DA SILVA
Data: 29/04/2024 14:01:09-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. Eduardo Krempser da Silva
(Avaliador Externo - FIOCRUZ)

VISTO:

Prof. Dr. Diego Lisboa Cardoso
(Coordenador do PPGEE/ITEC/UFPA)

ACKNOWLEDGMENTS

Firstly, I would like to thank God for all the help and inspiration throughout my life. I would also like to thank the following people and institutions who contributed to this work.

To my parents Josy and Eliezer, my grandmother Leia, and my siblings Sarah and Renan for all their support and help during my life, and all credulity on my work.

To my friends Lucas Vinicius, Daniel Victor, Marco Loureiro, Sandio Maciel, Igor Falcão, Caio Flexa, André Rosário, Roberto Xavier, Juan Vidal, Luana Gonçalves, Eduardo Carvalho, Rafael Rocha, Nikolas Carneiro, Renato Oliveira, Fabrício Oliveira, and Vitor Cirilo for all their motivation and collaboration during my doctoral period.

To the researchers Igor Matheus, Gabriel Aragão, Luiz Patrick, Juliana Auzier, Hannah Hirmz, and Evelien Wynendaele who contributed so much to the progress of this work.

I thank Prof. Dr. Kauê Santana for all his help, collaboration, inspiration, and orientation in executing this work.

I would like to thank Prof. Dr. Bart De Spiegeleer and his research team for their collaboration in providing the data from Brainpeps.

I thank Prof. Dr. Ronnie Alves and Profa. Dra. Renata Tedeschi for the opportunity to demonstrate and apply my knowledge in data science in other fields of science.

I thank Professors Dr. Claudomiro Sales and Dr. Anderson Lima, my doctoral advisors, for all their help, collaboration, orientation, experiences, and friendship.

I thank the support of the Coordination for the Improvement of Higher Education Personnel (CAPES).

ABSTRACT

Peptides comprise a versatile class of biomolecules with diverse physicochemical and structural properties, in addition to numerous pharmacological and biotechnological applications. Some groups of peptides can cross biological membranes, such as the cell membrane and the human blood-brain barrier. Researchers have explored this property over the years as an alternative to developing more powerful drugs, given that some peptides can also be drug carriers. Although some machine learning-based tools have been developed to predict cell-penetrating peptides (CPPs) and blood-brain barrier penetrating peptides (B3PPs), some points have not yet been explored within this theme. These points encompass the use of dimensionality reduction (DR) techniques in the preprocessing stage, molecular descriptors related to drug bioavailability, and data structures that encode peptides with chemical modifications. Therefore, the primary purpose of this thesis is to develop and test two frameworks based on DR, the first one to predict CPPs and the second to predict B3PPs, also evaluating the molecular descriptors and data structure of interest. The results of this thesis show that for the prediction of penetration in the cell membrane, the proposed framework reached 92% accuracy in the best performance in an independent test, outperforming other tools created for the same purpose, besides evidencing the contribution between the junction of molecular descriptors based on amino acid sequence and those related to bioavailability and cited in Lipinski's rule of five. Furthermore, the prediction of B3PPs by the proposed framework reveals that the best model using structural, electric, and bioavailability-associated molecular descriptors achieved average accuracy values exceeding 93% in the 10-fold cross-validation and between 75% and 90% accuracy in the independent test for all simulations, outperforming other machine learning (ML) tools developed to predict B3PPs. These results show that the proposed frameworks can be used as an additional tool in predicting the penetration of peptides in these two biomembranes and are available as free-to-use web servers.

Key-words: Peptides, Biomembranes, CPPs, B3PPs, Framework, Machine Learning.

RESUMO

Peptídeos compreendem uma classe versátil de biomoléculas com diversas propriedades físico-químicas e estruturais, além de inúmeras aplicações farmacológicas e biotecnológicas. Alguns grupos de peptídeos podem cruzar membranas biológicas, como a membrana celular e a barreira hematoencefálica humana. Pesquisadores tem explorado esta propriedade ao longo dos anos como uma alternativa ao desenvolvimento de novos medicamentos mais poderosos, tendo em vista que alguns peptídeos são carreadores de fármacos. Embora existam ferramentas baseadas em aprendizado de máquina desenvolvidas para prever *cell-penetrating peptides* (CPPs) e *blood-brain barrier penetrating peptides* (B3PPs), alguns pontos ainda não foram explorados dentro deste tema. Estes pontos abrangem o uso de técnicas de redução de dimensionalidade (RD) na etapa de pré-processamento, de descritores moleculares relacionados à biodisponibilidade de drogas, e de estrutura de dados que codificam peptídeos com modificações químicas. Portanto, a proposta principal desta tese é desenvolver e testar dois *frameworks* baseados em RD, o primeiro para prever CPPs e o segundo para prever B3PPs, avaliando também os descritores moleculares e estrutura de dados de interesse. Os resultados desta tese mostram que para a predição de penetração na membrana celular, o *framework* proposto atingiu 92% de acurácia no melhor desempenho em um teste independente, superando outras ferramentas criadas para o mesmo propósito, além de evidenciar a contribuição entre a junção de descritores baseado em sequência de aminoácidos e os relacionados a biodisponibilidade e citados na regra dos cinco de Lipinski. Além do mais, a predição de B3PPs pelo *framework* proposto revela que o melhor modelo que utiliza descritores moleculares estruturais, elétricos e associados a biodisponibilidade de compostos alcançou valores que superam 93% de acurácia média no 10-fold cross-validation e acurácia entre 75% e 90% no teste independente para todas as simulações, superando outras ferramentas de machine learning (ML) desenvolvidas para prever B3PPs. Estes resultados mostram que os *frameworks* propostos podem ser usado como ferramenta adicional na predição de penetração de peptídeos através dessas duas biomembranas e estão disponíveis como web servers gratuitos para uso.

Palavras-chaves: Peptídeos, Biomembranas, CPPs, B3PPs, Framework, Aprendizado de Máquina.

LIST OF FIGURES

Figure 1 – Amino acid structure.	32
Figure 2 – Peptide bond schematic.	33
Figure 3 – Schematic representation of the N- and C-terminal in a Glycylalanine peptide.	33
Figure 4 – FASTA file example containing the header and the sequence of peptides Opiorphin, Neurotensin, Arginine vasopressin 1-7, and PepH1.	34
Figure 5 – PDB file of Transportan peptide. A) PDB file containing information about the arrangement of atoms in 3D space. B) Visualization of the 3D confor- mation of Transportan using the PDB file.	35
Figure 6 – MDL file of Transportan peptide.	35
Figure 7 – Schematic representation of cell membrane showing its main chemical li- pidic and protein components. surface.	36
Figure 8 – Schematic representation of endocytosis and direct penetration mechanisms for cell-penetrating peptide internalization.	37
Figure 9 – Schematic representation of the blood-brain barrier, showing its main cell components (pericytes, astrocytes, and endothelial cells) and localization in the brain capillary wall.	38
Figure 10 – Schematic representation of penetration mechanisms for peptides internal- ization across the BBB.	39
Figure 11 – Representation of main categories of problems approached by ML, such as classification, regression, clustering, and dimensionality reduction.	44
Figure 12 – Schematic of dimensionality reduction process and low-dimensional data visualization.	44
Figure 13 – Schematic of the artificial neuron.	45
Figure 14 – Schematic of a MLP architecture.	46
Figure 15 – Example of non-linearly separable samples.	49

Figure 16 – Example of SVM using different kernels to classify non-linearly separable data. Left: SVM with a polynomial kernel of degree 3. Right: SVM with RBF kernel.	50
Figure 17 – Schematic of XGBoost algorithm.	51
Figure 18 – General structure of voting classifier.	53
Figure 19 – Schematic representation of sample-based neighbor computation in sLE. . .	55
Figure 20 – Diagram of CPPs and non-CPPs datasets construction. The first step represents the unification of CPPs and non-CPPs in their respective groups. The second step is the extraction of molecular descriptors	58
Figure 21 – Diagram of B3PPs and non-B3PPs datasets construction based on the employment of experimental pharmacokinetic indicators regarding the penetration of peptides into BBB to classify them into BBB+ or BBB-.	61
Figure 22 – 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.	62
Figure 23 – Diagram of molecular descriptors extraction to compose the four FCs in cell membrane case study using RDKit, PyBioMed, and Biopython packages. . .	64
Figure 24 – Fluxogram of dimensionality reduction and pattern learning stages in the proposed framework. The DR stage represents the projection of n -dimensional data to 3D using sLE algorithm. The pattern learning stage encompasses the use of XGBr to learn and generalize the sLE projection for new data.	66
Figure 25 – Proposed framework for CPPs and B3PPs prediction. Panel A) shows the stage of XGBc training with 3D data as input and peptides' labels as output. B) Illustrates the final pipeline of the proposed framework.	67
Figure 26 – Structure of voting classifier to predict peptides' biomembrane penetration. .	68
Figure 27 – Barplot of accuracy from 10-fold cross-validation of DPF-CPPred (purple), Vcf-CPP (orange), ANN (red), GPC (blue), and SVM (green) using PDB format.	70
Figure 28 – The 3D plot of the reduced PDB training dataset by the DPF-CPPred in each FC.	71

Figure 29 – Barplot of accuracy from 10-fold cross-validation of DPF-CPPred (purple), Vcf-CPP (orange), ANN (red), GPC (blue), and SVM (green) using FASTA format.	72
Figure 30 – The 3D plot of the reduced FASTA training dataset by the DPF-CPPred in each FC.	73
Figure 31 – Accuracy of ANN (red), GPC (blue), SVM (green), Vcf-CPP (orange), and DPF-CPPred (purple) by FCs evaluated in the independent test for PDB data format.	74
Figure 32 – Accuracy of ANN (red), GPC (blue), SVM (green), Vcf-CPP (orange), and DPF-CPPred (purple) by FCs evaluated in the independent test for FASTA data format.	76
Figure 33 – Barplot of accuracy achieved by DPF-3BPPred in 10-fold cross-validation analysis for all FCs and datasets.	79
Figure 34 – Barplot of accuracy achieved by the best model of each method in 10-fold cross-validation analysis among all FCs and datasets.	80
Figure 35 – LOOCV analysis employed in DPF-3BPPred.	83
Figure 36 – Dimensionality reduction result of BrainPepPass in pattern learning stage for FC-4. (a) Dataset 1. (b) Dataset 2. (c) Dataset 3.	84
Figure 37 – Home screen of the BChemRF-CPPred web server.	129
Figure 38 – Screen for uploading peptide files for permeability prediction.	129
Figure 39 – Home screen of the BrainPepPass web application in GitHub.	130
Figure 40 – Home screen of the BrainPepPass web application in Google Colab.	131

LIST OF TABLES

Table 1 – Dataset division of peptide samples according to cell membrane permeability and file format.	59
Table 2 – Feature composition for CPPs prediction analysis.	63
Table 3 – Feature composition for 3BPPs prediction analysis.	65
Table 4 – Comparison of accuracy, sensitivity, specificity, F1-score, and MCC obtained for ANN, GPC, SVM, Vcf-CPP, and DPF-CPPred in the independent test using FC-4 and PDB format.	75
Table 5 – Comparison of the performance of DPF-CPPred frameworks that used only natural peptides in the independent test. The comparison was performed between the frameworks based on the four feature compositions (FC-1 to FC-4) that use FASTA as input with the framework based on the FC-4 that uses the PDB as input.	76
Table 6 – Comparison of the performance of previous ML-based tools (MLCPP, CPPred-RF, and SkipCPP-Pred), FC-2 based Vcf-CPP, and FC-3 based DPF-CPPred using only natural peptides from the independent dataset (1 st experiment); as well as, the evaluation of the performance of Kelm-CPPpred and FC-4 based DPF-CPPred and Vcf-CPP from all independent dataset (2 nd experiment). . .	77
Table 7 – Independent test analysis for the best DPF-3BPPred, ANN, SVM, GPC, and Vcf-3BPP models by each FC.	82
Table 8 – Analysis of independent test comparing DPF-3BPPred and DPF-3BPPred-N with BBPpred, BBPpredict, and SCMB3PP algorithms using natural peptides.	85
Table 9 –	138

LIST OF ABBREVIATIONS AND ACRONYM

AI	<i>Artificial Intelligence</i>
AAC	<i>Amino Acid Composition</i>
AUC	<i>Area Under the Curve</i>
ANN	<i>Artificial Neural Network</i>
BBB	<i>Blood-Brain Barrier</i>
B3PPs	<i>Blood-Brain Barrier Penetrating Peptides</i>
CPPs	<i>Cell-Penetrating Peptides</i>
CNS	<i>central nervous system</i>
DR	<i>Dimensionality Reduction</i>
DT	<i>Decision Tree</i>
DPC	<i>Dipeptide Composition</i>
DPF-CPPred	<i>Dimensionality Reduction and Pattern Learning framework for CPP Prediction</i>
DPF-3BPPred	<i>Dimensionality Reduction and Pattern Learning Framework for B3PP Prediction</i>
DPF-3BPPred-N	<i>Dimensionality Reduction and Pattern Learning Framework for Natural B3PP Prediction</i>
FASTA	<i>Text-based Format for Representing Nucleotide or Amino Acid Sequences</i>
FC	<i>Feature Composition</i>
f(Arg)	<i>Fraction of Arginine Residues</i>
f(Lys)	<i>Fraction of Lysine Residues</i>
Fsp³	<i>Fraction of sp³-Hybridized Carbon Atoms</i>
GPC	<i>Gaussian Process Classifier</i>
HBA	<i>Hydrogen Bond Acceptors</i>

HBD	<i>Hydrogen Bond Donors</i>
high-d	<i>High dimensional</i>
kNN	<i>k-Nearest Neighbors</i>
low-d	<i>Low dimensional</i>
LogD	<i>Water-Octanol Distribution Coefficient</i>
LogP	<i>Water-Octanol Partition Coefficient</i>
LOOCV	<i>leave-one-out cross-validation</i>
LR	<i>Logistic regression</i>
ML	<i>Machine Learning</i>
MLP	<i>Multilayer Perceptron</i>
MCC	<i>Matthews Correlation Coefficient</i>
MDL	<i>Molecular Data File</i>
MW	<i>Molecular Weight</i>
NAR	<i>Number of Aromatic Rings</i>
NetC	<i>Net Charge</i>
NG	<i>Number of Guanidinium Groups</i>
NNCAA	<i>Number of Negatively Charged Amino Acid at pH = 7.4</i>
nN	<i>Nitrogen Count</i>
nO	<i>Oxygen Count</i>
n(N+O)	<i>Nitrogen and Oxygen Count</i>
NPA	<i>Number of Primary Amino Groups</i>
NRB	<i>Number of Rotable Bonds</i>
PCA	<i>Principal Component Analysis</i>
PDB	<i>Protein Database structure</i>

PseAAC	<i>Pseudo-Amino Acid Composition</i>
RBF	<i>Radial Basis Function</i>
RF	<i>Random Forest</i>
RO5	<i>Lipinski's Rule of Five</i>
sLE	<i>Supervised Laplacian Eigenmaps</i>
SVM	<i>Support Vector Machine</i>
TPSA	<i>Topological Polar Surface Area</i>
Vcf	<i>Voting Classifier</i>
Vcf-CPP	<i>Voting Classifier for CPPs Prediction</i>
Vcf-3BPP	<i>Voting Classifier for B3PPs Prediction</i>
XGBoost	<i>Extreme Gradient Boosting</i>
XGBr	<i>Extreme Gradient Boosting Regression</i>
XGBc	<i>Extreme Gradient Boosting Classifier</i>

CONTENTS

1	INTRODUCTION	17
1.1	The challenge and the importance of peptide's permeability prediction .	17
1.2	Motivation	19
1.3	Related works	20
1.3.1	Review of ML application in CPPs prediction	20
1.3.2	Review of ML application in B3PPs prediction	24
1.3.3	Review of dimensionality reduction application in computational chemistry .	26
1.3.4	Review of bioavailability properties in peptide	28
1.4	Contributions	28
1.5	Objectives	30
1.6	Thesis Organization	31
2	THEORETICAL BACKGROUND	32
2.1	On biochemical and computational aspects of peptides	32
2.1.1	Biochemical aspects of peptides	32
2.1.2	Computational coding of peptides	34
2.2	Cell membrane and cell-penetrating peptides	36
2.3	Blood-brain barrier and blood-brain barrier penetrating peptides . . .	37
2.4	Molecular descriptors and their influence on biomembranes uptake . . .	38
2.5	Machine learning algorithms	42
2.5.1	Dimensionality reduction	44
2.5.2	Artificial neural network	45

2.5.3	<i>Gaussian</i> process classifier	47
2.5.4	Support vector machine	48
2.5.5	Extreme gradient boosting	50
2.5.6	Voting classifier based on machine learning	52
2.5.7	Supervised Laplacian eigenmaps	53
2.6	Conclusion	56
3	PROPOSED METHOD	57
3.1	Peptides databases	57
3.1.1	Database for CPPs	57
3.1.2	Database for B3PPs	59
3.2	Molecular Descriptors	62
3.2.1	Molecular descriptors for CPPs prediction	62
3.2.2	Molecular descriptors for B3PPs prediction	64
3.3	Proposed framework to predict CPPs and B3PPs	65
3.4	Voting classifier to predict CPPs and B3PPs	67
3.5	Conclusion	68
4	RESULTS AND DISCUSSIONS	69
4.1	Prediction of CPPs	69
4.1.1	Cross-validation analysis in CPPs prediction	70
4.1.1.1	Cross-validation analysis for PDB encoding	70
4.1.1.2	Cross-validation analysis for FASTA encoding	72
4.1.2	Independent test analysis in CPPs prediction	74
4.2	Prediction of B3PPs	78
4.2.1	Cross-validation analysis in B3PPs prediction	78

4.2.2	Independent test analysis in B3PPs prediction	81
4.2.3	Leave-one-out cross-validation analysis	83
4.2.4	Performance comparison with web servers in B3PPs prediction	85
4.3	Conclusion	86
5	CONCLUSIONS AND FUTURE WORKS	87
5.1	Final remarks	87
5.2	Future works	89
5.3	Publications	91
	REFERENCES	92
A	APPENDIX A	111
A.1	Training dataset of CPPs and non-CPPs	111
A.2	Test dataset of CPPs and non-CPPs	114
B	APPENDIX B	116
C	APPENDIX C	117
D	APPENDIX D	123
E	APPENDIX E	127
F	APPENDIX F	128
F.1	Appendix F1	128
F.2	Appendix F2	128
G	APPENDIX G	129
G.1	BChemRF-CPPred web server	129
G.2	BrainPepPass web application	130
H	APPENDIX H	132

I	APPENDIX I	134
J	APPENDIX J	135
K	APPENDIX K	136
L	APPENDIX L	137
M	APPENDIX M	138
N	APPENDIX N	139
O	APPENDIX N	140
P	APPENDIX O	141

1 INTRODUCTION

1.1 The challenge and the importance of peptide's permeability prediction

Penetration into biological membranes is a desired characteristic for several bioactive molecules to reach the target site related to their molecular mode of action (DAINA; ZOETE, 2016; DOAK et al., 2014). The selective control through biomembranes has protected living organisms against the undesired and harmful effects of other organisms and exogenous molecules. However, these barriers have also been the main challenge to developing new potent compounds with therapeutic activity, and several strategies have been developed to overcome this obstacle (AHLAWAT et al., 2020).

Peptides are a class of bioactive molecules applied in several biological functions that contribute to human health. Since the discovery of oxytocin as a therapeutic agent in 1953, the biotech and pharmaceutical industries have invested time and money in discovering and developing new peptides with therapeutic effects (BAIG et al., 2018). Nowadays, there is a range of peptides with therapeutic effects, such as antioxidant (e.g., β -alanyl-L-histidine), antimicrobial (e.g., Defensins, Dermicidin, Melittin, and LL-37) (SONG; GROOT; SANSOM, 2019; MEMARIANI; MEMARIANI, 2019; NAGAOKA; TAMURA; REICH, 2020), and inhibitors for some neurodegenerative disorders (NDs) as Alzheimer's disease (e.g. Neurotrophins, Vasoactive Intestinal Peptides, A β (16-20) KLVFF, and Humanin) and Parkinson's disease (e.g., NAP, Neurotrophins, and Vasoactive Intestinal Peptides) (BAIG et al., 2018; ICHIM; TAUSZIG-DELAMASURE; MEHLEN, 2012; MATSUOKA, 2011). The use of peptides has become a crucial alternative to treat NDs since traditional therapies with commercial drugs have no high efficacy in crossing the blood-brain barrier (BBB), which hampers the transport of some molecules from blood vessels to brain parenchyma (DAI et al., 2021; ZHOU; SMITH; LIU, 2021).

Concerning small molecule drugs, the peptides have some advantages when used as a therapeutic agent, such as high biological activity, high specificity, and better membrane permeability, having the ability to enter into eukaryotic cells in a non-disruptive way (KUMAR; AGRAWAL, et al., 2018). Furthermore, peptides have the ability of cargo delivery, i.e., they can be used as drug carriers, which is a great strategy to overcome one of the most significant problems in drug development: the uptake of drugs through biological membranes (KARDANI et al., 2019).

Cell-penetrating peptides (CPPs) are molecules capable of crossing the cell membrane and achieving its interior and can be used to transport drugs, nucleic acids, and nanoparticles. They are essential to fulfill therapeutic effects against several diseases (MANAVALAN et al., 2018; LEE; HARRIS, et al., 2019). The prediction of new CPPs aided by artificial in-

telligence (AI) algorithms has become the aim of several pharmaceutical and biotechnological researchers due to their high throughput and low cost in the screening of large databases (LEE; HARRIS, et al., 2019; RÖCKENDORF; NEHLS; GUTSMANN, 2022). Some machine learning (ML) algorithms were designed to predict uptake across cell membranes using both primary structure (FASTA) or tertiary structure (PDB) of peptides as possibilities. FASTA format consists of a single-line description of a molecule based on a sequence of amino acid and nucleic acid codes standardized by IUB/IUPAC (INSTITUTE, 2020), while PDB format represents the three-dimensional structures of macromolecules, which generally are proteins or peptides aggregated with other molecules or ions (NAYARISSERI et al., 2014). Section 2.1.2 will provide more information about these file structures.

Similarly, peptides that can be uptaken by the BBB, also known as BBB-Penetrating Peptides (B3PPs), also have been explored by AI due to their essential applications in NDs treatments. However, the validated databases used to train the ML algorithms in each work were not so large in terms of validated peptides with BBB activity, which impacts the learning process of the techniques and compromises the results due to the phenomenon of underfitting (DAI et al., 2021; ZOU, 2021).

The machine learning tools applied in CPPs and B3PPs prediction have explored some properties such as physicochemical, structural-, and sequence-based descriptors as input information. They evaluated how these descriptors can affect the pharmacokinetics¹ properties regarding penetration of these molecules through cell membrane (PANDEY et al., 2018; DAMIATI et al., 2019), and through BBB (KUMAR; PATIYAL, et al., 2021; DAI et al., 2021; ZOU, 2021). Although these works have studied the impact of many features in biomembrane permeability, no previous research focused on how molecular descriptors related to the oral bioavailability² are correlated to the permeability of the peptides in comparison to others descriptors. These descriptors are used by industry as an auxiliary mechanism to define the drug-likeness of molecules. These properties have been researched over the years and had the first significant advance in 1997 with the works of Christopher Lipinski and colleagues, whose results are known as Lipinski's rule of five (RO5) (LIPINSKI et al., 2012)³. Over the years, other researchers as Veber et al. (2002)⁴ and Lovering (2013)⁵ complemented this theory with other molecular descriptors.

While some studies have explored the use of ML techniques for predicting these two classes of peptides, understanding which molecular descriptors have a direct correlation with the permeability of these molecules across the two biomembranes remains a challenge. From

¹Pharmacokinetics is the path that the drug follows in the body of living beings.

²Oral bioavailability is the fraction of a drug orally administered that reaches systemic circulation.

³Descriptors related to the Lipinski's rule of five are: molecular weight (MW); calculated octanol-water partition coefficient (LogP); the number of hydrogen bond acceptors (HBA); and the number of hydrogen bond donors (HBD).

⁴Veber et al. (2002) evaluated the number of rotatable bonds (NRB) and topological polar surface area (TPSA).

⁵Lovering (2013) evaluated the fraction of sp³-hybridized carbon atoms (Fsp³).

a computational point of view, performing data mining to investigate the correlation between hundreds of molecular descriptors available in softwares and programming language packages and the peptide permeability across the cell membrane or BBB can be a hard task that impacts both the exploratory data analysis time and the ML performance, since the initial set of analyzed descriptors may not have a clear correlation with this pharmacokinetic property. Furthermore, there is a challenge in selecting molecular descriptors that simultaneously provide relevant information for the performance of ML models and have a biochemical explanation for how their calculated values for a peptide correlate with its penetration into biomembranes. This is crucial for pharmaceutical and biotechnological development, as investigating these descriptors is of greater relevance. However, many molecular descriptors available for computational calculation lack clear explanations or have never been experimentally investigated regarding these biomembranes' permeability.

In addition, only one published work until present regarding the prediction of CPPs investigated how a dimensionality reduction (DR) algorithm employed in preprocessing stages can improve accuracy in classifying peptides according to permeability, while for B3PPs prediction, no work has reported this type of investigation. Therefore, this work aims to propose the development of two architectures of frameworks based on ML to predict CPPs and B3PPs, exploring the use of supervised DR algorithm to preprocess the high-dimensional (high-d) peptide data, and evaluating how this strategy can help to visualize the molecules that can cross or not the biomembranes using physicochemical, structural- and sequence-based properties as input information, including the molecular properties related to oral bioavailability.

1.2 Motivation

Peptides are a group of molecules that can act as direct therapeutic agents or as drug carriers, reaching regions inside the cell or in the central nervous system (CNS) that typically several drugs could not achieve, and performing this prediction can be a great ally to the discovery and development of new drugs and other biotechnological applications, mainly with application in neurodegenerative diseases, microbial infection, and gene therapy (ZHOU; SMITH; LIU, 2021; BAIG et al., 2018). The development of new therapies against neural diseases is crucial for the coming years of humanity, as the drugs approved so far are not as efficient to treat disorders such as Parkinson's and Alzheimer's, as well as the unregulated use of antibiotics is escalating a severe global crisis, providing the emergence of new strains of super-resistant bacteria and fungi.

According to a study conducted by Feigin et al. (2019), neurological disorders stood out as the primary contributor to disability-adjusted life years and emerged as the second most prevalent cause of global mortality, resulting in nine million deaths annually. Out of the nine million global deaths reported in 2016, the study evaluated 15 neurological disorders and identified the top three neurological causes as stroke (67.4%), Alzheimer's disease, and other related

dementias (20.3%), along with meningitis (3.7%). Parkinson's disease, encephalitis, traumatic brain injuries, multiple sclerosis, central nervous system cancers, and neuroinfectious diseases are other common neurological disorders explored in this investigation (FEIGIN et al., 2019; OWOLABI et al., 2023).

The World Health Organization (WHO) recognizes antimicrobial resistance as one of the top three significant threats to public health. According to WHO, antimicrobial-resistant infections rank third as the leading cause of mortality, following cardiovascular diseases. A substantial study published in January 2022 revealed that approximately 1.27 million deaths were attributed to antimicrobial-resistant infections in 2019 alone. Additionally, nearly 5 million deaths were associated with drug-resistant infections. Some projections indicate that this number may soar to 10 million per year by 2050, surpassing deaths from cancer (MURRAY et al., 2022; SALAM et al., 2023).

Therefore, the motivation of this thesis is the desire to contribute to the scientific and industrial community through the development of more efficient computational tools to predict the penetration of peptides into both biomembranes, which can contribute to the development of new therapeutic agents against several diseases more quickly and less expensively.

1.3 Related works

This section reviews of state-of-art related to the development of ML algorithms to predict the uptake of peptides across the cell membrane and BBB. Additionally, this proposal reviews some advances and applications of dimensionality reduction techniques in chemoinformatics problems. Furthermore, some works exploring the chemical space of peptides for oral bioavailability in drug discovery will be presented. The purpose here is to explore the main characteristics of algorithms and molecular descriptors approached over the years by researchers in this theme.

1.3.1 Review of ML application in CPPs prediction

The first statistical tools were developed to predict CPPs in 2005 by Hällbrink and collaborators (HÄLLBRINK et al., 2005) and posteriorly in 2008 by Hansen and collaborators (HANSEN; KILK; LANGEL, 2008). They used z-descriptors, representing the average of a group of physicochemical properties calculated from peptides. Since then, some improvements have been raised by using machine learning tools with different architectures of algorithms capable of learning nonlinear patterns from different kinds of peptides' features.

Each work aimed to use different classifiers to not only predict CPPs but also to explain how physicochemical, sequence-, and structural properties increase the information regarding the uptake of these molecules by the cell membrane, providing biochemical insights about these descriptors. Some of the first published works on this topic are cataloged below with a brief

explanation:

- (A. DOBCHEV et al., 2010): This was the first work approaching the use of a machine-learning algorithm to categorize if a peptide can translocate. The authors used an artificial neural network (ANN) to classify 101 peptides using 250 molecular features. The principal component analysis (PCA) was also employed to select the best descriptors to be used in ML. The results showed that descriptors as the topographic electronic index for all bonds and charged partial surface area have a high correlation with the intake of these molecules into cells, and the best-trained model achieved 83% of accuracy.
- (SANDERS et al., 2011): In this paper, the authors approached the use of support vector machine (SVM) to predict CPPs. Four datasets of peptides were evaluated with different groups of molecular descriptors, such as physicochemical and amino acid composition (AAC). The results proved that SVM achieved the best performance with the fourth dataset composed of 111 CPPs and 111 non-CPPs with redundancy, reaching an accuracy of 95.94% in 10-fold cross-validation. Furthermore, the work concluded that descriptors such as negative charge, isoelectric point, percent hydrophobic, water-octanol partition coefficient (LogP), percent negative, hydrophobicity, and AAC have a better correlation to this pharmacokinetic property.
- (GAUTAM et al., 2013): Similar to previous work, this paper explored the use of SVM as ML classifier to differentiate 708 CPPs from non-CPPs. The authors also investigated descriptors AAC, dipeptide composition (DPC), binary profile of patterns, and physicochemical properties as input to train the model. The results were evaluated by 5-fold cross-validation, leave one-out cross-validation, and independent test. The proposed method outperformed the accuracy of the models developed by Dobchev et al. and Sander et al using three datasets. Moreover, the result also concluded that DPC provides more information to SVM than other descriptors. The authors also developed CellPPD, a webserver based on trained SVM to perform CPP prediction.
- (CHEN, Lei et al., 2015): This paper investigated CPPs and non-CPPs encoded by pseudo-amino acid composition (PseAAC), which was used as features to train a random forest (RF) to predict cell penetration of peptides. The authors also employed minimum redundancy maximum relevancy (mRMR) to select the best features, and used incremental feature selection (IFS) to evaluate the performance of the ML model based on a subset of these selected features. The results showed that the proposed model outperformed Sander et al.'s method. Furthermore, this work also concluded that the encoded AAC, polarity, secondary structure, molecular volume, codon diversity, and electronic charge, all grouped by PseAAC, are more relevant to classify CPPs correctly.
- (DIENER et al., 2016): This work investigated the performances of SVM and RF to predict cell penetration by peptides using AAC and physicochemical properties, such as:

mean charge; sliding window range of charge; hydrophobicity; isoelectric point; sliding window range of the hydrophobic moment; LogP; and an approximation of the alpha-helical content in the sequence. The ML performances were evaluated according to 4-fold cross-validation and showed that the RF performed better and achieved accuracy near 90%. The proposed method is available in a webserver called DCF.

- (TANG et al., 2016): The authors developed in this work the C2Pred, an SVM-based webserver to predict CPPs using DPC as input features. The results showed that based on feature selection using ANOVA, the best model with 164 features achieved 83,5% accuracy in 5-fold cross-validation. The authors also compared the C2Pred with Sander et al.'s and Chen et al.'s method, and it outperformed the prediction capacity using 10-fold cross-validation as the metric.
- (WEI; XING, et al., 2017): This work focused on exploring the representation capability of 4 groups of features still not sufficiently explored: parallel correlation pseudo amino acid composition (PC-PseAAC), series correlation pseudo amino acid composition (SC-PseAAC), adaptive skip dipeptide composition (ASDC), and physicochemical properties. These features were applied in two feature selection algorithms to improve the representation of each class of peptide. To perform the prediction, the authors also proposed the CPPred-RF, a random forest-based framework with two layers to classify peptides in CPPs and non-CPPs, and measure the uptake efficiency of CPPs, respectively. The accuracies achieved in the jackknife test were 91.6%, and 71.1% in CPPs prediction and uptake efficiency.
- (WEI; TANG; ZOU, 2017): This work proposed the SKIPCPP-Pred, a framework based on RF and adaptive k-skip-2-gram algorithm, a technique developed to integrate distance information of amino acids into the traditional n-gram model. k-skip-2-gram extracts a 400-dimensional feature vector from peptides used by RF to classify the molecule in CPPs or non-CPP. The work results showed that the framework outperformed some state-of-art predictions, achieving 90,6% in the jackknife test.

Although the previously selected works have explored not only several molecular descriptors (mainly sequence-based ones) but also several machine learning techniques, more recent work (from 2018 to the present) has focused on developing tools dedicated both to using improved sequence-based features and exploring new descriptor selection strategies.

Qiang et al. (2018) proposed CPPred-FL, a tool that uses RF to predict CPPs using nine groups of features: Composition–Transition–Distribution (CTD); AAC; Parallel correlation-based pseudo-amino-acid composition (PC-PseAAC); Series correlation-based Pseudo-Amino-Acid Composition (SC-PseAAC); G-gap dipeptide composition; Adaptive skip dipeptide composition (ASDC); N+C-terminal approach; Twenty-bit features (BIT20); Twenty-one-bit features (BIT21); and overlapping property features. The method developed in this work has two

steps before defining the best-trained RF model. The first step is to use all the features and subdivide them to train 45 RF models (one model by feature). The second step ranks the trained models based on the mRMR algorithm, filtering only the best classifiers. The results show that CPPred-FL achieved superior results in 10-fold cross-validation based on the area under the curve (AUC), which reached 0.445, a higher value than other frameworks such as CPPRed-RF, SkipCPP-Pred, and CellPPD. This work is essential for being a pioneer in the development of more sophisticated mechanisms for selecting ML models and representative features within the line of research on CPPs prediction. However, the CPPred-FL model and feature selection steps lack clarity regarding some method characteristics, such as how the 45 subgroups of features were separated to train the same number of RFs.

Pandey et al. (2018) developed a tool called Kelm-CPPred was proposed, which uses an extreme learning machine (ELM) to predict the permeability of peptides in the cell membrane. This work focused on using AAC, DPC, PseAAC, and a hybrid approach of these three feature sets. The authors extensively analyzed the level of information that each feature provided to the model and compared the best one with other developed frameworks. The results showed that Kelm-CPPred reached accuracies between 85.20% and 86.64% in the 10-fold cross-validation and 87% in the independent test using the hybrid-AAC composition, and outperformed existing prediction models in almost all group of features using their datasets. Although this work is the first one to explore the use of ELM in this research field, which is efficient and computationally less expensive when compared with models previously used in other frameworks to predict CPPs, such as SVM and ANN, the authors do not add new information regarding the features that describe the peptides.

Kumar, Agrawal, et al. (2018) was the first to introduce the use of tertiary peptide structures in the study of penetration prediction of peptides through the cell membrane using ML. In this work, the SVM, RF, Naive Bayes (NB), J48, and SMO algorithms were evaluated for the predictive capacity of CPPs using structural molecular descriptors, such as atomic composition; diatomic composition; 2D descriptors; 3D descriptors; and molecular fingerprints. In addition to these structural descriptors, the authors also evaluated models with sequence-based features, such as AAC, DPC, and composition-based terminus. The results show that the SVM reached accuracies of 91.67% and 89.67% for the 5-fold cross-validation and the external validation, respectively. In summary, this work has a differential compared to the others, as it uses the tertiary structure of the peptides and evaluates the prediction of these molecules with chemically modified residues, differing from other tools that use only the primary structure of the peptides as features. Also, as a contribution, the authors developed CellPPDMod, a webserver to predict CPPs based on the best model developed. However, only one peptide at a time can be predicted by this tool.

In Manavalan et al. (2018), the authors proposed the MLCPP, a framework of two layers to predict whether a peptide can cross the cell membrane. The first layer predicts if a peptide

is a CPP and the second one predicts the level of CPP uptake (high or low). In this study, the algorithms RF, SVM, ERT, and kNN were independently selected to compose the two layers based on the Matthews Correlation Coefficient (MCC) index. The databases used in this work were collected from CPPsite 1 and C2PRed, and the features used were AAC, AAI, DPC, and physicochemical properties. The results showed that ERT and RF achieved better performance in cross-validation and were selected to construct the first- and second layers, respectively. Furthermore, the framework outperformed the state-of-art methods on the independent test, reaching 89.6% and 72.5% of accuracy in CPPs prediction and uptake efficiency, respectively. The differential of this work is that, in addition to surpassing the performance of other tools already developed, it is the second framework after the development of CPPred-RF that also evaluates the efficiency of capturing CPPs, resulting in an ML framework of 2 layers. The layers of the best MLCPP model were structured using sequence features like AAC and DPC, simpler and more interpretive than those applied in CPPred-RF.

In 2019, the SVM algorithm was again explored and evaluated for its ability to predict cell membrane penetration. In Fu et al. (2019), the authors use SVM to predict CPPs based on sequence properties such as Grouped Amino Acid Composition (GAAC); k-Spaced Amino Composition Acid Group Pairs (CKSAAGP); Grouped Di-Peptide Composition (GDPC); and Composition-transition-distribution (CTD). A resource selection process was also carried out to improve representative capacity using SVM recursive feature elimination (RFE) and correlation bias reduction (CBR). This research shows that the proposed method reached 92.3% accuracy in the knife test for the four resources used, a value superior to other state-of-the-art techniques. In addition, the authors concluded that the CTD feature exhibited the best effect on prediction performance. Although this work has achieved good results and proposed a straightforward methodology, it only contributes to using some features not previously explored in this context. In contrast, the algorithm used still fits into a technique widely studied, and the features were not employed to investigate the performance of other algorithms in CPPs prediction. Another disadvantage is that the developed tool was not used in an easy-to-use web application, as with other recent works.

1.3.2 Review of ML application in B3PPs prediction

Since 2002, some works have been published to disclose results of using machine learning techniques to predict the permeability of compounds through the BBB. In Doniger, Hofmann, and Yeh (2002) was used ANN and SVM to predict the penetration of 324 compounds across this barrier, and more recently, Plisson and Piggott (2019) applied gradient RF and logistic regression (LR) to predict the permeability of 471 marine products on the same membrane. Until 2019, 24 articles have been published on the topic of using machine learning to predict the penetration of compounds through the blood-brain barrier (SAXENA et al., 2019). However, it was only in 2021 that the first works related to using ML to predict the penetration of peptides through this biomembrane were published.

The first work on the use of ML to predict the penetration of B3PPs was published by Dai et al. (2021). This work uses peptide sequences from Brainpeps, SATPdb, PepBank, and other curated bases. The authors explored 16 sequence-based molecular descriptor classes. The 3-stage feature selection process was: F1-score classification obtained by ERT models trained for each descriptor individually; excluding redundant resources using Spearman's coefficient greater than 0.7; and selecting the best trait subgroup using direct sequential search (SFS). The machine learning models trained and evaluated were ERT, RF, SVM, multilayer perceptron (MLP), extreme gradient boosting (XGBoost), and LR. This last one obtained the best performances with accuracies of 77.5% and 78.95% for 10-fold cross-validation and independent tests, respectively. The difference in accuracies reached by the best model in each scenario can be explained by a possible distinction in the distribution of the molecular descriptors between the training and test datasets since the number of samples for the independent test is relatively much smaller compared to the training samples and the model may have suffered overfitting for a subset of training data with a feature distribution profile similar to that of the test. Furthermore, this work provides an essential contribution to this field of research as it is the first tool dedicated to predicting B3PPs. However, the training and test datasets raise questions about the quality of the models since the relatively small amount of samples may have induced the overfitting of the techniques.

Similarly, Kumar, Patiyl, et al. (2021) proposed B3Pred, a RF-based computational tool to predict BBB penetration of peptides using an optimized subset of sequence-based features from a total of 15 different feature classes. A subset of these descriptors was selected in two steps, the first using the SVC-L1 algorithm to select the most correlated features and the second step ranking the features based on their importance in classification using a light gradient boosting machine (LightGBM). The decision tree (DT), RF, LR, kNN, *Gaussian* naive bayes (GNB), XGBoost, and SVM algorithms were evaluated as classifiers using the selected features for three datasets of peptides. The results show that the RF obtained the best performance with 85.08% of correct answers using 80 selected descriptors. Compared to the previous work, B3Pred is a tool built with a slightly more extensive database and has evaluated many more features, and the final model is based on a smaller number of selected descriptors. However, this work does not discuss the biochemistry correlation of each selected feature with the BBB permeability.

Unlike other works published in this same line of research, Zou (2021) used several techniques to select the best descriptors among the ten based on the physicochemical properties analyzed: hydrophobicity; hydrophilicity; side-chain mass; pK1 (C α -COOH); pK2 (NH₃); PI (25 °C); average buried volume; molecular weight; side-chain volume; and mean polarity. The author used Pearson's and the maximum correlation coefficients to select the best descriptors. Selecting features based on these two criteria is then merged using the similarity network fusion algorithm. Then, a new subgroup of descriptors is chosen by applying Fisher's algorithm. The classifiers used in this work to predict the B3PPs are DT, RF, kNN, NB, and SVM, and the

results show a better performance achieved by the SVM with an accuracy of 89.47% for the independent test, indicating pK2 (NH3) as the most relevant property to predict the penetration of peptides through the BBB concerning all ten analyzed features. One of Zou's main contributions is that it is the first to focus on physicochemical properties and expand new feature selection strategies. However, some points in its methodology could be more transparent, such as how the variables are used in Pearson's correlation selection.

Xue Chen et al. (2022) published a paper proposing the BBPpredict, an ML-based web-server to predict the permeability of peptides across the BBB. The authors used a dataset of 652 sequences equally distributed in positive and negative samples, while for the independent test, 198 sequences were distributed equally. The features used to predict the peptides were AAC, DPC, PseAAC, CKSAAGP, and GAAC, and F-score method was implemented to select the most informative descriptors among all the properties extracted. The ML classifiers trained and tested were DT, RF, SVM (with linear and radial basis function), kNN, adaptive boosting (AdaBoost), gentle adaptive boosting (GentleBoost), adaptive logistic regression (LogitBoost), and long-short term memory (LSTM). This paper shows that for five-fold cross-validation, the RF achieved the best performance with an accuracy of 81.90%. At the same time, the RF also outperformed the other techniques with 77.78% accuracy for the independent test. This work contributed to one more available web tool to predict B3PPs and outperformed the BBPpred and B3Pred with the independent test dataset. However, there is no significant contribution in terms of mechanism to preprocess the dataset before training or chemical space relating the most informative sequence-based descriptors and the classes BBB+ and BBB-.

Most recently, Charoenkwan et al. (2022) developed the SCMB3PP, a computational tool used to predict B3PPs. SCMB3PP uses a scoring card method-based predictor (SCM) for generating propensity scores of amino acids and dipeptides and a genetic algorithm for optimization of the propensity scores. The peptides are classified as B3PPs and non-B3PPs according to a threshold established for SCM scores. The results achieved in this work indicated that SCMB3PP achieved accuracies between 83% and 95.1% for 10-fold cross-validation analysis and accuracies between 88% and 94.4% for independent testing. The contribution of this work relies on an alternative web tool to predict B3PPs. Furthermore, SCMB3PP also outperformed other state-of-art tools such as BBPpred, B3PPred, iBBP, and MIMML. However, this work does not contribute to ML development for predicting B3PPs, since the proposed method only uses a unique value to differentiate B3PPs from non-B3PPs. Furthermore, the work is unclear in explaining how physicochemical properties were extracted from analyzed peptides.

1.3.3 Review of dimensionality reduction application in computational chemistry

The review's works presented above focused on how supervised classifiers predict CPPs and B3PPs. However, these works did not explore dimensionality reduction algorithms as a strategy to process the high-d datasets composed of many molecular descriptors. DR algorithms

can be a powerful tool to overcome issues related to high-d datasets, mapping the original data on a low-dimensional (low-d) space and making it possible to explore the features most correlated to the sample classes, or even improve the modeling of the problem Chao, Luo, and Ding (2019).

Although these DR algorithms have not been explored explicitly in the context of the problem approached in this thesis, some works related to computational chemistry used this tool. Nasser et al. (2022) used the autoencoder to reduce the dimensionality of a dataset of small molecules to perform a similarity search. This method outperforms traditional methods such as the Tanimoto Similarity Method, Adapted Similarity Measure of Text Processing, and Quantum-Based Similarity Method. The three autoencoder architectures proposed achieved better recall metrics in almost all cases evaluated, proving that dimensionality reduction can be auxiliary in virtual screening using low dimensional representation.

In the same field, Mostafa, Salem, and Mohamed (2022) proposed an ML-based framework composed of a feature selection algorithm and a DR technique, where these two stages are responsible for filtering the best descriptors and mapping the result onto a low-d space, respectively, before the stage of classification. The authors tested principal component analysis (PCA), uniform manifold approximation and projection (UMAP), and Neighborhood Components Analysis (NCA) as DR algorithms and SVM, kNN, and LR as classifiers to predict inhibitors (antidepressant medicines) and inducers molecules. The results proved that feature selection with UMAP achieved the best accuracy with the value of 99% for SVM.

Similarly, Jinuraj et al. (2018) explored virtual screening to select molecules for experimental tests against *Leishmania Mexicana*. Here was proposed the use of PCA to reduce the number of molecular descriptors to apply in a self-organizing map (SOM) and in the Eli Lilly MedChem rule filter to refine the screening of molecules. Using PCA improved results' accuracy and helped select two molecules for experimental tests.

Dimensionality reduction has also been applied in molecular dynamics to represent the molecular structures more simply. Zhou, Wang, and Tao (2018) evaluated how t-distributed stochastic neighbor embedding (t-SNE) could construct low-d descriptors to represent the free energy landscape of Vivid (a photosensitive circadian clock protein) related to the switching between the dark and light states. The results show that even for one dimension, t-SNE outperforms the PCA and Time-Structure Based Independent Component Analysis (t-ICA) according to RMSD metrics. Furthermore, t-SNE could retain the structural and dynamical information with minimum information loss compared to other commonly used DR methods.

In summary, supervised and unsupervised dimensionality reduction algorithms have been used as an alternative to overcome high-d problems pertinent to many issues in various chemical informatics fields and theories. Mainly, virtual screening of molecules makes the problem easier to solve and more interpretative, allowing advances in drug design.

1.3.4 Review of bioavailability properties in peptide

Until this point, some papers related to the application of ML in CPPs and B3PPs were reviewed, focusing mainly on the techniques and some aspects of molecular descriptors used as features to predict the uptake of peptides. Nevertheless, it is essential to highlight that some molecular descriptors are desirable for a molecule to be more drug-like, whose properties are related to oral bioavailability. This thesis will also explore these descriptors as features to predict peptides' permeability.

Although many published works approached the use of descriptors related to oral bioavailability, few works explored how these features affect the peptides. Santos, Ganesan, and Emery (2016) evaluated the chemical space for all FDA-approved drugs from 2012 to 2016 according to MW, NRB, LogP, HBA, HBD, TPSA, and Fsp³, noting that only Fsp³ fits the criteria described by the literature (average < 0.47), while the remainder descriptors broke the rules of oral availability. Similarly, Díaz-Eufracio et al. (2018) explored the chemical space of pentapeptides of six datasets over the effect of cyclization and N-methylation. The descriptors analyzed were MW, NRB, LogP, HBA, HBD, and TPSA, and the result shows that N-methylation and cyclization change the peptides' chemical space toward the FDA-defined one, representing a promising source to explore novel and biologically relevant intervals of these descriptors.

1.4 Contributions

Over the years, many works have been published approaching the development of new ML tools for CPPs prediction and some for B3PPs prediction. These tools have become a great ally in drug discovery and development against several diseases due to decreased costs- and time-related to research.

The problem of how peptides can cross the cell membrane and blood-brain barrier can involve many variables related to physicochemical and structural molecular properties, turning this into a high-dimensional challenge. In terms of framework architecture, all the ML-based tools developed until this moment to predict this permeability focused on pipelines involving direct feature selection according to classes of molecular descriptors (MANAVALAN et al., 2018; PANDEY et al., 2018) or using some statistical strategies, such as remotion of redundancy (FU et al., 2019; PANDEY et al., 2018), information gain from tree-based ML (DAI et al., 2021), performance by minimal Redundancy Maximum Relevance (WEI; XING, et al., 2017), Pearson's correlation coefficient (WEI; XING, et al., 2017), and maximal information coefficient (ZOU, 2021). However, dimensionality reduction algorithms have not been explored until now by research in the context of these two applications as a mechanism of data preprocessing. The DR can reduce the number of features to a low-d representation of the peptides' chemical space that can be used as reduced data in machine learning classifiers. Therefore, the use of DR algorithms can increase the performance of CPPs and B3PPs prediction with ML models. They can reveal how different and most significant molecular descriptors can cluster the peptides, which

constitutes trivial information to design new pharmaceutical and biotechnological applications with these molecules.

Among the types of DR algorithms, there are supervised manifold techniques, which have been gaining prominence due to their ability to preserve discriminative information in the projection into low-d based on the labels of each sample class, in addition to their ability to process non-linear information among the samples (CHAO; LUO; DING, 2019). These two characteristics are essential in the process of classifying molecules based on various molecular properties, which usually do not have a linear relationship with their classes. However, the use of some of these algorithms has limitations because their canonical structure is not capable of generating a mathematical model that can be reused with new samples, preventing the use of these techniques in classification pipelines. Therefore, developing strategies capable of learning the projection pattern of these DR algorithms is an essential step to overcome this problem, besides contributing to the performance of ML-based pipelines in predicting new CPPs and B3PPs.

Analyzing the correlation that the penetration of these two classes of peptides has with the molecular properties involved in compound bioavailability, especially those described in Lipinski's Rule of 5, is essential for the planning of potential therapeutic agents by the pharmaceutical industry (BENET et al., 2016; MULLARD, 2018). However, a comprehensive analysis on a statistical level or of the information gain that these molecular descriptors have for predicting CPPs or B3PPs using machine learning models has not been conducted. Since the publication of A. Dobchev et al. (2010) up to the work by Fu et al. (2019)⁶ regarding the prediction of CPPs using ML algorithms, the works concentrated majority on investigating features based on the amino acid residue sequence, except the works by Manavalan et al. (2018) and Kumar, Agrawal, et al. (2018) that assessed some physicochemical properties. Furthermore, studies regarding the use of ML to predict B3PPs published from 2021 up to the writing of this thesis⁷ also focused on the use of sequence-based descriptors, except the work by Zou (2021), which also investigated some physicochemical properties. However, no one of these published works investigated and compared the benefits of using the molecular descriptors associated with the bioavailability of compounds.

Another essential aspect that has been underexplored in this research field is the development of tools for predicting the permeability in the cell membrane or the BBB of non-natural peptides or those with chemical modifications. These molecules occur naturally in nature or can be obtained through synthesis, and they have various biotechnological applications. Unlike natural peptides, there is no canonical way to represent such molecules solely based on

⁶Fu et al. (2019) was the last work published before the paper **Predicting Cell-Penetrating Peptides Using Machine Learning Algorithms and Navigating in Their Chemical Space**, which was published by the author of this thesis and collaborators in 2021.

⁷In 2023, the author of this thesis and collaborators published the paper **BrainPepPass: A Framework Based on Supervised Dimensionality Reduction for Predicting Blood-Brain Barrier-Penetrating Peptides**.

their primary structure, making the investigation of these molecules challenging using only sequence-based molecular descriptors. This necessitates obtaining and processing the tertiary structure of the peptide. Among the published works on the prediction of CPPs, only the tool CellPPd-Mod utilized the tertiary structure of peptides (KUMAR; AGRAWAL, et al., 2018). However, this tool has a practical limitation in its web version to predict the permeability of one structure at a time. Regarding the prediction of B3PPs with chemical modifications, no work published before the writing of this thesis has investigated the use of ML to predict this class of molecules.

The use of a curated and validated database is essential for the analysis of phenomena and the development of reliable prediction models. Many works involving the development of machine learning models to predict CPPs have utilized samples from CPPSite 2.0, one of the largest curated and validated databases for this class of peptides (AGRAWAL et al., 2016). On the other hand, although there is a curated database for peptides with BBB penetration activity, named BrainPeps (VAN DORPE et al., 2012), investigations into the creation of ML models capable of predicting B3PPs have predominantly used computationally synthesized molecules. Many of these lack experimental validation of penetration into the blood-brain barrier, compromising the model's reliability for testing real molecules.

Therefore, this thesis aims to contribute to this research line's state of the art, addressing some unexplored points. The main contributions of this thesis are highlighted below:

- Development of a ML-based framework to predict natural or chemically modified CPPs and B3PPs using supervised manifold dimensionality reduction as a preprocessing strategy.
- Investigation of the structural and physicochemical properties associated with the oral bioavailability of compounds described in Lipinski's rule of five.
- Use of experimentally validated database related to uptake of peptides by the blood-brain barrier.
- Development of free-to-use web servers to predict CPPs and B3PPs based on the best models.

1.5 Objectives

The general thesis proposal is the development of machine learning-based tools to predict the permeability of CPPs and B3PPs, investigating how physicochemical, structural- and sequence-based descriptors influence the process of penetration of these molecules in each biomembrane. Some specific objectives were designed and listed below to fulfill the general proposal of this thesis:

- a) Collect the database of peptides tested to cell membrane permeability in servers as CPP-Site 2.0 and C2Pred and from some published works, and construct datasets in PDB and FASTA format. Also, collect samples of peptides tested to BBB penetration from Brain-Peps in the MDL format.
- b) Extract the physicochemical, structural-, and sequence-based descriptors using computational tools to construct the datasets according to feature compositions (FC), where for cell membrane case, the descriptors will be calculated from both PDB and FASTA format, and for BBB problem they will be extracted from MDL file.
- c) Develop two machine learning based frameworks using sLE in preprocessing stage, and XGBoost regression and classifier. The first framework is dedicated to predicting CPPs and the second to predicting B3PPs.
- d) Evaluate the proposed frameworks according to 10-fold cross-validation and compare with voting classifiers (Vcf) grouping baseline algorithms ANN, SVM, and GPC in the prediction of CPPs and B3PPs, respectively.
- e) Perform independent test for cell membrane case and compare the proposed framework with the voting classifier and other state-of-art tools according to metrics of accuracy, sensitivity, specificity, F1-Score, area under the curve (AUC), and Matthews Correlation Coefficient (MCC).
- f) According to the metric results, investigate how the descriptors in each FC impact the permeability prediction of peptides in both biomembranes. Besides, evaluating how the molecular descriptors related to bioavailability affect the permeability prediction.

1.6 Thesis Organization

The subsequent sections of this thesis are structured in the following manner. **Chapter 2** approaches the theoretical background about the pharmacokinetics and biochemistry aspects of cell membranes and the blood-brain barrier, and how peptides can cross them. Also, this chapter explains the theory of the machine learning algorithms used in our methodology.

Chapter 3 describes the database of CPPs and B3PPs used in each case study, besides the feature extraction from the peptides' structures. Furthermore, this chapter approaches the architecture of the voting classifier used to solve the CPPs problem and the architecture of the framework used to solve the B3PPs problem.

Chapter 4 presents the result of applying each designed framework to predict the permeability of CPPs and B3PPs, showing the performance of each tool and some aspects of the impact of each descriptor. Finally, **Chapter 5** presents the conclusions about the results achieved. In addition, this chapter also explains the next steps of this research and the published works during the doctorate period.

2 THEORETICAL BACKGROUND

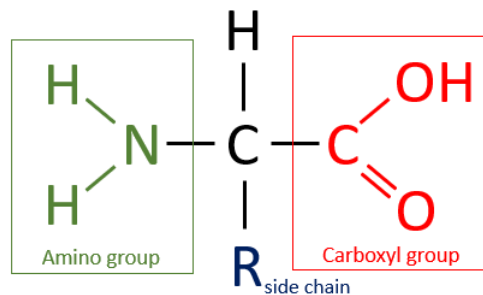
This chapter presents general aspects of the biochemical subjects related to this thesis, covering the structure of a peptide, the cell membrane, and the blood-brain barrier. The chapter also approaches some fundamental aspects of machine learning to understand this thesis's purpose, which encompasses dimensionality reduction and the sLE algorithm up to the architecture of a voting classifier and the algorithms ANN, GPC, SVM, and XGBoost.

2.1 On biochemical and computational aspects of peptides

2.1.1 Biochemical aspects of peptides

Peptides are organic molecules formed by linking two up to 50 amino acids through a peptide bond. Amino acids are the fundamental structures that constitute a peptide and are represented by the formula $R-CH(NH_2)COOH$. These structures comprise an amino group, a carboxyl group, hydrogen, an asymmetric carbon, and an R group representing a side chain, as shown in Figure 1. The peptide bond links two amino acids using a covalent bond, where the carbon of carboxyl groups of one amino acid links to the nitrogen of the amino group of the other amino acid by a dehydration reaction (LANGEL et al., 2009; FORBES; KRISHNAMURTHY, 2021), as shown in Figure 2.

Figure 1 – Amino acid structure.



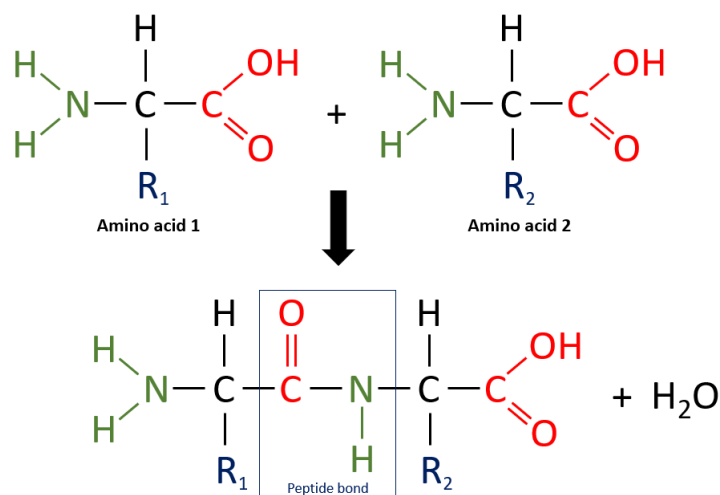
Source: Author's own.

There are 20 naturally occurring amino acids that make part of the structure of almost all peptides and proteins. These amino acids can be divided into two large groups, polar and non-polar. The amino acids aspartate (D)¹, glutamate (E), arginine (R), lysine (K), serine (S), threonine (T), cysteine (C), methionine (M), asparagine (N), and glutamine (Q) constitute the polar group. The group of non-polar amino acids encompasses the following residues: glycine (G), alanine (A), valine (V), leucine (L), isoleucine (I), proline (P), phenylalanine (F), tyrosine (Y), histidine (H), tryptophan (W). These amino acids are considered natural because

¹The letter beside the residue name represents the one-letter codification of amino acid.

the human genome encodes them. Other residues such as 5-hydroxylysine, selenocysteine, 4-hydroxyproline, and 6-N-metil lysine are non-natural amino acids that can be found in peptides (VERLI, 2014).

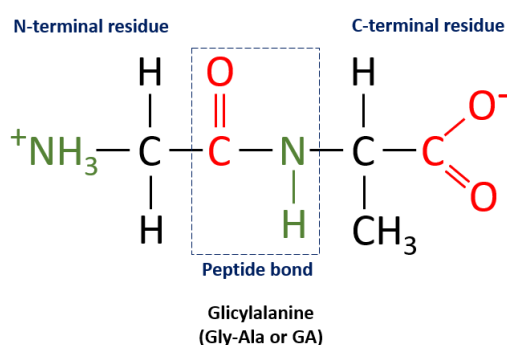
Figure 2 – Peptide bond schematic.



Source: Author's own.

A peptide has two ends: the end with a free amino group (NH_2) is called the N-terminal amino acid residue. The end with a free carboxyl group (COOH) is called the C-terminal amino acid residue. These groups are protonated² when the peptide is in an aqueous solution (OUELLETTE; RAWN, 2018). Peptides are named from the N-terminal acid residue to the C-terminal amino acid. Figure 3 illustrates the Glycylalanine peptide with its N- and C-terminal.

Figure 3 – Schematic representation of the N- and C-terminal in a Glycylalanine peptide.



Source: Author's own.

Bioactive peptides play an essential role in physiological and biochemical processes related to human health, mainly affecting metabolic functions and the digestive, endocrine, cardiovascular, immune, and nervous systems (Sanchez, 2016). These molecules also display hormone and drug-like activities such as antimicrobial, anti-inflammatory, antihypertensive,

²Loss or gain of a proton H^+ .

antithrombotic, opioid, immunomodulatory, antioxidative, and can be used as drug-delivery (BHANDARI et al., 2019; BAIG et al., 2018).

Peptides must reach their molecular target to perform their therapeutic effect as commercial drugs. It is not a simple task because organisms have several biological membranes that constitute physical barriers to cells, organelles cells, tissues, and solid organs (CHANTEMARGUE, 2018). These structures have different functions such as protecting cells and organs from xenobiotics, maintaining their biochemical integrity; hosting bioactive molecules such as receptors, enzymes, ion channels, functional proteins, or even groups of cells with specific biological functions; controlling the traffic of molecules between the two sides of a cell or organs' barrier (PIGNATELLO, 2013). The interaction of peptides with biomembranes is a complex phenomenon, and some characteristics of both structures can affect the penetration of these molecules.

2.1.2 Computational coding of peptides

The computational coding of a molecule represents how this structure can be stored, read, and processed by computers, based on a set of chemical information that is desired to be stored (OLIVEIRA, 2018). Currently, there are some possibilities for coding peptides in computational files, such as FASTA, SDF, PDB, MDL, SMILES formats, and others. Each file structure has its particularities regarding the type of information it encodes in the molecule. It is essential to highlight this information because, among the formats mentioned, FASTA is the only one that uses the representation of the primary structure of the molecule, that is, the structure that depends only on the chain of natural amino acid residues. Figure 4 shows an example of a FASTA file containing four peptides, where each peptide in this file is represented by a header and the sequence of amino acid residues.

Figure 4 – FASTA file example containing the header and the sequence of peptides Opiorphin, Neurotensin, Arginine vasopressin 1-7, and PepH1.

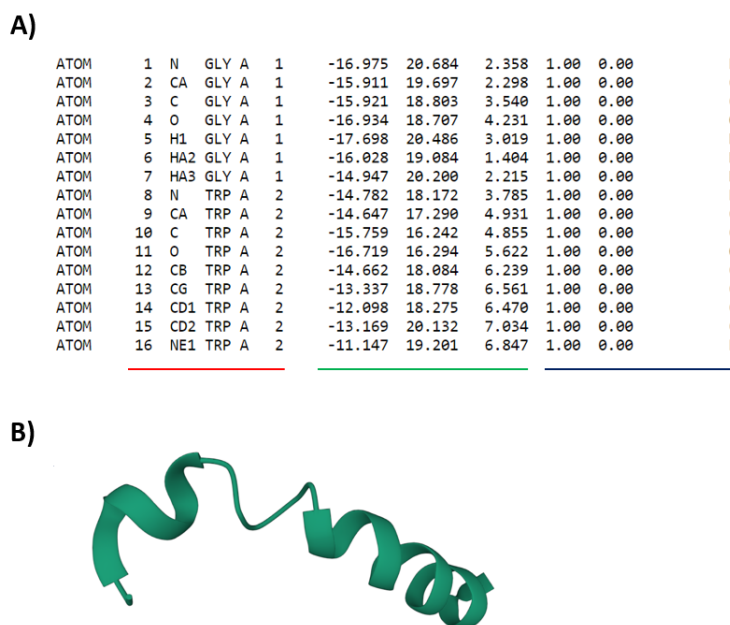
```
>Opiorphin
QRFSR
>Neurotensin
ELYENLPRRPYIL
>Arginine vasopressin 1-7
CYFQNCP
>PepH1
VQQLTKRFSL
```

Source: Author's own.

On the other hand, PDB and MDL are some file formats that encode the tertiary structure of the peptide, that is, the three-dimensional conformation of the molecule, taking into account the position of the amino acid atoms in 3D space, including non-natural amino acids and chemical modifications. Panel A of Figure 5 shows an example of part of the PDB file of the peptide Transportan (GWTLSAGYLLGKINLKALAALAKKIL). The red line indicates

the atom number, atom type, amino acid residue, chain, and residue number, respectively. The green line indicates the XYZ coordinates of the respective atom, and the blue line highlights the occupancy³, beta factor⁴, and the atom element, respectively. Panel **B** of Figure 5 shows the 3D conformation of the Transportan according to its PDB file.

Figure 5 – PDB file of Transportan peptide. **A)** PDB file containing information about the arrangement of atoms in 3D space. **B)** Visualization of the 3D conformation of Transportan using the PDB file.



Source: Adapted from Protein Data Bank (PDB ID: 1SMZ).

Figure 6 shows a part of the MDL file of the Transportan. The red line indicates the XYZ coordinates of the atoms, the green line indicates the atom symbol, and the blue line indicates other information of the atoms such as nonstandard isotope, charge valence, etc.

Figure 6 – MDL file of Transportan peptide.

```
transportan
431433 0 0 0 0 0 0 0 0999 V2000
-0.9080 1.9825 -0.1171 H 0 0 0 0 0 15 0 0 0 0 0 0
0.0000 1.5760 0.0000 N 0 0 0 0 0 15 0 0 0 0 0 0
1.2480 1.9019 0.6835 C 0 0 0 0 0 15 0 0 0 0 0 0
1.5690 0.8741 1.7431 C 0 0 0 0 0 15 0 0 0 0 0 0
0.8390 -0.0885 1.9570 O 0 0 0 0 0 15 0 0 0 0 0 0
-0.5810 0.7020 0.2554 H 0 0 0 0 0 15 0 0 0 0 0 0
2.0810 1.9486 -0.0409 H 0 0 0 0 0 15 0 0 0 0 0 0
1.1770 2.8995 1.1547 H 0 0 0 0 0 15 0 0 0 0 0 0
1.4781 -0.2728 1.5522 N 0 0 0 0 0 15 0 0 0 0 0 0
2.7241 -1.0017 1.7702 C 0 0 0 0 0 15 0 0 0 0 0 0
3.0791 -1.8393 0.5646 C 0 0 0 0 0 15 0 0 0 0 0 0
2.3721 -1.8642 -0.4427 O 0 0 0 0 0 15 0 0 0 0 0 0
2.5661 -1.9164 3.0168 C 0 0 0 0 0 15 0 0 0 0 0 0
```

Source: Author's own.

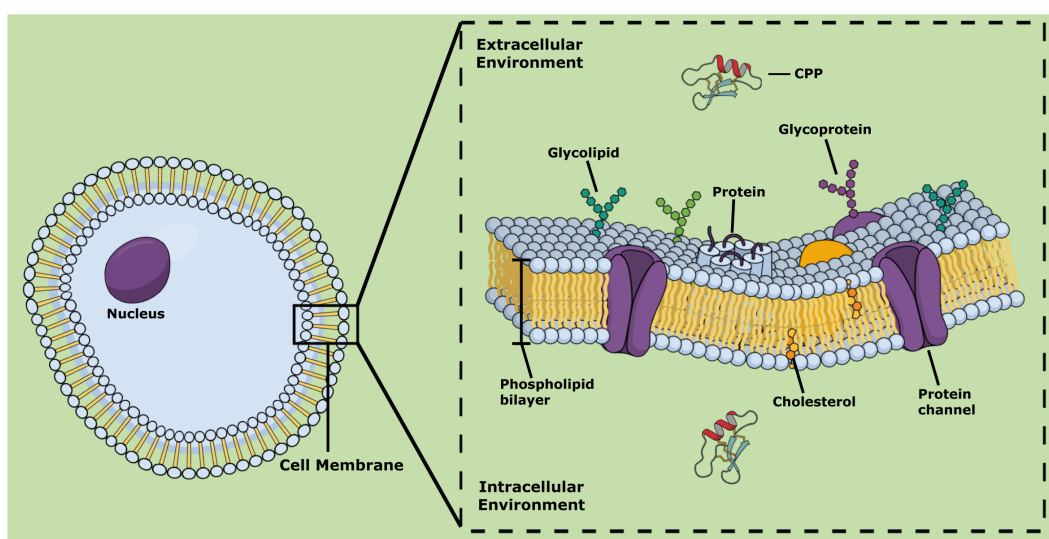
³Fraction of atoms that appear at that location.

⁴Average displacement of atom.

2.2 Cell membrane and cell-penetrating peptides

The cell membrane, also known as the cytoplasmic membrane, separates the cell from the exterior environment (YANG; HINNER, 2015). This biomembrane consists of a phospholipid bilayer that contains cholesterol between phospholipids that maintain their fluidity (SZLASA et al., 2020). Figure 7 illustrates the structure of the cell membrane better.

Figure 7 – Schematic representation of cell membrane showing its main chemical lipidic and protein components. surface.



Source: Author's own.

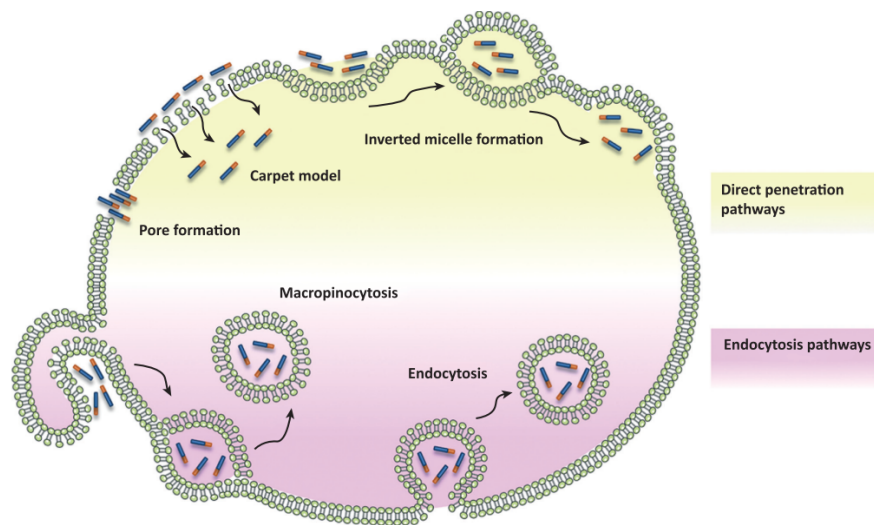
The cell membrane controls the passage of organic molecules and ions inside the cell, maintaining its homeostasis (DERAKHSHANKHAH; JAFARI, 2018). This membrane contains several transmembranes, peripheral, and lipid-anchored proteins that perform various molecular functions, including ion transportation, cell adhesion, cell signaling, and catalysis (YANG; HINNER, 2015). This biochemical structure is crucial to protect the cytoplasm and intracellular components from invading organisms and xenobiotics.

Cell-penetrating peptides (CPPs), also known as peptide transduction domains (PTD), is a class of positively charged short peptides with 5–30 amino acids that have been reported to traverse the cell membrane (DERAKHSHANKHAH; JAFARI, 2018; YANG; HINNER, 2015; MANAVALAN et al., 2018). In 1988 and 1991 were identified the first CPPs TAT and Penetratin, which were derived from the transactivator protein (Tat) of human immunodeficiency virus type 1 (HIV-1) and the *Drosophila* antennapedia homeobox protein (pAntp), respectively (KARDANI et al., 2019). Regarding their application, these peptides can as much act directly as is the case with antimicrobial peptides (ANNUNZIATO; COSTANTINO, 2020), as cargo-delivery, binding to other drugs or molecules that have difficulty of crossing the cell membrane (e.g., DNA, siRNA, protein, and peptide) (GUIDOTTI; BRAMBILLA; ROSSI, 2017).

The exact transport mechanism of cell-penetrating peptides through the membrane is

still a subject much studied by scientists, despite descriptions in the literature. Currently, the studies describe three ways for a peptide to enter the cell: direct penetration, endocytosis, and translocation by forming a transitional structure, as shown in Figure 8. Each of these routes has factors that influence the level of uptake, such as the level of concentration of peptides, the sequence of amino acids, the lipid components in each membrane, and physicochemical properties (BOLHASSANI, 2011; GUIDOTTI; BRAMBILLA; ROSSI, 2017).

Figure 8 – Schematic representation of endocytosis and direct penetration mechanisms for cell-penetrating peptide internalization.



Source: Guidotti, Brambilla, and Rossi (2017)

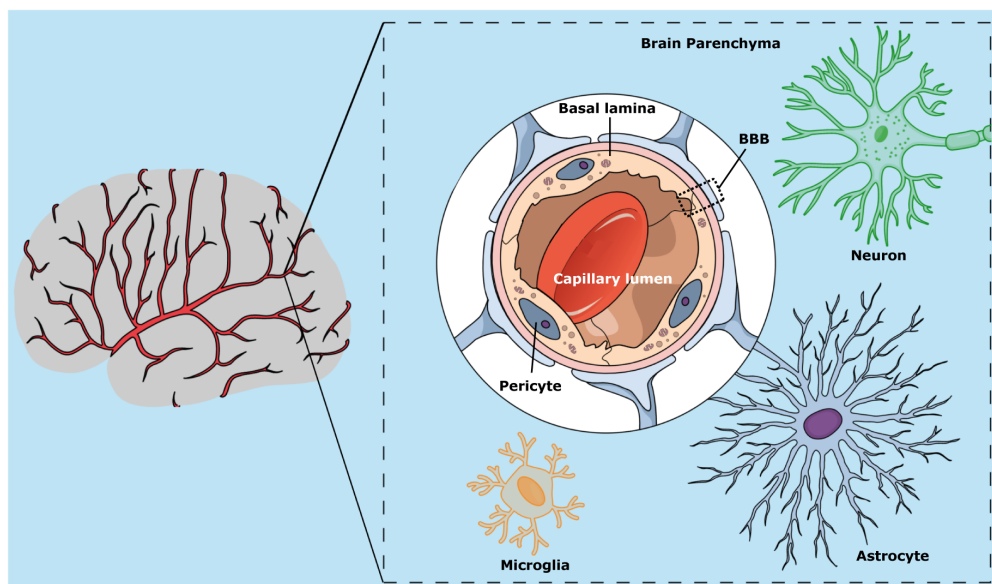
2.3 Blood-brain barrier and blood-brain barrier penetrating peptides

The blood-brain barrier (BBB) is a selective biomembrane that acts as a physical and chemical barrier of molecules of the central nervous system (CNS), controlling the homeostasis of the brain (OLLER-SALVIA et al., 2016). This biomembrane acts as a functional barrier between the brain's interstitial fluid and blood, maintaining a controlled biochemical environment necessary for neural function (LEE; JAYANT, 2019).

The BBB is mainly composed of endothelial cells on the brain capillary walls forming tight junctions among adjacent cells. Other cell types present in the BBB include astrocytes and pericytes (ZARAGOZÁ, 2020). The BBB restricts the passage of pathogens and toxins while allowing the diffusion of some solutes present in the blood to the cerebrospinal fluid (DANEMAN; PRAT, 2015). Figure 9 illustrates the structure of the blood-brain barrier.

Blood-brain barrier penetrating peptides (B3PPs), also known as brain-penetrating peptides or BBB shuttle peptides, represent oligopeptide chains with permeability into the BBB that represent interesting biotechnological applications due to their favoring the increase in the brain uptake of large molecular cargoes in a non-selective way (DÍAZ-PERLAS et al., 2018; OLLER-SALVIA et al., 2016). These peptides have been extensively investigated, aiming for

Figure 9 – Schematic representation of the blood-brain barrier, showing its main cell components (pericytes, astrocytes, and endothelial cells) and localization in the brain capillary wall.



Source: Author's own.

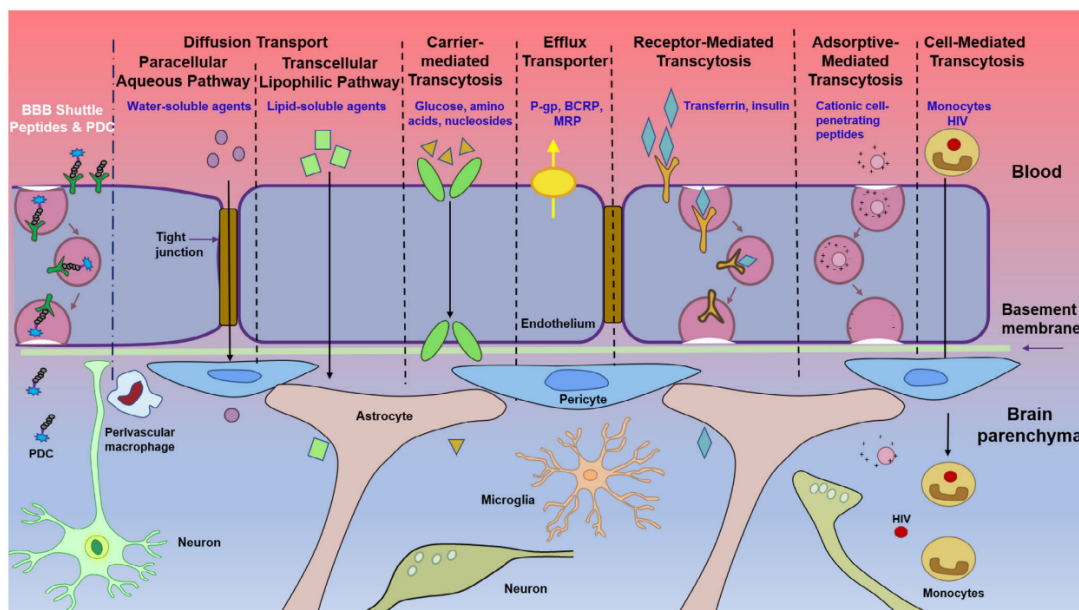
the development of new chemotherapeutic compounds due to their antiviral (JACKMAN et al., 2018), anticancer (CHEN, Long et al., 2019), and neuroprotective activities (MELONI et al., 2014; BAIG et al., 2018). B3PPs have some interesting characteristics, such as a strong affinity toward a specific receptor that is often expressed in the luminal side of brain vasculature to trigger internalization; the capacity to mediate transcytosis; and the ability to facilitate cargo transport into brain parenchyma via a noninvasive way without affecting the integrity of the BBB (ZHOU; SMITH; LIU, 2021).

The BBB is a very complex organic structure that selectively allows the passage of molecules from the blood to the brain parenchyma. Currently, the literature describes penetration mechanisms in this barrier, such as passive diffusion, carrier-mediated transport, receptor-mediated transcytosis, adsorptive-mediated transcytosis, and cell-mediated transcytosis (ZHOU; SMITH; LIU, 2021; BAGCHI et al., 2019). Figure 10 illustrates these mechanisms. These mechanisms have complex particularities, making it difficult to understand which properties contribute more to the uptake of peptides.

2.4 Molecular descriptors and their influence on biomembranes uptake

The molecular descriptors are properties calculated computationally or experimentally from molecules and can describe them according to structural and physicochemical features. Some of these descriptors have been studied and used by industry as a filter to select oral druggable compounds, relating their pharmacokinetics properties with bioavailability (DOAK et al., 2014), and posteriorly were investigated specifically for peptides (DÍAZ-EUFRACIO et

Figure 10 – Schematic representation of penetration mechanisms for peptides internalization across the BBB.



Source: Zhou, Smith, and Liu (2021)

al., 2018; MANAVALAN et al., 2018; QIANG et al., 2018; PANDEY et al., 2018).

The structural and physicochemical descriptors approached in the cell membrane case are: molecular weight (MW), topological polar surface area (TPSA), water-octanol partition coefficient (LogP), Number of hydrogen bond acceptors (HBA) and donors (HBD), number of aromatic rings (NAR), number of rotatable bonds (NRB), Fraction of sp^3 -hybridized carbon atoms (F_{sp^3}), number of guanidinium groups (NG), net charge (NetC), number of negatively charged amino acids (NNCAA) at pH = 7.4.

MW is a measure of the sum of the atomic weight values of the atoms in a molecule. TPSA is a 2-dimensional approximation of the surface sum over all polar atoms in molecules, primarily oxygen and nitrogen, also including their attached hydrogen atoms (VEBER et al., 2002). The LogP represents the partition coefficient, which is the ratio of the concentration of the compound in octanol to its concentration in water (ARNOTT; PLANEY, 2012). HBA is a measure of the hydrogen-bonding ability of a molecule expressed in terms of the number of possible hydrogen-bond acceptors, while HBD is the same measure for possible hydrogen-bond donors (ARUNAN et al., 2011). The NAR counts the number of benzene rings, while NRB is the number of bonds that allow free rotation around themselves (VEBER et al., 2002). F_{sp^3} is the relation of sp^3 -hybridized carbon atoms by total carbon atoms (LOVERING, 2013). NG is the count of guanidine molecules $HNC(NH_2)_2$ (HANNON; ANSLYN, 1993). NetC and NNCAA represent the charge of a molecule and the sum of glutamic acids, aspartic acids, respectively, and c-terminals, which are negatively charged residues at pH 7.4.

The other group of molecular descriptors approached in the cell membrane case is the sequence-based descriptors such as amino acid composition (AAC), dipeptide composi-

tion (DPC), and pseudo-amino acid composition (PseAAC). The AAC is the percentage of an individual amino acid in the given sequence, this study focuses on the composition of arginine ($f(\text{Arg})$) and lysine ($f(\text{Lys})$), which can be computed respectively by

$$f(\text{Arg}) = \frac{\text{number of arginine residues}}{\text{total residues count}}, \quad (1)$$

$$f(\text{Lys}) = \frac{\text{number of lysine residues}}{\text{total residues count}}. \quad (2)$$

The DPC comprises the total number of i -th dipeptide normalized against all possible combinations of dipeptides in a given peptide sequence. DPC can be computed by

$$\text{DPC}(i) = \frac{\text{total number of dipeptides (i)}}{\text{total number of all possible dipeptides}}. \quad (3)$$

The PseAAC encompasses the relation between the frequency of a given amino acid residue and physicochemical properties such as hydrophobicity, hydrophilicity, and side-chain mass along with the local sequence order (CHOU, 2001). PseAAC can be computed by

$$\text{PseAAC}_j = \frac{1}{L-j} \sum_{i=1}^{L-j} \theta(R_i, R_{i+j}), \quad (4)$$

where L is the total residues content in peptide, R_i is the i -th amino acid residue, and $j \in [1; 20+\lambda_p]$ is the j -th descriptor of PseAAC. λ_p is the correlation factor that reflects the sequence order of all the most contiguous residues along a protein chain. The correlation function $\theta(\cdot)$ of the amino acid residues can be calculated by

$$\theta(R_i, R_{i+j}) = \frac{1}{3} \{ [H_1(R_i) - H_1(R_{i+j})]^2 + [H_2(R_i) - H_2(R_{i+j})]^2 + [M(R_i) - M(R_{i+j})]^2 \}, \quad (5)$$

where H_1 is the hydrophobicity of the i -th residue, H_2 represents the hydrophilicity of the residue, and M is the side-chain mass of the residue.

Some of the physicochemical and structural descriptors selected to evaluate the permeation into the cell by peptides comprise a list of features investigated by other researchers in oral drug design. The influence of molecular weight, TPSA, and lipophilicity in cell membrane permeation of peptides by passive diffusion is reviewed in the work of Dougherty, Sahni, and Pei (2019). The properties of RO5 also were investigated for cell permeability of drugs and clinical candidates in the work of Doak et al. (2014). In many works, sequence-based descriptors have been explored due to their association with peptide charge, which can be an interesting property to correlate with the polarity of the cell membrane in permeation mechanisms. Furthermore, to justify the choice of fractions of arginine and lysine in AAC, these two residues

are predominant in cationic CPPs (GUIDOTTI; BRAMBILLA; ROSSI, 2017; KARDANI et al., 2019). The descriptors used to predict the B3PPs also encompass physicochemical and structural properties extracted from the peptides and can be divided into two groups. The first group comprises molecular properties such as LogP, octanol-water distribution coefficient at pH 7.4 (LogD), TPSA, HBA, HBD, oxygen count nO, nitrogen count nN, and nitrogen and oxygen count n(N+O). The second group is composed of 1428 descriptors from Mordred package (MORIWAKI et al., 2018), where five descriptors are highlighted on the results of this thesis: 12-or-greater-membered aromatic hetero ring count (nG12Ring), Geary coefficient of lag 8 weighted by polarizability (GATS8p), Geary coefficient of lag 5 weighted by polarizability (GATS5p), mean topological charge index of order 9 (JGI9), and MOE⁵ Charge VSA Descriptor 4 (PEOE-VSA4).

The Mordred descriptors are composed of a variety of structural, physicochemical, and topological descriptors, which also were used to evaluate the permeation of small molecules across the BBB (LI et al., 2005) and correlated to cell membrane uptake of peptides (STALMANS; WYNENDAELE, et al., 2013). The Geary coefficient (GATS) is a general index of 2D-autocorrelation applied to a molecular graph, which describes the topology of the peptide in association with atomic masses, polarizabilities, and Sanderson electronegativities. This index is calculated using the topological distance and the Kronecker delta (TODESCHINI; CONSONNI, 2000; STALMANS; WYNENDAELE, et al., 2013). Charge index (JGI) is a topological descriptor with the ability to describe the molecular charge distribution. It was proposed to evaluate the charge transfer between pairs of atoms and, therefore, the global charge transfer in the molecule (GALVEZ et al., 1994). PEOE-VSA represents the partial charge descriptor calculated by the sum of the proximate accessible van der Waals surface area (VSA) for each atom over all the atoms (REDDY; KUMAR; GARG, 2010).

The first group of descriptors selected to investigate their correlation to BBB uptake encompasses features tested as a chemical filter to predict the permeability of a range of small molecules across this biomembrane using the LogBB⁶ as indicator (DICHIARA et al., 2020). Other studies also tried to establish a general filter to predict this pharmacokinetics property for small molecules evaluating intervals for MW, weak hydrogen bond, LogP, and TPSA (ABRAHAM, 2004; LALATSA; SCHATZLEIN; UCHEGBU, 2014; DAINA; ZOETE, 2016). The descriptors from the second group were explored recently in the work of Plisson and Piggott (2019), whose features TPSA and PEOE-VSA achieved the best values of importance in the prediction of blood-brain barrier permeability of marine-derived kinase inhibitors.

⁵MOE is the acronym for Molecular Operating Environment software.

⁶LogBB is the concentration of drug in the brain divided by the concentration in the blood

2.5 Machine learning algorithms

Artificial intelligence is a branch of computer science that explores the automation of human intelligent behavior to perform tasks (CHOWDHARY, 2020). Currently, AI a variety of subfields dedicated to solving specific problems by developing computer algorithms based on human cognitive aspects. These aspects include the ability to learn from past experiences, recognize patterns, perform reasoning, make decisions, understand natural language, and interact adaptively in diverse contexts (RUSSELL; NORVIG, 2021).

Machine learning (ML) is one of the fields covered within the study of AI. This field can be defined as the automatic process of extracting patterns through observation of data (KELLERHER; MAC NAMEE; D'ARCY, 2020). In other words, machine learning is a field of study that aims to develop computational models capable of learning patterns through data observation and using these models to solve problems (RUSSELL; NORVIG, 2021). The use of ML models has considerably increased in the last decades and can be used in problems for which some solutions require a very extensive set of rules, for complex problems for which there is no good solution using traditional approaches, in problems that require adaptation of the model depending on the variability of the problem, or even a search for insights into the problems addressed and large amounts of data (GÉRON, 2019; GARG; MAGO, 2021).

The field of machine learning can be divided based on the form of learning that will be used in the model, which is directly related to the nature of the problem and the type of data used in the learning process (CHOWDHARY, 2020; MÜLLER; GUIDO, 2016). The subfields covered in the study of ML models can be divided into the following categories according to the type of learning process: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning.

In supervised learning the ML model observes input-output pairs aiming at learning a mapping that properly correlates them (RUSSELL; NORVIG, 2021). In other words, a set of descriptive features from a database is used as input, and their corresponding target feature (label) is used as output for training a model. After this learning process, the trained model can be used for predicting the output generated by other samples using the same set of features. Classification and regression are the two major classes of problems approached in the supervised machine learning field (MÜLLER; GUIDO, 2016).

Unsupervised learning encompasses the learning process of ML with data based exclusively on the input information without supporting a target feature. In essence, the class of algorithms learns a pattern from the data without making an association with targets (JAMES et al., 2013). Clustering, anomaly detection, dimensionality reduction, and association learning are the main classes of problems approached by unsupervised learning models (MÜLLER; GUIDO, 2016; PATEL, 2019).

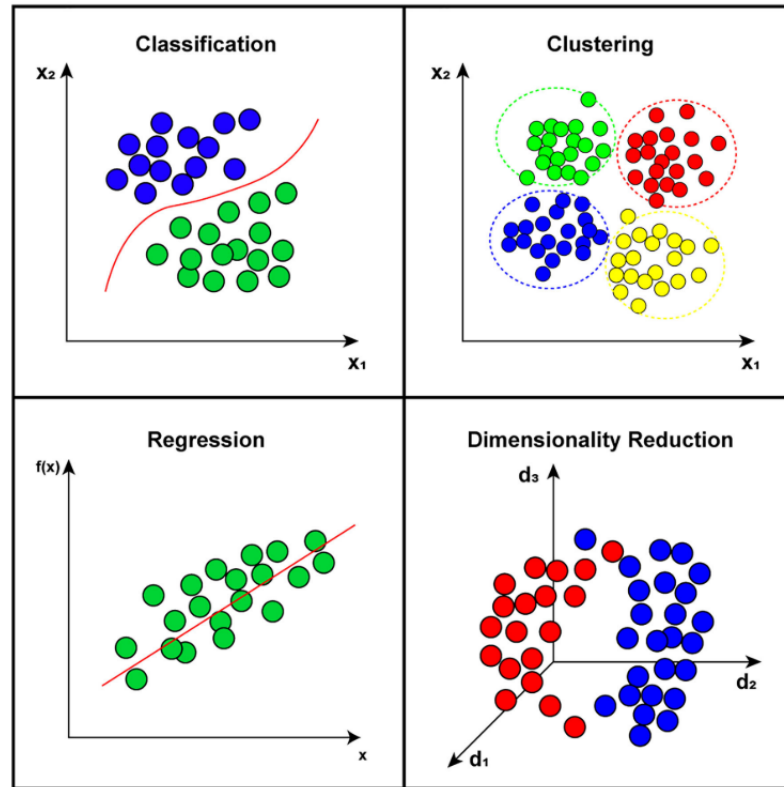
Semisupervised learning is the field of machine learning that approaches strategies us-

ing a combination of supervised and unsupervised learning models to deal with partially labeled data, usually with few labeled data and many unlabeled data. The ML are trained with labeled and unlabeled data to learn the pattern and perform predictions. Classification, regression, constrained clustering, and dimensionality reduction are examples of problems approached in semisupervised learning (ZHU; GOLDBERG, 2009).

Reinforcement learning is a field distinct from supervised and unsupervised learning. This learning process is based on an agent that can observe the environment, select and perform actions, and get rewards or penalties to learn a pattern (RUSSELL; NORVIG, 2021; MOREIRA, 2022). The agent must then learn by itself what is the best strategy to get the most reward over time in a given situation (GÉRON, 2019).

Different computational problems have been approached and solved by ML algorithms based on the type of learning. Classification, time series regression, natural language processing, optimization, clustering, and dimensionality reduction (DR) are the main classes of problems approached by machine learning. For example, classification and regression problems can be solved with artificial neural network (ANN), deep learning (DL), k-nearest neighbors (kNN), support vector machine (SVM), decision tree (DT), and random forest (RF) (T.K.; ANNAVARAPU; BABLANI, 2021). Clustering problems can be treated using k-means, hierarchical cluster analysis (HCA), and DBSCAN. Visualization and dimensionality reduction problems can be solved using principal component analysis (PCA), locally-linear embedding (LLE), and t-distributed stochastic neighbor embedding (t-SNE) (ROOHI et al., 2020; USAMA et al., 2019). Figure 11 illustrates some of the aforementioned problems approached and solved by ML models.

Figure 11 – Representation of main categories of problems approached by ML, such as classification, regression, clustering, and dimensionality reduction.

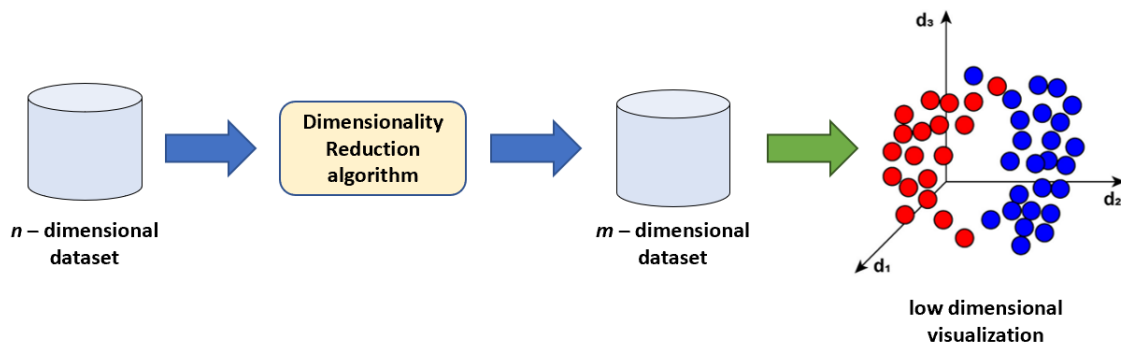


Source: Author's own.

2.5.1 Dimensionality reduction

Dimensionality reduction can be defined as a subfield of machine learning that addresses the development of algorithms dedicated to reducing the number of features in a database to a smaller number of features. In summary, DR algorithms are capable of mapping an n -dimensional dataset into an m -dimensional subspace, i.e., $\mathbb{R}^n \rightarrow \mathbb{R}^m \{n, m \in \mathbb{Z}^+ \mid n > m\}$ (GHOJOGH et al., 2023). Figure 12 illustrates an example of the DR process and the visualization of the low-dimensional projected data.

Figure 12 – Schematic of dimensionality reduction process and low-dimensional data visualization.



Source: Author's own.

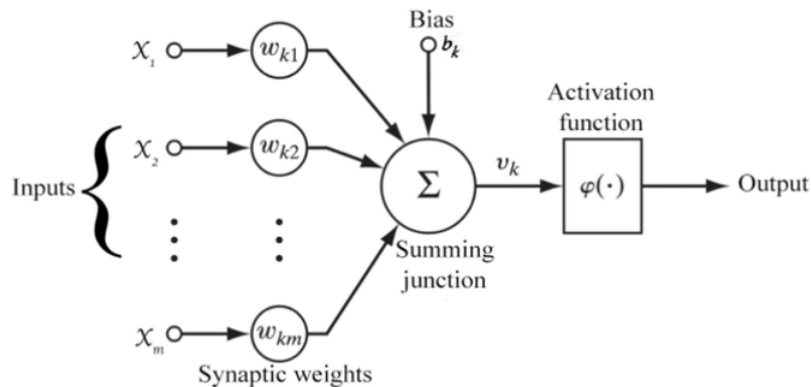
Dimensionality reduction is very useful in extracting the most relevant characteristics from high-dimensional data, besides enabling projection into a subspace capable of being visualized. DR is employed in many fields of quantitative science to obtain a representative feature space that, by being more susceptible to the problem domain, best describes data and preserves important structural properties such as dissimilarities, cluster shapes, probability distributions, and neighboring relationships, besides allowing better visualization (FLEXA et al., 2019). Currently, there are some algorithms that perform DR, i.e., PCA, kernel PCA, LLE, multidimensional scaling (MDS), and linear discriminant analysis (LDA) (GÉRON, 2019).

Manifold learning is a subfield of dimensionality reduction that involves the development of DR algorithms capable of exploring the geometric properties of data projected into low dimensions. Unlike the algorithms mentioned above, which focus on preserving variance, manifold learning assumes that high-d data has a low-d projection. Therefore, its algorithms attempt to find a low-d representation of the data that is representative and preserves the geometric and nonlinear characteristics of the high-d data (CHAO; LUO; DING, 2019; LUNGA et al., 2014). Isomap, LLE, t-SNE, Laplacian eigenmaps, and Hessian eigenmaps are some examples of algorithms approached in studying manifold learning (FU, 2011).

2.5.2 Artificial neural network

Artificial neural network (ANN) is a ML model introduced in 1943 by the neurophysiologist Warren McCulloch and the mathematician Walter Pitts, where they proposed a computational model based on the functioning of biological neurons to perform complex computations using propositional logic (MCCULLOCH; PITTS, 1943). In 1957, Frank Rosenblatt created the artificial neuron called *Perceptron*, one of the simplest ANN architectures. An artificial neuron is a basic unit of information processing in a ANN (HAYKIN, 1998). Figure 13 illustrates the block diagram of the model of the artificial neuron, which can process the information by performing the sum of the inputs associated with weights and passing this result through a function.

Figure 13 – Schematic of the artificial neuron.



Source: Adapted from Haykin (1998)

The output ($y_k \in \mathbb{R}$) of a single neuron represented in Figure 13 can be computed by

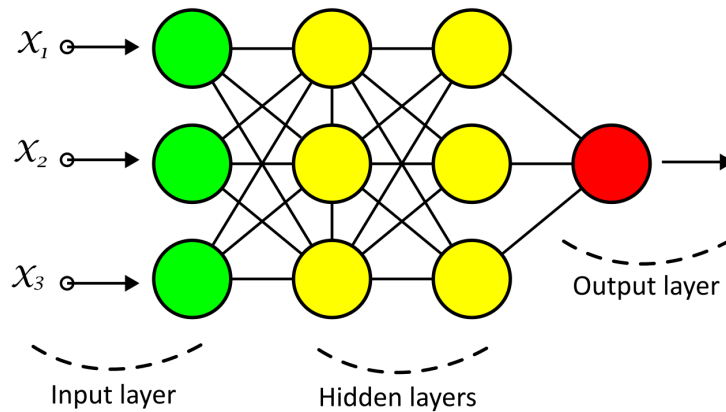
$$y_k = \phi\left(\sum_{j=1}^m w_{kj}x_j + b_k\right), \quad (6)$$

where $w_{kj} \in \mathbb{R}$ is the synaptic weight of the k -th neuron related to j -th input ($x_j \in \mathbb{R}$), $b_k \in \mathbb{R}$ represents the bias and $\phi(\cdot)$ is the activation function. In summary, the synaptic weights are responsible for weighting their respective input values, while the bias provides more flexibility for the sum of the weighted inputs. The summation unit is responsible for creating a linear combination of the weighted inputs and the activation function is employed for constraining the output of a neuron.

Currently, there are many possibilities of activation functions to use in a neuron: hyperbolic tangent, sigmoid, ReLu, softmax, linear, hard limit (EMMERT-STREIB et al., 2020; HAGAN et al., 2014). In general, these activation functions are responsible for limiting the output of the neuron to the closed unit interval $[0, 1]$ or the interval $[-1, 1]$. The bias is a mathematical term used to increase or decrease the liquid value of the input to the activation function (HAYKIN, 1998).

The multilayer perceptron (MLP) is one of the most famous ANN architecture and is the structure selected to perform prediction of CPPs and B3PPs in this thesis. This architecture is based on the connection of many single neurons in layers where the signals flow only from the inputs to the outputs, so this architecture is an example of a feedforward neural network (GÉRON, 2019). Figure 14 illustrates a schematic of a generic MLP, which is composed of an input layer, hidden layers, and an output layer. The green circle represents an input of the network, the yellow circle symbolizes an individual artificial neuron fully connected in a hidden layer, and the red circle represents the output of the MLP.

Figure 14 – Schematic of a MLP architecture.



Source: Author's own.

The adjustment of all synaptic weights and biases of each neuron present in a MLP

is performed by a training algorithm. This is an iterative process capable of making a neural network learn from its environment and adjusts its parameters with the objective of improving performance in solving a given problem. In 1986, Rumelhart, Hinton, and Williams introduced the *Backpropagation*, the first algorithm capable of efficiently training multilayer networks (RUMELHART, DAVID E. HINTON; WILLIAMS, 1986). This algorithm uses gradient descent to adjust the weights and biases of the network by minimizing the error signal generated by the difference between the current output of the network and the desired output and back-propagated against the direction of the network (HAYKIN, 1998). Currently, other algorithms are also employed for training MLPs, such as the Newton method, quasi-Newton optimization, and Levenberg-Marquardt (KINGMA; BA, 2015; CÖMERT; KOCAMAZ, 2017).

2.5.3 Gaussian process classifier

A *Gaussian* process is a conditional probability on a multivariate *Gaussian* distribution used to solve regression and classification issues (OPPER; WINTHER, 2000). The *Gaussian* process classifier (GPC) is a supervised ML algorithm based on Bayesian probability theory that assumes that there is a relationship between the input samples and its label based on a *Gaussian* distribution (EBDEN, 2015; RASMUSSEN; WILLIAMS, 2006). In summary, GPC models the probability of a sample belonging to a class based on a conditional probability modeled by a *Gaussian* distribution described by

$$P(y|\mathbf{x}) = \mathcal{N}(\eta, \sigma^2), \quad (7)$$

where $P(y|\mathbf{x})$ represents the conditional probability of a sample $\mathbf{x} \in \mathbb{R}^n$ belonging to class y , and $\eta \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}$ are the mean and variance of the distribution, respectively. The mean of the *Gaussian* distribution can be computed by

$$\eta = \mathbf{k}_* \mathbf{K}^{-1} y, \quad (8)$$

where $\mathbf{k}_* \in \mathbb{R}^{N_s}$ and $\mathbf{K} \in \mathbb{R}^{N_s \times N_s}$ are the kernel vector and the kernel matrix, respectively. The variance of the distribution can be calculated by

$$\sigma^2 = k_{**} - \mathbf{k}_* \mathbf{K}^{-1} \mathbf{k}_*^T, \quad (9)$$

where $k_{**} \in \mathbb{R}$ is the prior covariance (RASMUSSEN; WILLIAMS, 2006). The vector \mathbf{k}_* stores the information of similarity between a new sample $\mathbf{x}^* \in \mathbb{R}^n$ related to all the training samples, as shown in

$$\mathbf{k}_* = [k(\mathbf{x}^*, \mathbf{x}_1), k(\mathbf{x}^*, \mathbf{x}_2), \dots, k(\mathbf{x}^*, \mathbf{x}_{N_s})], \quad (10)$$

where $k(\cdot)$ represents a kernel function and $N_s \in \mathbb{Z}^+$ is the total of training samples. The kernel matrix \mathbf{K} stores the covariance among all the training samples using a kernel function, as demonstrated by

$$\mathbf{K} = \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & k(\mathbf{x}_1, \mathbf{x}_2) & \dots & k(\mathbf{x}_1, \mathbf{x}_{N_s}) \\ k(\mathbf{x}_2, \mathbf{x}_1) & k(\mathbf{x}_2, \mathbf{x}_2) & \dots & k(\mathbf{x}_2, \mathbf{x}_{N_s}) \\ \vdots & \vdots & \ddots & \vdots \\ k(\mathbf{x}_{N_s}, \mathbf{x}_1) & k(\mathbf{x}_{N_s}, \mathbf{x}_2) & \dots & k(\mathbf{x}_{N_s}, \mathbf{x}_{N_s}) \end{bmatrix}. \quad (11)$$

The prior covariance k_{**} is computed using a kernel function for the new sample, as shown by

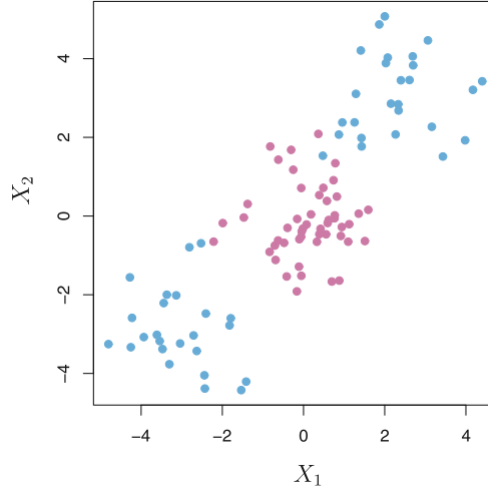
$$k_{**} = k(\mathbf{x}_*, \mathbf{x}_*). \quad (12)$$

Currently, there are many kernel functions employed in GPC algorithm, such as radial basis function (RBF), Matérn, rational quadratic, exponential sine squared, polynomial of degree d , dot product (LUNDERMAN et al., 2018; RASMUSSEN; WILLIAMS, 2006).

2.5.4 Support vector machine

The SVM algorithm is a supervised machine learning method used in classification and regression problems. Specifically for classification problems, which is the focus of this thesis, SVM can be understood as an extension of the maximal margin classifier algorithm, which computes the farthest separating hyperplane that separates the training observations classes (JAMES et al., 2013). SVM utilizes a kernel to generate a boundary function capable of separating groups of samples that are not linearly separable in order to classify them. Figure 15 exemplifies a group of samples belonging to two classes (blue and red), which cannot be linearly separated by a hyperplane but can be separated by other non-linear functions. In summary, SVM is the algorithm that provides an optimal boundary function capable of separating groups of samples that normally can not be separated by hyperplane.

Figure 15 – Example of non-linearly separable samples.



Source: Adapted from James et al. (2013)

Mathematically, given $T = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_{N_s}, y_{N_s})\}$ as the training set with input samples and their respective labels, the SVM algorithm employed in classification problems can be defined as an optimization problem described by

$$\text{minimize}_{\alpha} \quad \frac{1}{2} \sum_{i,j=1}^{N_s} y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \alpha_i \alpha_j - \sum_{j=1}^{N_s} \alpha_j, \quad (13)$$

$$\begin{aligned} \text{subject to} \quad & \sum_{j=1}^{N_s} y_j \alpha_j = 0, \\ & 0 \leq \alpha_i \leq C, \quad i = 1, \dots, N_s, \end{aligned} \quad (14)$$

where $\mathbf{x}_i \in \mathbb{R}^n$ represents the i -th input sample, $y_i \in \{0, 1\}$ is the class of the i -th sample, and $N_s \in \mathbb{Z}^+$ represents the number of samples. The term α_i is the i -th element of the Lagrange multipliers vector $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_{N_s})^T$, $k(\cdot)$ is the kernel function that computes the similarity between two samples \mathbf{x}_i and \mathbf{x}_j , and $C \in \mathbb{R}^+$ is the penalty parameter that controls the trade-off between maximizing margin and minimizing classification error (DENG; TIAN; ZHANG, 2013). The solution for the optimization problem expressed above is a boundary function ($g(\mathbf{x})$), as shown by

$$g(\mathbf{x}) = \sum_{i,j=1}^{N_s} y_i \alpha_i^* k(\mathbf{x}_i, \mathbf{x}_j) + b^*, \quad (15)$$

where $\boldsymbol{\alpha}^* = (\alpha_1^*, \dots, \alpha_{N_s}^*)^T$ is a Lagrange multipliers vector chosen by the algorithm for calculating the term b^* by

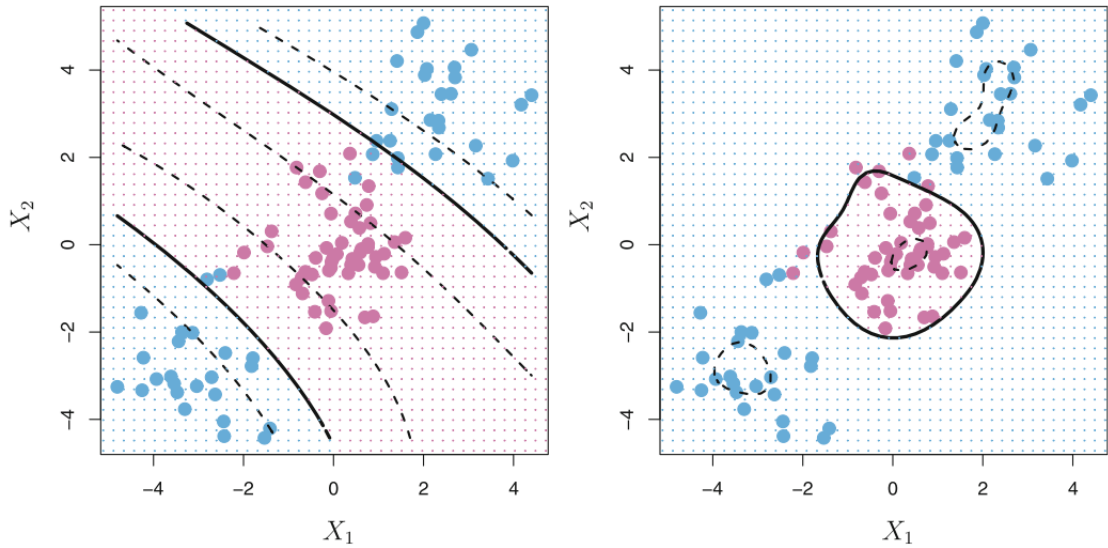
$$b^* = y_j - \sum_{i,j=1}^{N_s} y_i \alpha_i^* k(\mathbf{x}_i, \mathbf{x}_j). \quad (16)$$

The decision function ($f(\mathbf{x})$) used to classify the sample according to the result obtained by the boundary function can be expressed by

$$f(\mathbf{x}) = \begin{cases} 1, & g(\mathbf{x}) \geq 0; \\ 0, & g(\mathbf{x}) < 0; \end{cases} \quad (17)$$

Currently, many kernel functions have been employed in SVM to improve the performance in non-linear classification problems. Some examples of these functions are: polynomial of degree p , sigmoid, Dirichlet, Gaussian RBF (GHOLAMI; FAKHARI, 2017). Figure 16 illustrates an example of applying the SVM with a polynomial kernel of degree 3 and RBF kernel to generate the boundary function to classify the samples shown in Figure 15.

Figure 16 – Example of SVM using different kernels to classify non-linearly separable data. Left: SVM with a polynomial kernel of degree 3. Right: SVM with RBF kernel.



Source: Adapted from James et al. (2013)

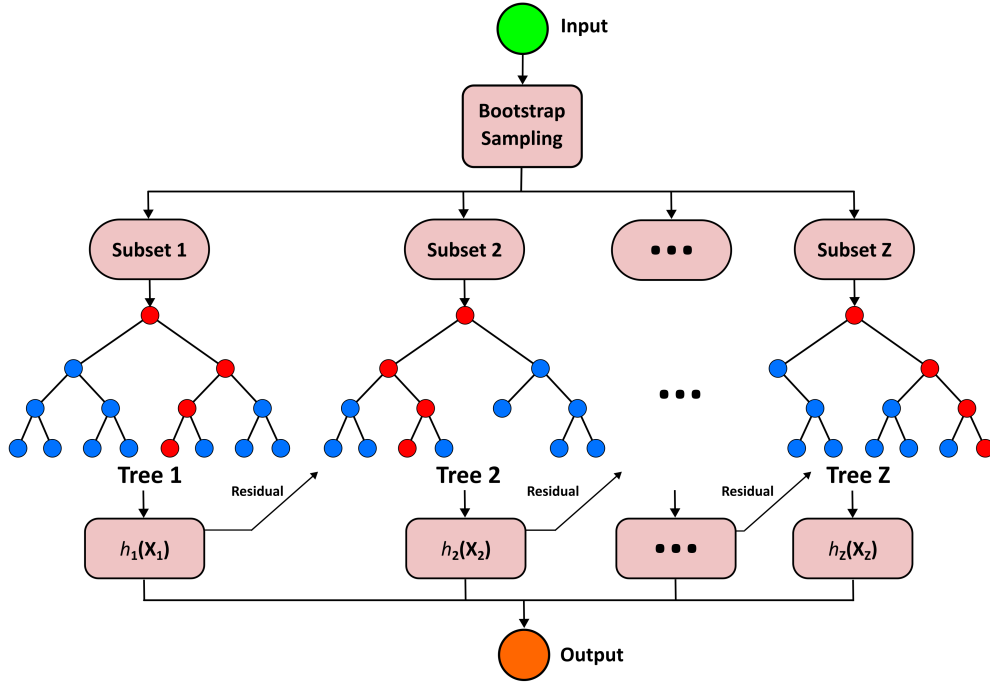
2.5.5 Extreme gradient boosting

Extreme gradient boosting (XGBoost) is a decision-tree-based ML ensemble⁷ technique that uses a gradient-boosting framework for grouping estimators. This algorithm was proposed by Chen and Guestrin (2016) as a scalable machine learning system for solving regression, classification, and ranking problems. Figure 17 illustrates an example of XGBoost composed of Z decision trees.

⁷Ensemble corresponds to techniques that combine multiple machine learning models to improve overall system performance.

In summary, XGBoost uses Z subsets of input information, which were generated by bootstrap sampling⁸, for training its trees. However, starting from the second tree, each DT consecutively uses the residual error from the previous tree's prediction to optimize its training in a known boosting process (ARIF ALI et al., 2023). In the end, in a classification problem, the XGBoost prediction result is based on the majority class predicted among all decision trees.

Figure 17 – Schematic of XGBoost algorithm.



Source: Author's own.

Mathematically, given $T = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_{N_s}, y_{N_s})\}$ as the training set with input samples and their respective labels, the training process of each tree in the XGBoost algorithm can be understood as an optimization problem that seeks to minimize the objective function based on a set of leaf weights of the tree, as shown by

$$\text{minimize}_{w_m} \sum_{j=1}^Z l(y_j, h_j(\mathbf{x}_j) + \hat{h}_{j-1}^{(t-1)}) + \Omega, \quad (18)$$

where $l(\cdot)$ represents the convex loss function that measures the difference between the target $y_j \in \mathbb{R}$ and the prediction $h_j(\mathbf{x}_j) \in \mathbb{R}$ from the j -th tree. The vector $\mathbf{x}_j \in \mathbb{R}^n$ represents the input information provided by the j -th subset for $j=1, \dots, Z$. The term $\hat{h}_{j-1}^{(t-1)} \in \mathbb{R}$ represents the previous tree prediction in the t -th iteration. Ω is the complement of the objective function shown in Equation 18 and represents the regularization term used to penalize the complexity

⁸Bootstrap sampling is a statistical technique used to draw random samples with replacement from the original dataset repeatedly.

of the model, which helps to smooth the final learned weights to avoid over-fitting (CHEN; GUESTRIN, 2016). Ω can be computed by

$$\Omega = \sum_{q=1}^{L_q} \left(\gamma_g L_q + \frac{1}{2} \lambda_g w_m^2 \right), \quad (19)$$

where $\gamma_g \in \mathbb{R}$ and $\lambda_g \in \mathbb{R}$ are coefficients used to control the model complexity and the output of the objective function. The number of leaf nodes in the tree is represented by $L_q \in \mathbb{Z}^+$ and $w_q \in \mathbb{R}$ is the leaf weight of the q -th leaf node in the tree for $q = 1, \dots, L_q$.

The XGBoost algorithm uses a second-order Taylor approximation of the loss function and speeds up the process of searching for the global minimum through the first and second derivatives of the loss function (LIANG et al., 2021). This approximation helps to find the optimum value for the q -th leaf weight of the j -th tree, which can be computed by

$$w_q = \frac{\sum_{q=1}^j G_q}{\sum_{q=1}^j H_q + \lambda_g}, \quad (20)$$

where $G_q \in \mathbb{R}$ and $H_q \in \mathbb{R}$ are the first and the second derivative of the loss function ($l(\cdot)$), which can be respectively calculated by

$$G_q = \frac{\partial l(y_q, \hat{h}_{q-1}^{(t-1)})}{\partial \hat{h}_{q-1}^{(t-1)}}, \quad (21)$$

$$H_q = \frac{\partial^2 l(y_q, \hat{h}_{q-1}^{(t-1)})}{\partial^2 \hat{h}_{q-1}^{(t-1)}}. \quad (22)$$

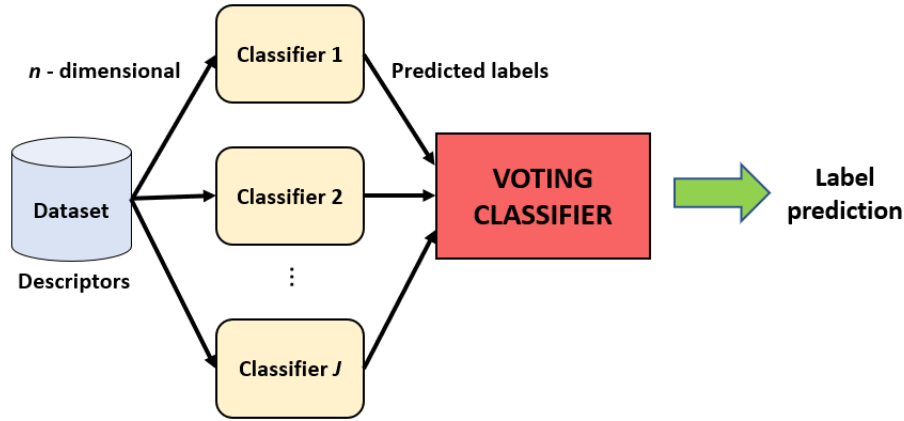
Finally, after the training of the XGBoost algorithm using the steps described above, the predicted output ($\hat{h}(\mathbf{x})$) for a given input can be computed for each tree. In a classification problem, the XGBoost algorithm classifies the sample according to the majority class label predicted by all the trees.

Other important XGBoost parameters that can be highlighted are the learning rate, the number of trees, and the depth of each tree. The learning rate is a factor applied to each tree and performs as a step-size shrinkage used in the update to prevent overfitting. The number of trees is the count of the model's decision tree, and the depth is related to the number of consecutive nodes in a tree (WADE, 2020).

2.5.6 Voting classifier based on machine learning

The voting classifier (Vcf) is a ML architecture to group classifiers to improve the classification task, using a majority vote or the average predicted probabilities (soft vote) to predict the class labels (GÉRON, 2019). Figure 18 illustrates the general architecture of the Vcf.

Figure 18 – General structure of voting classifier.



Source: Author's own.

In summary, the final result for the prediction by Vcf is the statistical mode of the set of predicted labels by individual classifiers. For instance, considering five binary classifier groups in a voting classifier, if three models predict the data as belonging to class 1, while the other two classifiers predict label 0 for the same data, the voting classifier considers as the final result the most frequent result, that is, the final result is label 1.

2.5.7 Supervised Laplacian eigenmaps

The supervised Laplacian eigenmaps (sLE) is a supervised manifold dimensionality reduction technique proposed by Raducanu and Dornaika (2012). This algorithm uses class labels to guide the non-linear mapping of the high-dimensional data to the embedded space by large margin concept. The labels in this algorithm allow splitting the graph Laplacian associated with the data into two components: within-class and between-class graphs. This proposal provides important properties when compared with canonical Laplacian eigenmaps, such as adaptive estimation of the local neighborhood surrounding each sample based on data density and similarity, besides maximizing the local margin between heterogeneous samples and pushing the homogeneous samples closer to each other simultaneously by the objective function. In summary, the sLE perform the dimensionality reduction of labeled data, clustering closer data with the same labels and distancing samples from different labels. All the steps of this algorithm are described below.

Mathematically, given the training set $T = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_{N_s}, y_{N_s})\}$ with input samples and their respective labels, the first step of sLE is the calculation of average similarity ($AS(\mathbf{x})$). This coefficient indicates the level of proximity of an input sample in relation to the other training samples in n -dimensional space. $AS(\mathbf{x})$ can be calculated by

$$AS(\mathbf{x}_i) = \frac{1}{N_s} \sum_{j=1}^{N_s} \exp \left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\beta} \right), \quad (23)$$

where $\mathbf{x}_i \in \mathbb{R}^n$ is the i -th input sample, $N_s \in \mathbb{Z}^+$ is the number of samples in the training set and $\beta \in \mathbb{R}$ is the average of squared distances between all pairs of samples.

The second step encompasses the great contribution of this algorithm, where two subsets of samples are computed using the similarity between pairs of samples. The first subset is the $N_w(\mathbf{x}_i)$, which represents the set of within-class neighbors samples in relation to the \mathbf{x}_i . This subset can be expressed by

$$N_w(\mathbf{x}_i) = \left\{ \mathbf{x}_j | y_j = y_i, \exp \left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\beta} \right) > AS(\mathbf{x}_i) \right\}, \quad (24)$$

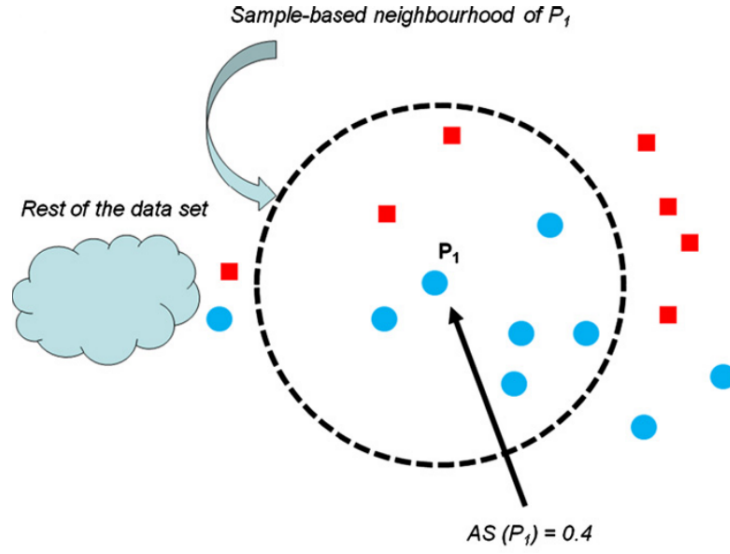
where y_j represents the class label of the j -th sample. In summary, N_w represents the subset of neighbors of the \mathbf{x}_i , which is composed of all the samples \mathbf{x}_j that belong to the same class of \mathbf{x}_i ($y_j = y_i$) and have similarity with this sample greater than $AS(\mathbf{x}_i)$.

The second subset is the $N_b(\mathbf{x}_i)$, which represents the set of between-class neighbor samples in relation to the \mathbf{x}_i . In summary, N_b is composed of all the samples that do not belong to the same class of \mathbf{x}_i ($y_j \neq y_i$) and have similarity with this sample greater than $AS(\mathbf{x}_i)$. This subset can be expressed by

$$N_b(\mathbf{x}_i) = \left\{ \mathbf{x}_j | y_j \neq y_i, \exp \left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\beta} \right) > AS(\mathbf{x}_i) \right\}. \quad (25)$$

Figure 19 illustrates an example of the process for constructing the subsets $N_w(\mathbf{x}_i)$ and $N_b(\mathbf{x}_i)$, where the sample \mathbf{P}_1 (blue ball) has an average similarity equal to 0.4. The other blue balls inside the dashed circle have similarities greater than $AS(\mathbf{P}_1)$. Therefore, these samples are included in $N_w(\mathbf{x}_i)$, while the red square samples, which also have a similarity greater than average similarity and have distinct labels from the blue ball, are included in the subset $N_b(\mathbf{x}_i)$. This process exemplifies how the neighborhoods are formed in this algorithm according to each label.

Figure 19 – Schematic representation of sample-based neighbor computation in sLE.



Source: Raducanu and Dornaika (2012)

The construction of N_w and N_b in the previous step is essential to calculate the weight matrices $\mathbf{W}_w \in \mathbb{R}^{N_s \times N_s}$ and $\mathbf{W}_b \in \mathbb{R}^{N_s \times N_s}$, which represent how the similarity weighting is distributed for each subset and are used in the equation of sLE algorithm. \mathbf{W}_w and \mathbf{W}_b can be respectively computed using by

$$\mathbf{W}_{w,ij} = \begin{cases} \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\beta}\right) & \text{if } \mathbf{x}_j \in N_w(y_i) \text{ or } \mathbf{x}_i \in N_w(\mathbf{x}_j), \\ 0 & \text{otherwise,} \end{cases} \quad i, j = 1, \dots, N_s \quad (26)$$

$$\mathbf{W}_{b,ij} = \begin{cases} 1 & \text{if } \mathbf{x}_j \in N_b(\mathbf{x}_i) \text{ or } \mathbf{x}_i \in N_b(\mathbf{x}_j), \\ 0 & \text{otherwise.} \end{cases} \quad i, j = 1, \dots, N_s \quad (27)$$

The next step of supervised Laplacian eigenmaps relies on calculating the Laplacian matrices $\mathbf{L}_w \in \mathbb{R}^{N_s \times N_s}$ and $\mathbf{L}_b \in \mathbb{R}^{N_s \times N_s}$ according to

$$\mathbf{L}_w = \mathbf{D}_w - \mathbf{W}_w, \quad (28)$$

$$\mathbf{L}_b = \mathbf{D}_b - \mathbf{W}_b, \quad (29)$$

where $\mathbf{D}_w \in \mathbb{R}^{N_s \times N_s}$ and $\mathbf{D}_b \in \mathbb{R}^{N_s \times N_s}$ are the diagonal weight matrices and are formed by the sum of column (or row) of \mathbf{W}_w and \mathbf{W}_b , respectively.

The final mathematical expression of the sLE that calculates the low-d data in m -dimensional space ($m < n$) is shown in

$$\mathbf{B}\mathbf{Z}_D = \boldsymbol{\lambda}\mathbf{D}_w\mathbf{Z}_D, \quad (30)$$

where $\mathbf{Z}_D \in \mathbb{R}^{N_s \times m}$ and $\boldsymbol{\lambda} \in \mathbb{R}^m$ are the eigenvectors matrix and their respective eigenvalues vector calculated for the matrices $\mathbf{B} \in \mathbb{R}^{N_s \times N_s}$ and \mathbf{D}_w . In summary, each eigenvector of the matrix $\mathbf{Z}_D = (\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_m)$ computed by Equation 30 represents each dimension of the desired low-d data in the embedded space. The matrix \mathbf{B} represents a filter between the contribution of \mathbf{L}_b and \mathbf{W}_w and can be calculated by

$$\mathbf{B} = \gamma_s \mathbf{L}_b + (1 - \gamma_s) \mathbf{W}_w, \quad (31)$$

where γ_s is a real scalar hyperparameter that belongs to $[0, 1]$, and the terms \mathbf{L}_b and \mathbf{W}_w are the matrices calculated by Equations 29 and 26, respectively.

2.6 Conclusion

This chapter presented the main biochemical and computational aspects of peptides, addressing their composition and representation computation and covering some theoretical aspects related to the structure of the cell membrane and the BBB. The chapter also addresses the theoretical aspects related to the molecular descriptors used to predict the penetration of peptides, besides discussing their biochemical relationship with permeability in biomembranes. Another point covered in this chapter is the machine learning techniques, covering some general aspects of classifiers and dimensionality reduction models, besides focusing on the explanation of each of the algorithms used as well in the proposed framework for predicting CPPs and B3PPs as in the other comparable models. The next chapter addresses the methodology of this thesis, describing the construction of the datasets and the architecture of the proposed framework.

3 PROPOSED METHOD

This chapter presents general aspects related to the proposed ML-based framework to predict CPPs and B3PPs. The origin and the preprocessing stages for peptide databases as the molecular descriptors used in the cell membrane and BBB problems are described here. In addition, this chapter also displays the general pipeline for the proposed ML-based framework to predict the CPPs and B3PPs.

3.1 Peptides databases

Databases with experimentally validated information on biomembrane-penetrating peptides are a valuable source for obtaining structure- and sequence-based data for developing computational models to predict these structures' permeability. The present section exhibits the databases of CPPs and B3PPs used in this thesis, besides the preprocessing steps applied to them.

3.1.1 Database for CPPs

For the problem of cell membrane penetrating prediction, this thesis proposes using datasets of peptide structures obtained from curated CPP databases. The CPP structures were obtained from CPPsite2.0, a chemo-structural database with more than 1800 validated experimental CPPs with different structural properties (linear/cyclic; and modified/non-natural residues) and a wide range of applications for cargo transportations into the cell (AGRAWAL et al., 2016). Moreover, 411 CPPs and 411 non-CPPs were obtained from the C2Pred server (TANG et al., 2016). Additionally, 31 CPP and 21 non-CPP structures were obtained from previously published works and pharmaceutical catalogs (SANDERS et al., 2011; PONNAPPAN; CHUGH, 2017; ANASPEC, 2010). Figure 20 illustrates better this stage, where it is possible to see how the peptides were divided according to the labels (CPPs and non-CPPs) and preprocessed before their use in ML models for training and test.

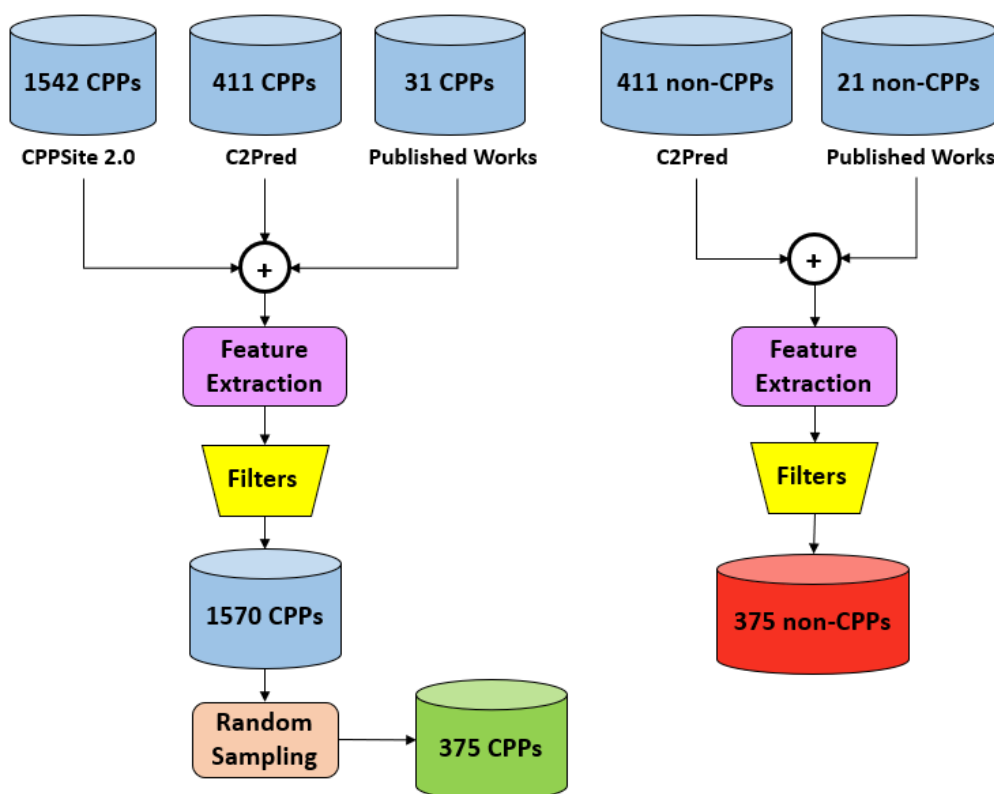
This thesis uses peptides in FASTA and PDB formats, which correspond to primary and tertiary structures, respectively. There is an importance in the use of both FASTA and PDB since the first one encompasses only natural and more accessible peptides' structures. On the other hand, PDB includes information about chemical modifications, although their structure is more complex to obtain. All the peptides from CPPsite2.0 have both primary and tertiary structures, while the remainder has only a primary one. For these molecules that do not have a tertiary structure, their PDB files were obtained using the PEP-FOLD3 server (LAMIABLE et al., 2016).

Figure 20 summarizes the steps related to the construction of the datasets of this case

study. After joining the databases from CPPSite 2.0, C2Pred, and published works according to the peptide labels, the molecular descriptors (features) are extracted from the peptides to compose the datasets of CPPs and non-CPPs.

Figure 20 – Diagram of CPPs and non-CPPs datasets construction. The first step represents the unification of CPPs and non-CPPs in their respective groups. The second step is the extraction of molecular descriptors

. The third step is the filters based on peptide chain length, duplicated structures, and peptides containing descriptors with outliers. Random sampling is employed to select an equivalent number of CPPs regarding non-CPPs.



Source: Author's own.

The next step of the dataset construction relies on filtering the peptides according to some criteria, which is represented by the yellow box in Figure 20. Each dataset is filtered regarding peptide length in the preprocessing stage, limited to 5 up to 30 amino acid residues, besides removing the duplicates among CPPs and non-CPPs (GAUTAM et al., 2013; QIANG et al., 2018). Another filter used in preprocessing was the removal of outliers structures using z-score > 3. This technique calculates the normalized distribution of each feature using the average and the standard deviation of the feature distribution (COUSINEAU; CHARTIER, 2010). According to this criterion, all the peptides with at least one feature with z-score > 3 were removed from the dataset.

After applying the filters, the number of selected CPPs (1570) is greater than non-CPPs (375). To balance the two datasets and avoid overfitting of the ML models, random

sampling is applied to 1570 CPPs hlsamples to select from them 375 peptides. Finally, the evaluation setup has the training dataset with 300 CPPs and 300 non-CPPs and an independent test with 75 CPPs and 75 non-CPPs for PDB format. Specifically for FASTA format experiment, only natural peptides were considered from the same sample space, and the peptides with chemical modification or non-natural residues were removed from the training and test datasets, resulting in a training dataset with 241 CPPs and 300 non-CPPs and test samples with 60 CPPs and 75 non-CPPs. These datasets are available in Appendix A. The division of the number of samples for training and independent test datasets by type of structure is summarized in Table 1.

Table 1 – Dataset division of peptide samples according to cell membrane permeability and file format.

	PDB		FASTA	
	Training	Independent Test	Training	Independent Test
CPPs	300	75	241	60
non-CPPs	300	75	300	75

Source: Author's own.

In summary, the flowchart of Figure 20 shows that all individual databases were collected and grouped, and their descriptors are extracted to be evaluated according to some filters, which encompass the limit for amino acids, removal of duplicates, and outliers. Finally, using random sampling, the final number of CPPs is equalized to the number of non-CPPs. This equalization balances the training and test dataset for the ML framework.

3.1.2 Database for B3PPs

For the problem of blood-brain barrier penetrating prediction, this thesis proposes using the Brainpeps database, a comprehensive data obtained from literature information about B3PPs, which includes primary structure, sequence, physicochemical properties, and information related to the experimental method applied in the validation of the penetration (VAN DORPE et al., 2012).

Unlike CPPs database, Brainpeps does not provide clear information about whether each peptide can cross or not the BBB, i.e., no labels are referencing the penetration of each peptide available in this database. Some experimental researches established classification for peptide permeability across the BBB using numerical limits for pharmacokinetic indicators.

Stalmans, Gevaert, et al. (2015) established five BBB penetration classes according to the level of influx rate constant (K_{in}) and BBB permeability (P). The indicator K_{in} represents the clearance of blood from the peptides after a single passage of the brain, expressed in mL/(g.min), while P indicates the permeability of a peptide acquired using a brain microvessel endothelial cell culture model (BMEC), expressed in cm/s (STALMANS; GEVAERT, et al., 2015; VAN DORPE et al., 2012). Stalmans and collaborators defined the five classes for in-

flux using a box and whisker plot, where classes 1 to 4 were determined by the 25, 50, and 75 percentiles as well as the lower and higher whiskers. BBB influx class 5 comprised the peptides with outlying BBB influx data. This method was used to classify the penetration into brain parenchyma for peptide as conotoxin cVc1.1 (POTH et al., 2021); somatropin, NOTA-conjugated somatropin, and gallium-labeled NOTA-conjugated somatropin (BRACKE et al., 2020); PapRIV (JANSSENS et al., 2021). The classes and limit values for each indicator are shown in the table available in Appendix B.

Another experimental parameter used to classify the penetration of peptides across the BBB is the endothelial effective permeability coefficient obtained with the parallel artificial membrane permeability assay (PAMPA) (P_e), expressed in 10^{-6} cm/s. Di et al. (2003) proposed the high penetration for molecules that has $P_e(10^{-6}) > 4.0$, low penetration for $P_e(10^{-6}) < 2.0$, and permeation uncertain $2.0 < P_e(10^{-6}) < 4.0$. These ranges were derived empirically from experiments to investigate high-throughput assay for BBB permeation prediction using porcine polar brain lipid. This method and its analogue on a logarithmic scale were used to classify the permeability of compounds in the brain, such as 3-hydroxy-2-pyridineal doxime compounds (ZORBAZ et al., 2018); furosemide, ranitidine, donepezil, and tacrine (ROSSI et al., 2021); platyphyllenone and alnusone (FELEGYI-TÓTH et al., 2022); gingerol and shogaol derivatives (SIMON et al., 2020).

As mentioned above, several works that experimentally investigated the penetration of peptides and other compounds into the blood-brain barrier also used the classification by the permeability level of the molecules based on established limits for K_{in} , P , or P_e . For example, Rossi and colleagues used the criterion of P_e to discover multi-target-directed ligands for Alzheimer's Disease (ROSSI et al., 2021). Based on the criteria and evidence highlighted in the works above, this thesis suggests a preliminary classification of the peptides of the Brainpeps database using the influx and permeability metrics.

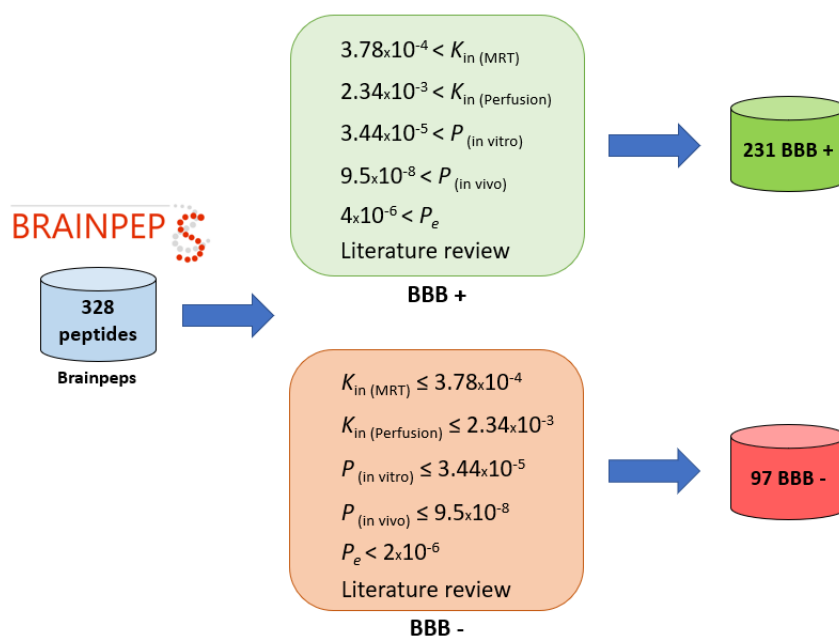
Regarding the classes established by Stalmans and collaborators using K_{in} and P values, it is proposed to classify the peptides of groups 1 and 2 (very low and low influx) as BBB-, while the other classes (medium, high, and very high influx) are defined as BBB+. Concerning the P_e metric, the same criterion created by Di and collaborators was used to classify the peptides, where high penetration compounds were labeled as BBB+, low penetration as BBB-, and uncertain penetration was classified according to the proximity to each threshold, i.e., $P_e(10^{-6}) > 3.0$ is BBB+ and $P_e(10^{-6}) < 3.0$ is BBB-. Figure 21 summarizes the process of previous classification according to the threshold values and type of experiment (MRT¹, perfusion, in vitro, or in vivo) for each parameter.

The implementation of the proposed criteria for classifying peptides based on their permeability in the BBB using pharmacokinetic indicators in this thesis yielded a database containing 231 BBB+ and 97 BBB- peptides. However, the database needed to be more balanced

¹Multiple-Time Regression method.

regarding 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.

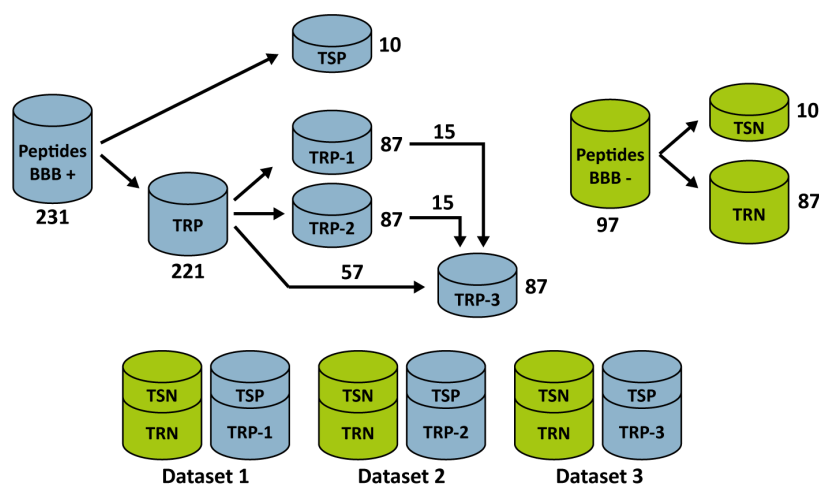
Figure 21 – Diagram of B3PPs and non-B3PPs datasets construction based on the employment of experimental pharmacokinetic indicators regarding the penetration of peptides into BBB to classify them into BBB+ or BBB-.



Source: Author's own.

To solve this problem of unbalanced datasets regarding the BBB+ and BBB- samples, three datasets were constructed with an equal number of 97 peptides in the BBB+ dataset and 97 samples in the BBB- dataset, as illustrated in the diagram of Figure 22. The dataset composed of 97 BBB- samples is the same for the three datasets (Dataset 1, Dataset 2, and Dataset 3) for training and testing the models, where ten samples were randomly selected to constitute the test samples of non-permeable molecules (TSN) and the 87 remaining samples are defined as the training set (TRN). Similarly, ten samples of BBB+ peptides were randomly selected to constitute the test samples of permeable molecules (TSP), and the remaining 221 peptides were divided to compose the training set of the three datasets. The training set of the Dataset 1 (TRP-1) is composed of 87 peptides randomly selected. The training set of Dataset 2 (TRP-2) contains 87 peptides selected randomly from the remaining samples. The training set of Dataset 3 (TRP-3) is composed of 87 peptides, of which 57 samples came from the remaining BBB+ samples, 15 are peptides randomly selected from TRP-1 and 15 were randomly selected from TRP-2.

Figure 22 – 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.



Source: Author's own.

In summary, 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 training and testing and that the models developed using these datasets were adequately trained and tested. Appendix C provides information regarding the peptides used in each dataset.

3.2 Molecular Descriptors

The molecular descriptors are the chemical features used to describe the peptides' structure, composition, and physicochemical properties. These features are used to train and test the ML models besides characterizing the chemical space of the molecules. The following sections describe the molecular properties proposed to predict CPPs and B3PPs.

3.2.1 Molecular descriptors for CPPs prediction

The properties selected to investigate the permeability of peptides across the cell membrane are divided into structural and physicochemical properties and sequence-based properties. The first group encompasses: MW, NRB, TPSA, Fsp³, LogP, number of aromatic rings (NAR), HBD, HBA, number of primary amino groups (NPA), number of guanidinium groups (NG), net charge (NetC), and number of negatively charged amino acids (NNCAA) at pH = 7.4. These descriptors are used because some are related to the oral bioavailability of compounds, some are approached in RO5, and others are associated with improving penetration in cell membrane (SANTOS; GANESAN; EMERY, 2016; SANDERS et al., 2011).

The second group encompasses the two AAC descriptors: fraction of arginine and lysine residues ($f(\text{Arg})$ and $f(\text{Lys})$). This group also encompasses the 40 DPC descriptors (dipeptides: RR, KK, KR, RQ, RK, WR, WK, NR, KW, WF, RS, FQ, RW, RI, QR, GR, RM, IW, RL, QN, ET, CN, PG, PL, GI, TV, FC, FG, GP, LS, SE, CV, GT, FL, CC, VC, GA, LG, GF, and GL) and 22 features from the PseAAC (CHOU, 2001). These sequence-based descriptors were selected because they were evaluated in previous studies and presented relevance in predicting cell membrane uptake (MANAVALAN et al., 2018; PANDEY et al., 2018; SANDERS et al., 2011). Section 2.4 of this thesis discusses in more detail the biochemical characteristics that led these physicochemical, structural, and sequence-based descriptors to predict CPPs.

The prediction of CPPs is analyzed according to feature composition FC to investigate which subgroup of features gives more discriminating information to classify the molecules correctly. Table 2 shows how all the descriptors were grouped into four different feature compositions, FC-1 to FC-4. FC-1 grouped only amino acid composition and sequence-based descriptors ($f(\text{Lys})$, $f(\text{Arg})$, 40 most correlated DPC, and PseAAC), which are important to evaluate the framework performance using only primary structure information of peptides. FC-2 comprises the twelve physicochemical and structure-based properties related to the oral bioavailability of compounds. FC-3 is the grouping of all analyzed descriptors, i.e., this feature composition combines FC-1 and FC-2. Finally, FC-4 encompasses the combination of the fraction of arginine and lysine, all the descriptors from PseAAC, the nine most well-correlated physicochemical and structure-based properties, and the ten most well-correlated DPCs. It is important to highlight that DPC descriptors in FC-1, FC-3, and FC-4 were selected using Kendall's correlation, as well as the nine physicochemical and structural descriptors in FC-4. The selection of 40 descriptors from 400 DPCs aimed to minimize the dimension of the original dataset, using only the dipeptides with more correlation to the labels. Regarding the selection applied in FC-4, the purpose was to construct a feature composition optimized concerning FC-3. The results of Kendall's correlation for the descriptors employed in CPPs prediction can be seen in Appendix D.

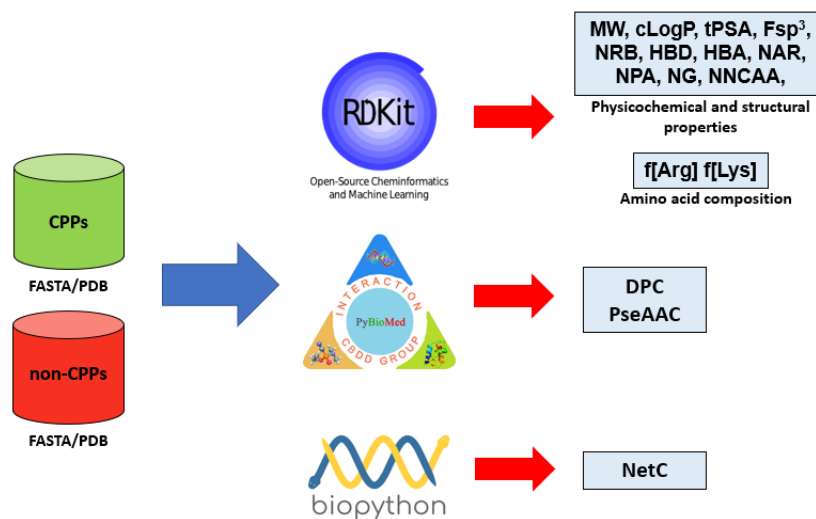
Table 2 – Feature composition for CPPs prediction analysis.

Feature Composition	Molecular Descriptors				Number of Descriptors
	Structural	AAC	DPC	PseAAC	
FC-1	-	$f[\text{Lys}]$, $f[\text{Arg}]$	40 DPCs	22 PseAACs	64
FC-2	MW, cLogP, tPSA, Fsp3, NRB, HBD, HBA, NAR, NPA, NG, NetC, NNCAA	-	-	-	12
FC-3	MW, cLogP, tPSA, Fsp3, NRB, HBD, HBA, NAR, NPA, NG, NetC, NNCAA	$f[\text{Lys}]$, $f[\text{Arg}]$	40 DPCs	22 PseAACs	76
FC-4	MW, cLogP, Fsp3, HBA, NAR, NPA, NG, NetC, NNCAA	$f[\text{Lys}]$, $f[\text{Arg}]$	10 DPCs	22 PseAACs	43

All the molecular descriptors displayed in Table 2 were extracted using Python packages. RDKit package (LOVRIĆ; MOLERO; KERN, 2019) was used to calculate the physicochemical and structure-based descriptors, besides the $f(\text{Lys})$ and $f(\text{Arg})$. PyBioMed package (DONG et al., 2018) was applied to extract the features from DPC and PseAAC. Net charge (NetC) of the peptides was calculated from structures using Biopython package (COCK et al.,

2009). All these properties were extracted from CPPs and non-CPPs libraries using both PDB and FASTA format. Figure 23 summarizes the feature extraction process explained here, relating the descriptors to their respective Python package used to calculate.

Figure 23 – Diagram of molecular descriptors extraction to compose the four FCs in cell membrane case study using RDKit, PyBioMed, and Biopython packages.



3.2.2 Molecular descriptors for B3PPs prediction

The molecular properties selected to investigate the permeability of peptides across the blood-brain barrier are divided into four FCs. The first feature composition (FC-1) comprised several key descriptors including molecular weight (MW), calculated water-octanol partition coefficient (cLogP), 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) (GELDENHUYS et al., 2015; MIKITSH; CHACKO, 2014; DICHARA et al., 2020). Some of these descriptors are also related to the oral bioavailability of compounds approached in RO5 (LOVERING, 2013; LIPINSKI et al., 2012; DOAK et al., 2014).

The second feature composition (FC-2) comprised 749 Mordred's molecular descriptors, which consist of a combination of structural and physicochemical descriptors. Mordred is a Python library for molecular descriptor calculations encompassing 2D, 3D, constitutional, and electronic descriptors (MORIWAKI et al., 2018). 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. The Table 3 summarizes the molecular descriptors by FC, and the complete list of Mordred's descriptors modules can be seen in Appendix E.

Table 3 – Feature composition for 3BPPs prediction analysis.

Feature Composition	Origin	Descriptors	Number of descriptors
FC-1	Dichiara's	MW, cLogP, LogD, tPSA, HBA, HBD, nO, nN, n(N+O)	9
FC-2	Mordred	2D, 3D, constitutional, and electronic descriptors	749
FC-3	Mordred selection by Kendall	JGI5, JGI6, JGI7, JGI9, EState-VSA5, GATS3d, nAcid, RotRatio, Lipinski, GhoseFilter	10
FC-4	FC-1 + FC-3	JGI5, JGI6, JGI7, JGI9, EState-VSA5, GATS3d, nAcid, RotRatio, Lipinski, GhoseFilter, MW, cLogP, LogD, tPSA, HBA, HBD, nO, nN, n(H+O)	19

Source: Author's own.

Unlike the cell membrane problem, the prediction of B3PPs proposed in this thesis does not use sequence-based descriptors because approximately 94.81% of the Brainpeps database used in this work comprises chemically modified peptides or the amino acid sequence is missing. These two factors make it infeasible to use sequence information as an input feature to a predictor.

All descriptors described in this section were calculated using Python packages Mordred, RDKit, and Instant JChem software. LogD descriptor was calculated using Instant JChem, while the remaining Dichiara's descriptors were calculated using RDKit. The peptide file format used in this analysis is MDL format. This format was selected firstly due to its capability to aggregate information about molecular conformation and chemical modifications, and secondly due to its availability in the Brainpeps repository.

3.3 Proposed framework to predict CPPs and B3PPs

The ML-based framework proposed in this thesis to predict CPPs and B3PPs is a generic pipeline that is also flexible regarding the choice of internal algorithms. This tool is composed of supervised Laplacian eigenmaps (sLE), a manifold dimensionality reduction algorithm used to reduce the high-d molecular descriptors dataset to three dimensions (3D) aiming as much to visualize the peptides in low-d as clustering the molecules according to their classes, CPP and non-CPP for cell membrane problem and BBB+ and BBB- for blood-brain barrier case.

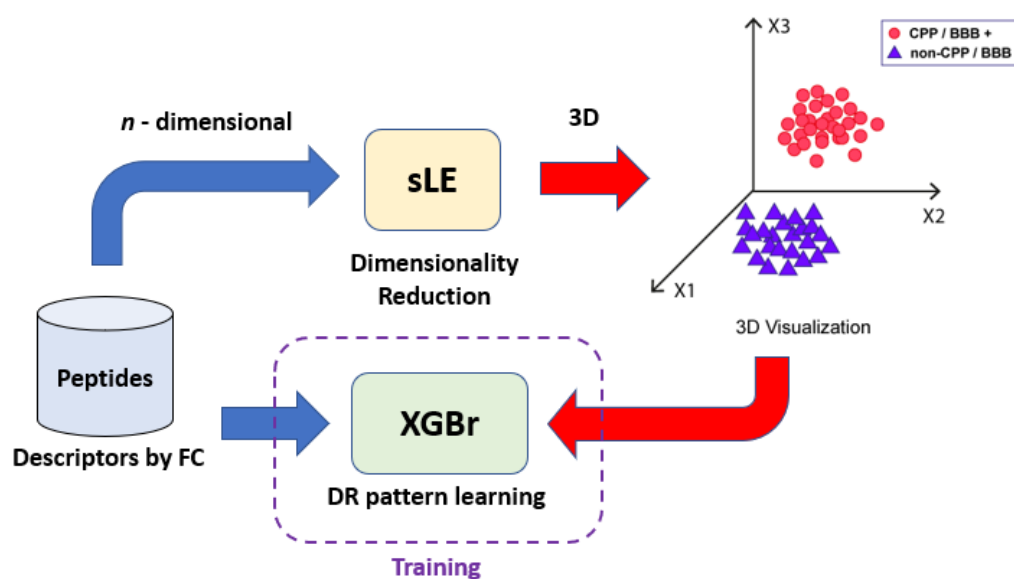
The sLE algorithm was chosen to compose the pipeline of the proposed framework due to its ability to perform dimensionality reduction in a supervised manner, adding information from the classes of molecules in the process of projecting dimensions into 3D space. In addition, sLE is capable of dealing with non-linear distribution data. Another decisive point for choosing this algorithm is that sLE is entirely based on deterministic equations and does not require an iterative training process, which could increase the computational cost of the framework.

However, the original sLE does not perform dimensionality reduction in an independent dataset, since this algorithm does not generate a model, which makes it infeasible to use this algorithm in the framework to predict the permeability of new peptides. Then, it is proposed

the use of a XGBoost regression (XGBr) algorithm to overcome this issue. Figure 24 illustrates the first stage of the proposed framework, where the n -dimensional dataset, based on molecular descriptors by FC, is reduced to 3D. The same n -dimensional data and its 3D reduction are respectively used as input and target in XGBr present in DR pattern learning stage. The XGBr model is trained and optimized by grid-search² to learn the reduction pattern generated by sLE. The DR pattern learning is an important step to overcoming the problem of reducing new peptides' data dimension. After this step, the next one is the training of a XGBoost classifier XGBc using the output from XGBr and the peptides' labels, as shown in panel A of Figure 25.

The final pipeline of the proposed framework to predict the penetration across the biomembranes is a pipeline with XGBr and XGBc, where the first algorithm learns the DR pattern from sLE and the second one is responsible for predicting if the molecule can penetrate or not the biomembrane, as shown in panel B of Figure 25. Furthermore, the framework can provide the 3D visualization of new data, which is essential for analyzing how distant the peptide is from its original cluster. Also, it is important to highlight that the proposed tool is trained and tested separately for cell membrane and BBB.

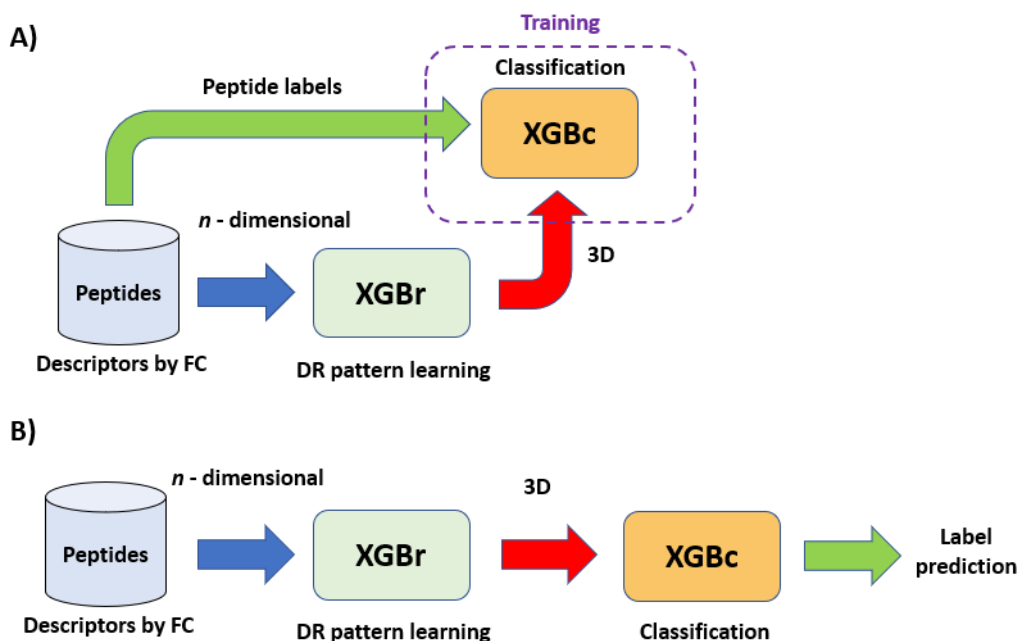
Figure 24 – Fluxogram of dimensionality reduction and pattern learning stages in the proposed framework. The DR stage represents the projection of n -dimensional data to 3D using sLE algorithm. The pattern learning stage encompasses the use of XGBr to learn and generalize the sLE projection for new data.



Source: Author's own.

²Grid-search is a method applied for optimization of hyperparameters using cross-validation over exhaustive search in a parameter grid.

Figure 25 – Proposed framework for CPPs and B3PPs prediction. Panel A) shows the stage of XGBc training with 3D data as input and peptides' labels as output. B) Illustrates the final pipeline of the proposed framework.



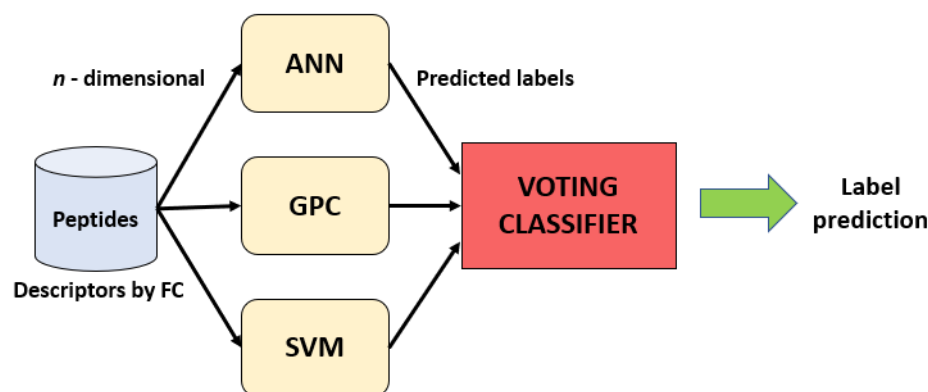
Source: Author's own.

The range of the searching hyperparameters adjusted for XGBr, XGBc, and sLE are shown in Appendix F.1. All the algorithms incorporated in the framework and their configuration processes were implemented using the scikit-learn and XGBoost package (CHEN; GUESTRIN, 2016) for Python language.

3.4 Voting classifier to predict CPPs and B3PPs

The voting classifier described in Section 2.5.6 is used in this thesis as a ML architecture to group classifiers to improve the prediction of CPPs and B3PPs and compare with the results achieved by the framework described in the previous section. The proposed Vcf groups ANN (MLP architecture), SVM, and *Gaussian* process classifier (GPC). Each classifier receives descriptors according to each FC depending on the biomembrane. Figure 26 illustrates the architecture of the voting classifier, where each ML model is trained separately using the grid-search to find the best combination of hyperparameters according to the best accuracy in 10-fold cross-validation by feature composition.

Figure 26 – Structure of voting classifier to predict peptides' biomembrane penetration.



Source: Author's own.

In summary, the proposed Vcf to predict CPPs and B3PPs is composed of three single classifiers optimized and trained by FC. Like the framework, Vcf is trained and tested separately for the cell membrane and BBB. The range of the searching parameters adjusted for each ML-based algorithm is shown in Appendix F.2. All the algorithms incorporated in the framework and their configuration processes were implemented using the scikit-learn package for Python language.

It is essential to highlight that the BChemRF-CPPred, a free-to-use web server created by the author of this thesis and collaborators to predict specifically CPPs is based on the architecture of voting classifier presented above Oliveira et al. (2021). More information about the BChemRF-CPPred web server is provided in Appendix G.

3.5 Conclusion

This chapter presented the main aspects related to the proposed methods, approaching the database used in both problems, prediction of CPPs and B3PPs, as the molecular descriptors used as features to perform the classification. The chapter also approaches the general aspects of the proposed framework, describing how the ML algorithms are linked inside the developed pipeline, and describes the voting classifier used to predict peptides' permeability through biomembranes. The next chapter describes and discusses the results of the frameworks and Vcf in each problem.

4 RESULTS AND DISCUSSIONS

This chapter presents the results achieved by the proposed methods to predict CPPs and B3PPs in this thesis. The first section focuses on describing the performance of the developed framework to predict peptides' permeability across the cell membrane and comparing it with a voting classifier and its classifiers using 10-fold cross-validation analysis and independent test. The second section approaches the same analysis for predicting peptides' permeability across the blood-brain barrier. The final section summarizes the results presented in this chapter.

4.1 Prediction of CPPs

This section presents the cross-validation and independent test analysis for the proposed framework with DR algorithm in CPPs prediction. The performance of this tool is compared with the classifiers ANN, SVM, GPC, and with a voting classifier joining these three models, as mentioned in the chapter on methods. The voting classifier to predict CPPs is referenced in this thesis as Vcf-CPP (*voting classifier for CPP prediction*) and the proposed framework is named as DPF-CPPred (*dimensionality reduction and pattern learning framework for CPP prediction*).

In this thesis, the prediction of CPPs is analyzed using both PDB and FASTA format according to sample distribution showed in Table 1 in the previous chapter. The use of PDB aims to evaluate how the models perform in the prediction of permeability using peptides with chemical modifications and non-natural amino acids, while FASTA focuses on investigating the prediction capacity of the techniques using only natural peptides.

Regarding molecular descriptors, the prediction of CPPs in this thesis investigated two class of molecular descriptors: (1) the structure-based descriptors that include structural and physicochemical properties related to the permeation of molecules into the biological membranes which are obtained from the molecular structures of peptides—MW, TPSA, Fsp³, LogP, HBA, HBD, NAR, NRB, and NetC—, as well as, some properties related to the polar charged groups—NPA, NG, NNCAA—that could influence in their permeability; and, (2) sequence-based descriptors, i.e., information calculated from the primary structure of the peptide—AAC, PseAAC, and DPC. Regarding the sequence-based descriptors, the amino acid compositions f[Arg] and f[Lys] were applied in the framework. The sequence-based descriptors selected have relation with the charge, hydrophobicity, and hydrophilicity of the peptides, which are important properties related to interaction between these molecules and the cell membrane.

All the selected descriptors were divided into four FCs as mentioned in Section 3.2.1 for both PDB and FASTA. This division allows us to evaluate how different groups of descriptors impact the performance of the algorithms. This segmented analysis is essential because

physicochemical and structural descriptors are directly correlated with the oral bioavailability of compounds. At the same time, those based on the sequence do not have a direct correlation with this property. On the other hand, it is much easier to filter or develop peptides based on sequence filters than to meet physicochemical and structural criteria.

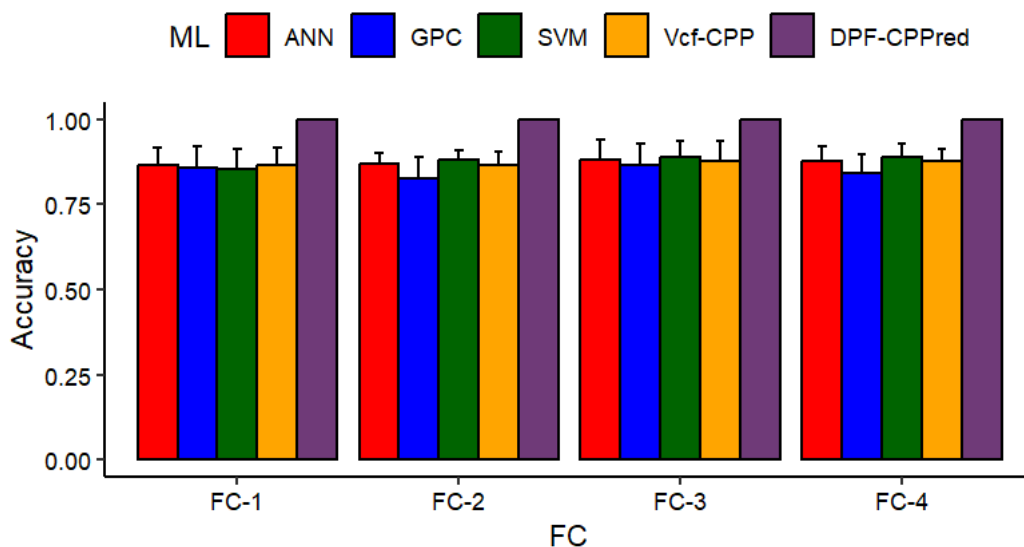
4.1.1 Cross-validation analysis in CPPs prediction

The 10-fold cross-validation is applied to evaluate the generalization of the proposed framework and other ML models using the training dataset with both data formats. The obtained results for performance of the DPF-CPPred, Vcf-CPP, and its algorithms an ANN, GPC, and SVM using the PDB dataset are shown in Figure 27 and for FASTA is presented in Figure 29. The used hyper-parameters of all the algorithms by FC and for both file encoding are shown in Appendix H.

4.1.1.1 Cross-validation analysis for PDB encoding

The performance of each estimator shown in Figure 27 indicates that the DPF-CPPred achieved an average accuracy of 100% in cross-validation for all FCs using PDB, highlighting how different the performance obtained by the framework is from the Vcf-CPP and its algorithms, which achieved the best result by SVM in FC-3 with 89% of accuracy.

Figure 27 – Barplot of accuracy from 10-fold cross-validation of DPF-CPPred (purple), Vcf-CPP (orange), ANN (red), GPC (blue), and SVM (green) using PDB format.



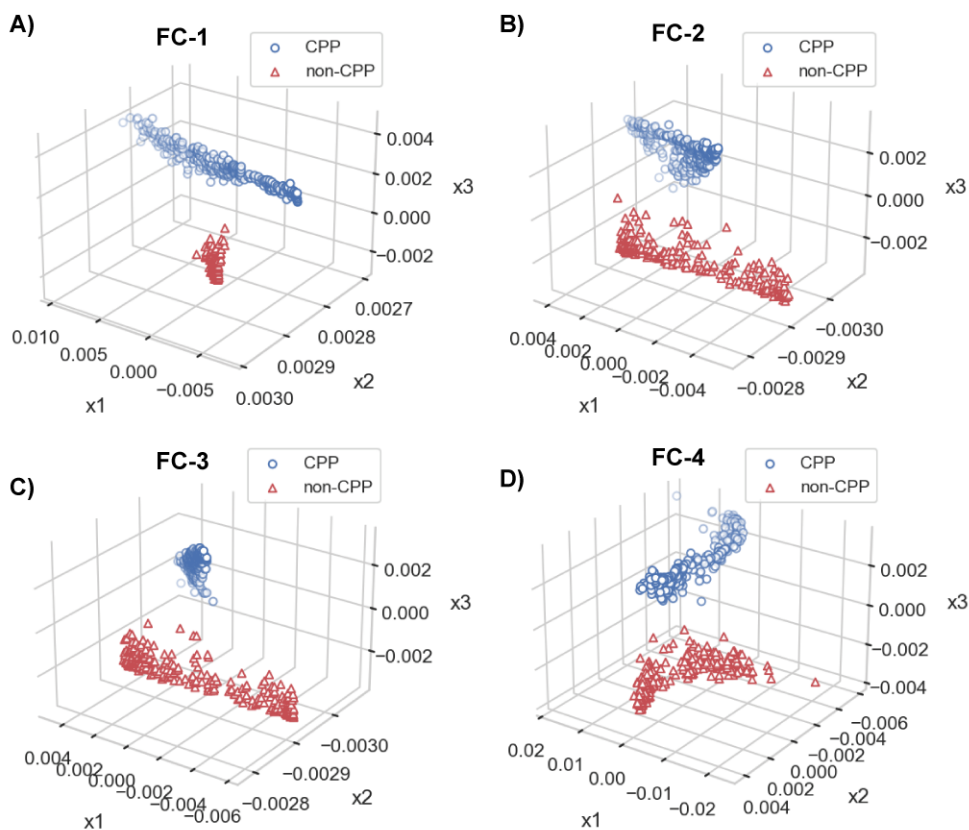
Source: Author's own.

These results indicate that the XGBoost model trained in the pattern learning stage of the framework reached good performance in generalizing the dimensionality reduction of the training dataset. This performance can be explained by level of sample clustering provided by sLE, which can accurately group the CPPs samples and differentiate them from non-CPPs,

as shown in panels on Figure 28. As can be seen, the 3D reduced training dataset exhibits significant differentiation between the two classes for the four FCs.

It is important to note that, for the four FCs, the framework segregated the training dataset of peptides very well between those that can cross the cell membrane and those that can not, using the sLE algorithm. Although encouraging, these results can not provide significant information about which group of molecular descriptors offers more information to correctly differentiate CPPs from non-CPPs, i.e., based only on cross-validation results, the DPF-CPPred can not indicate if physicochemical and structure-based descriptors provide more correlated to prediction than sequence-based features, or if the specific FC is more informative than other because all the feature compositions reached the maximum performance. This differentiation among the descriptors could contribute to designing new peptides capable of crossing the cell membrane.

Figure 28 – The 3D plot of the reduced PDB training dataset by the DPF-CPPred in each FC.



Source: Author's own.

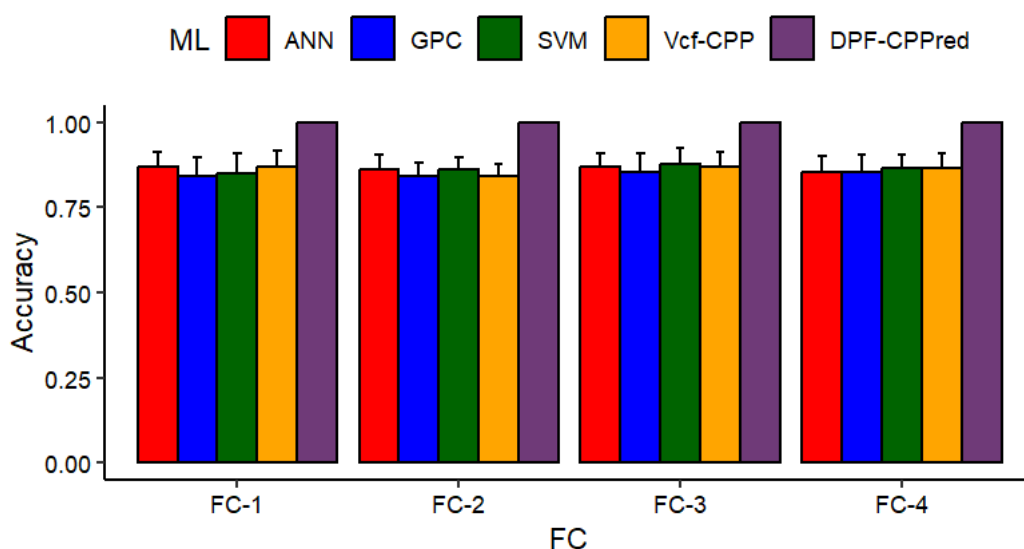
Regarding Vcf-CPP, the results obtained with FC-3 had reached an average accuracy of 0.17% greater than FC-4. However, the Kruskal–Wallis H test (p -value = 0.820) showed no statistically significant difference between the accuracies obtained by these two FCs. Furthermore, the voting classifier using the FC-4 (43 descriptors) is less complex than those that use FC-3 (76 descriptors). It is also important to highlight that although the descriptors from FC-1 (containing only sequence-based descriptors) and from FC-2 (only physicochemical and

structure-based descriptors) have shown relevant correlation to CPPs' prediction according to Kendall's correlation analysis, these descriptors isolated do not provide enough information to predict the permeability of these peptides satisfactorily into the cell membranes, as shown by the Vcf-CPP performance in Figure 27. These results are important not only to highlight the differences among the feature compositions in the prediction of CPPs but also indicate that the optimized combination of physicochemical, structure- and sequence-based descriptors (FC-4) better predict natural and synthetic peptides than other analyzed FCs.

4.1.1.2 Cross-validation analysis for FASTA encoding

Different from PDB format, the dataset composed only of FASTA files is easier to obtain and manipulate. However, until the writing of this thesis, there is no reliable representation of chemically modified peptides using the FASTA format, i.e., this encoding can only represent natural peptides. The performance of the proposed framework and other tools in natural CPPs prediction is evaluated separately here for this file format. Figure 29 shows the result of 10-fold cross-validation analysis in this case study.

Figure 29 – Barplot of accuracy from 10-fold cross-validation of DPF-CPPred (purple), Vcf-CPP (orange), ANN (red), GPC (blue), and SVM (green) using FASTA format.



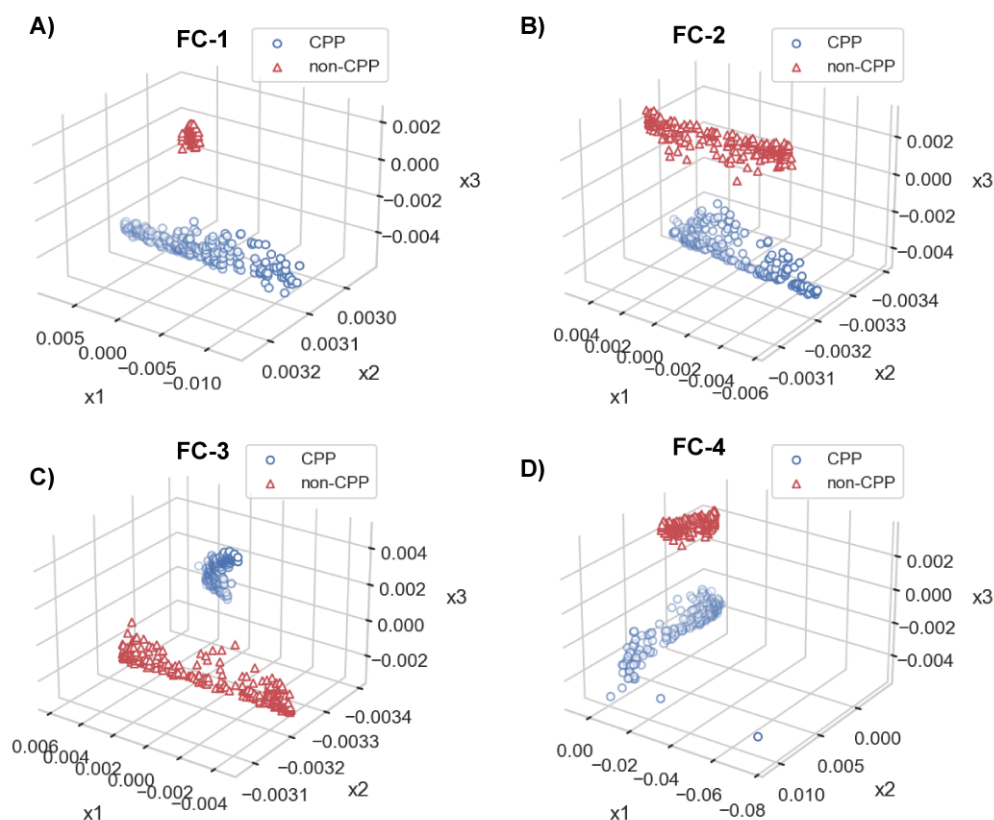
Source: Author's own.

Like the previous experiment, the DPF-CPPred achieved 100% of average accuracy in cross-validation analysis in predicting only natural peptides that can cross the cell membrane. Figure 30 shows the result of 3D dimensionality reduction by the framework for each FC, where it is possible to see that the proposed tool can segregate the two classes of molecules. Regarding Vcf-CPP performance, the model that used the FC-3 reached the best performance with an average accuracy of 86.9%, while FC-1, FC-2, and FC-4 achieved values between 84.13 and 86.71%, respectively. However, the Kruskal–Wallis H test (p -value = 0.675) again showed

no statistically significant difference between the performance of FC-3 and FC-4, evidencing the contribution of the optimized selection of physicochemical, structural, and sequence-based features to predict the uptake of natural peptides by the cell membrane.

In a general analysis, all these results display, mainly for Vcf-CPP and its classifiers, the difference between the use of PDB and FASTA format. The first one aggregates more chemical information, and impacts in CPPs prediction as shown in average accuracy between the two experiments. Another notorious difference is in feature composition, where the joining of sequence- and structural-based descriptors (FC-3 and FC-4) provided more information to classify peptides correctly in both experiments. It reveals the importance of combining molecular descriptors related to oral bioavailability and the composition and arrangement of amino acids to predict the uptake of peptides by the cell membrane.

Figure 30 – The 3D plot of the reduced FASTA training dataset by the DPF-CPPred in each FC.



Source: Author's own.

The proposed framework performed very well in cross-validation analysis, reaching 100% in average accuracy for all FCs, since it uses the ability of sLE to reduce the original dimension of the dataset and cluster the samples into each class, outperforming the other compared ML tools. However, as well as in the PDB experiment, it is impossible to confirm with cross-validation results which FC impacts the prediction of peptides' permeability across the cell membrane using the proposed framework. Then, an independent test will evaluate the im-

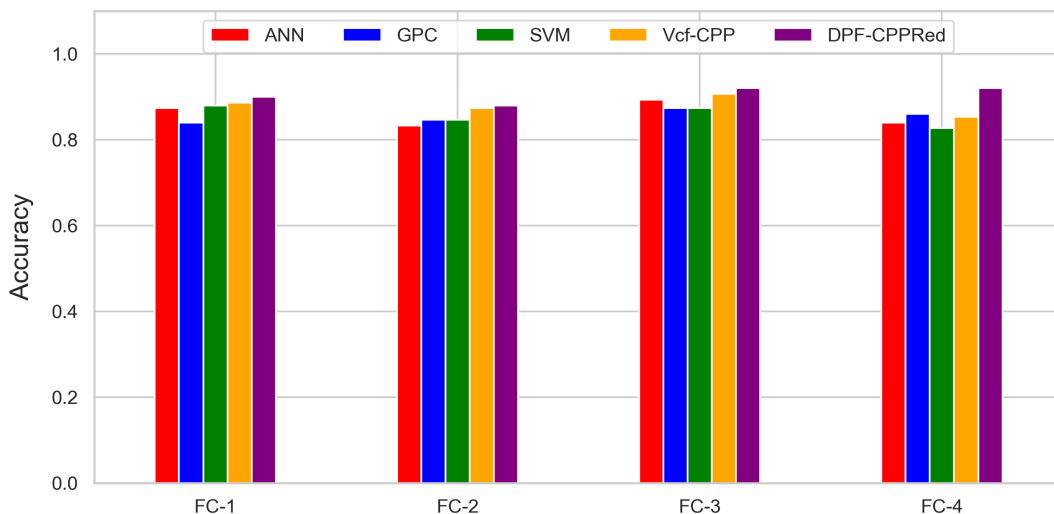
pact of each feature composition in correctly classifying CPPs using the DPF-CPPred.

4.1.2 Independent test analysis in CPPs prediction

This section approaches the independent test performed with the same ML tools described in the previous section and with the independent dataset, as described in Table 1. Here will be evaluated the performance of each model in predicting CPPs for both file formats (PDB and FASTA).

The analyses revealed that the DPF-CPPred based on feature compositions with more information (FC-3 and FC-4) and using PDB data, obtained an accuracy equal to 92%. At the same time, for FC-1 and FC-2, the framework exhibited an accuracy of 90% and 88%, respectively, as shown in Figure 31. These results indicate that the DPF-CPPred outperformed the Vcf-CPP and its models for almost all metrics in this test since the best result achieved by voting classifier was 90.6% in accuracy for FC-4 and the other three classifiers demonstrated performance of less than 90%, as shown in Table 4.

Figure 31 – Accuracy of ANN (red), GPC (blue), SVM (green), Vcf-CPP (orange), and DPF-CPPred (purple) by FCs evaluated in the independent test for PDB data format.



Source: Author's own.

The results presented in Figure 31 also highlight some interesting points regarding feature composition and framework performance. Although the cross-validation analysis (see the Section 4.1.1) revealed that the DPF-CPPred obtained 100% of average accuracy for all FCs, the independent test proved that the molecular descriptors based on the fusion of physicochemical, structure- and sequence-based properties contribute more to correctly classifying the peptides than using the same descriptors isolated (FC-1 and FC-2). These findings corroborate with previous works that associate the permeability of peptides across the cell membrane with the

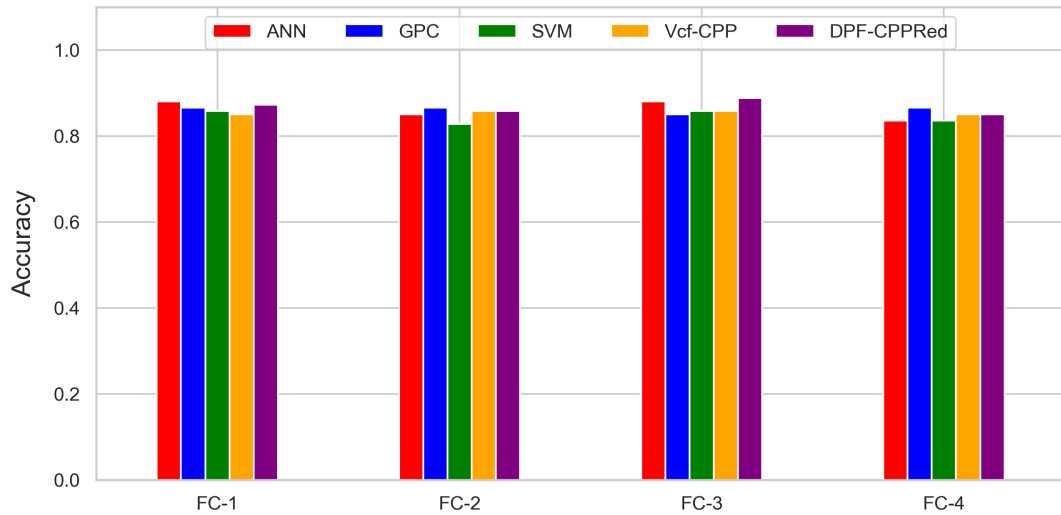
presence of a fraction of arginine and lysine (sequence-based descriptors) since these residues provide a positive charge to a peptide that contributes to the intersection between the peptide and the outer surface of the cell membrane, which is generally negatively charged due to the presence of phosphate groups in phospholipids. This interaction can facilitate the binding of peptides to the membrane and promote their internalization (SU et al., 2009; COPOLOVICI et al., 2014). These results also corroborate with the contribution of PseAAC that correlates a group of residues with hydrophobicity and hydrophilicity of peptides, important properties in molecular interaction between peptide and cell membrane. Furthermore, the results also highlight the contribution of descriptors associated with bioavailability of oral drugs, when combined with sequence-based descriptors, increase the capacity of prediction of CPPs by the DPF-CPPred, indicating that these descriptors could be used as a molecular filter for discovery and development of new potential CPPs.

Table 4 – Comparison of accuracy, sensitivity, specificity, F1-score, and MCC obtained for ANN, GPC, SVM, Vcf-CPP, and DPF-CPPred in the independent test using FC-4 and PDB format.

Method	Sensitivity	Specificity	Accuracy	F1-score	MCC	AUC
ANN	0.880	0.906	0.893	0.891	0.786	0.950
GPC	0.853	0.893	0.873	0.870	0.747	0.934
SVM	0.853	0.893	0.873	0.870	0.747	0.943
Vcf-CPP	0.893	0.920	0.906	0.905	0.813	0.953
DPF-CPPred	0.920	0.920	0.920	0.920	0.840	0.920

Regarding the prediction of cell-penetrating peptides using FASTA format, the DPF-CPPred achieved different performances compared to the model trained with peptides in PDB format. Figure 32 summarizes the performance of the ML tools evaluated in this test. The best performance was obtained by DPF-CPPred using FC-3, which reached an accuracy of 88.8% and F1-score of 87.2%. While for FC-1 DPF-CPPred obtained 87.3% and 86%, for FC-2 85.8% and 84%, and for FC-4 achieved 85.1% and 83.3% for accuracy and F1-score, respectively. The Vcf-CPP obtained the best result using the FC-2 with 85.07% of accuracy and 86.2% of F1-score, while the other feature compositions achieved inferior performance. These results indicate that the proposed framework trained with only natural peptides did not achieve the best performance for all FCs when compared with the other ML tools trained with the same data. Furthermore, the results presented in table 5 also indicated that this version of DPF-CPPred also did not outperform the version trained with chemically modified peptides.

Figure 32 – Accuracy of ANN (red), GPC (blue), SVM (green), Vcf-CPP (orange), and DPF-CPPred (purple) by FCs evaluated in the independent test for FASTA data format.



Source: Author's own.

Table 5 – Comparison of the performance of DPF-CPPred frameworks that used only natural peptides in the independent test. The comparison was performed between the frameworks based on the four feature compositions (FC-1 to FC-4) that use FASTA as input with the framework based on the FC-4 that uses the PDB as input.

Format	FC	Sensitivity	Specificity	Accuracy	F1-score	MCC
FASTA	FC-1	0.881	0.867	0.873	0.860	0.745
FASTA	FC-2	0.847	0.867	0.858	0.840	0.731
FASTA	FC-3	0.864	0.907	0.888	0.872	0.773
FASTA	FC-4	0.847	0.853	0.851	0.833	0.699
PDB	FC-4	0.920	0.920	0.920	0.920	0.840

The proposed framework was also compared with some available state-of-the-art web servers in CPPs prediction. This comparison was divided into two experiments according to the nature of peptides using the same independent dataset described in Table 1. The choice of the models was based on the FC that provides the best performance in terms of accuracy and F1-score in the independent test showed previously for PDB and FASTA formats.

The first experiment corresponds to ML tools that process only peptides in FASTA format, i.e., they were trained with only natural peptides. This experiment compares the proposed DPF-CPPred based on FC-3 with Vcf-CPP using descriptors from FC-2 and some other ML-based tools previously mentioned in Section 1.3.1: the ML-based predictor of CPPs (MLCPP) (MANAVALAN et al., 2018), the CPP predictor based on RF (CPPred-RF) (WEI; XING, et al., 2017), and the adaptive k-skip feature CPP predictor (SkipCPP-Pred) (WEI; TANG; ZOU, 2017). The second experiment compares the proposed framework and the voting classifier, using

FC-4, with the kernel extreme learning machine based CPP prediction model (Kelm-CPPpred) (PANDEY et al., 2018). The models used in the second experiment were trained with natural and non-natural peptides. Table 6 shows the results obtained in the first and second experiments.

The first experiment reveals that DPF-CPPred achieved the best performance in the prediction of natural CPPs, outperforming the other state-of-the-art tools with 88.8% accuracy and 87.2% of F1-score, using a combination of molecular features provided by FC-3. The proposed model also obtained the best specificity (90,7%), indicating its significative capacity to indicate the non-CPPs correctly. The Vcf-CPP based on FC-2 also obtained superior results compared to CPPred-RF and SkipCPP-Pred in the prediction of natural peptides that can cross the cell membrane, achieving accuracy and F1-score of 85% and 86,2%, respectively.

Table 6 – Comparison of the performance of previous ML-based tools (MLCPP, CPPred-RF, and SkipCPP-Pred), FC-2 based Vcf-CPP, and FC-3 based DPF-CPPred using only natural peptides from the independent dataset (1st experiment); as well as, the evaluation of the performance of Kelm-CPPpred and FC-4 based DPF-CPPred and Vcf-CPP from all independent dataset (2nd experiment).

Method	Sensitivity	Specificity	Accuracy	F1-score	MCC
First Experiment					
MLCPP	0.966	0.786	0.866	0.865	0.752
CPPred-RF	0.983	0.453	0.688	0.737	0.495
SkipCPP-Pred	0.966	0.520	0.625	0.753	0.525
Vcf-CPP (FC-2)	0.847	0.853	0.850	0.862	0.698
DPF-CPPred (FC-3)	0.864	0.907	0.888	0.872	0.773
Second Experiment					
Kelm-CPPpred	0.906	0.866	0.886	0.888	0.773
Vcf-CPP (FC-4)	0.893	0.920	0.906	0.905	0.813
DPF-CPPred (FC-4)	0.920	0.920	0.920	0.920	0.840

Concerning the second experiment, the DPF-CPPred and the Vcf-CPP, both using FC-4, also overcome the performance of Kelm-CPPpred in the prediction of chemically modified CPPs. The proposed framework reached 92% of accuracy against 88.6% achieved by Kelm-CPPpred. These results show that, for these two scenarios, the DPF-CPPred performed better using a combination of all groups of molecular descriptors.

A general analysis of independent test results shows that the framework architecture developed in this thesis involving DR with sLE and pattern learning performed better in the penetration prediction of peptides across the cell membrane when compared with other ML tools in almost all simulations. These results also corroborated with the hypothesis that the combination of physicochemical, structure-, and sequence-based descriptors can increase the prediction of the proposed framework, overcoming some state-of-the-art web servers dedicated to predicting CPPs.

4.2 Prediction of B3PPs

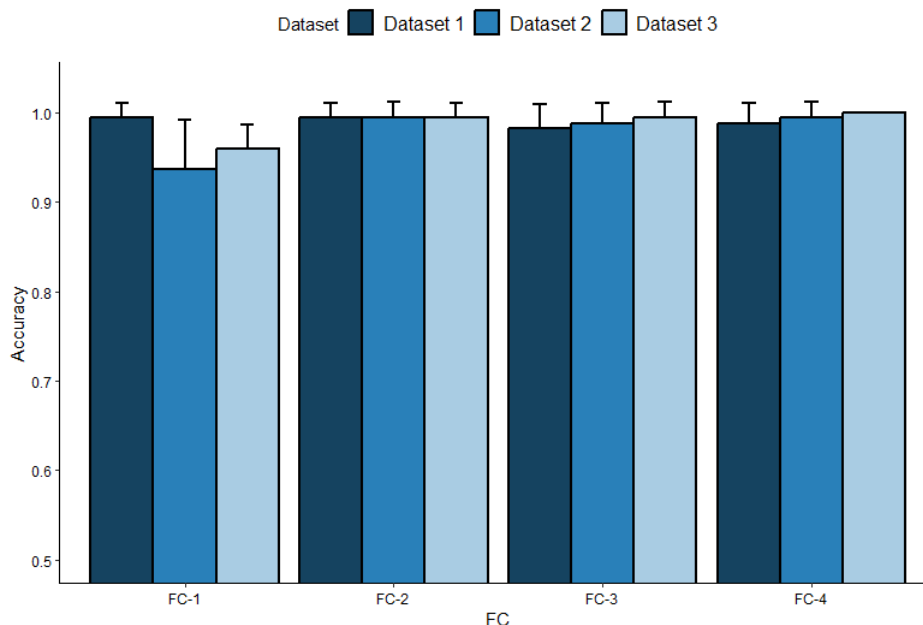
This section presents the performance analysis for the proposed framework DPF-3BPPred (*dimensionality reduction and pattern learning framework for 3BPP prediction*) to predict the penetration of peptides across the blood-brain barrier. The performance in this case study is evaluated according to 10-fold cross-validation and in an independent test. The proposed tool is also compared with Vcf-3BPP (*voting classifier for 3BPP prediction*) and its classifiers (ANN, SVM, and GPC).

In this thesis, the file format of peptides used in this analysis is MDL, since this is the original format provided by Brainpeps, whose structure can encode as much natural as chemically modified peptides. The three datasets used in cross-validation analysis are balanced and encompass 97 peptides that can cross the blood-brain barrier (BBB +) and 97 that can not (BBB -). The descriptors used in this case study are divided into four feature compositions. The FC-1 is based on 2D molecular properties provided by the Mordred package. FC-2 comprises the five best-correlated Mordred descriptors concerning peptides' labels according to Kendall's correlation (see Appendix I). FC-3 encompasses the descriptors approached in Dichiara et al. (2020), and FC-4 combines the descriptors from FC-2 and FC-3, as shown in Table 3 (see the Section 3.2.2). The division of these FCs is based on the concept of using a range of physicochemical and structural molecular properties not investigated to correlation with BBB permeability yet, where FC-1 comprises all these descriptors and FC-2 represents an optimized selection of these features. The descriptors with proven association with blood-brain barrier penetration (FC-3) are compared with the other unusual properties. FC-4 is used to evaluate the performance of the models when trained with the combination of molecular descriptors.

4.2.1 Cross-validation analysis in B3PPs prediction

This section approaches the predictive capacity of DPF-3BPPred according to 10-fold cross-validation. This metric was applied to the training portion of the three dataset samples and the four FCs. Seventy-two 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 33 – Barplot of accuracy achieved by DPF-3BPPred in 10-fold cross-validation analysis for all FCs and datasets.



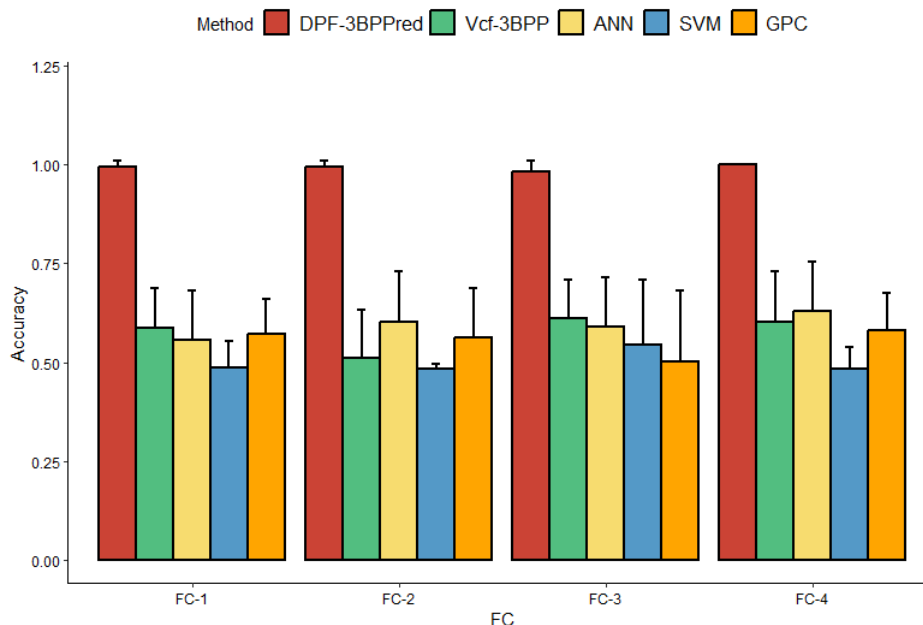
Source: Author's own.

The results presented in Figure 33 demonstrate the contribution of each group of descriptors in predicting B3PPs using the proposed ML-based framework. DPF-3BPPred achieved values greater than 93% of average accuracy for all FCs. The FC-1 model 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. The FC-3 model obtained values between 97.68% and 98.86%, whereas the one based on FC-4 merged both FC-1 and FC-3 descriptors, obtaining accuracy values ranging from 98.8% to 100%.

In order to evaluate the difference in predictive performance of DPF-3BPPred using molecular properties related to bioavailability (FC-2) and the fusion with selected descriptors from Mordred (FC-4), an ANOVA test was applied to the accuracy values obtained for each fold of the 10-fold cross-validation performed on the three datasets. 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 DPF-3BPPred trained with FC-2 requires the calculation of 749 descriptors, whereas the framework based on FC-4 requires only 19.

The cross-validation analysis was also applied to compare the performance of the best DPF-3BPPred model with the ANN, SVM, and GPC algorithms, in addition to the voting classifier composed of these last three models (Vcf-3BPP). Figure 34 shows the performance of the best models by FC and the results reveal that the performance of the proposed framework surpassed the other techniques for all FCs.

Figure 34 – Barplot of accuracy achieved by the best model of each method in 10-fold cross-validation analysis among all FCs and datasets.



Source: Author's own.

The results of Figure 34 reveal that for FC-1, the ANN, SVM, GPC, and Vcf-3BPP models achieved average accuracy values of 55.7%, 48.8%, 57%, and 58.7%, respectively. For the FC-2 descriptors, the same models obtained average accuracy values of 60.33%, 48.24%, 56.33%, and 51.11%, respectively. The results for the selected Mordred descriptors (FC-3) show that the ANN, SVM, GPC, and Vcf-3BPP models obtained average values accuracy of 59.12%, 54.38%, 50.26%, and 61.01%, respectively. Comparatively, the same models based on FC-4 descriptors achieved average accuracy values of 62.94%, 48.2%, 58.2%, and 60.16%. The used hyper-parameters of all the algorithms by FC are shown in Appendix J.

The cross-validation analysis highlights some characteristics of the B3PPs prediction with the proposed framework. From a computational perspective, DPF-3BPPred surpassed the performance of the other evaluated models for all groups of molecular descriptors, which can be explained by the capacity of the sLE to reduce and cluster the peptides in their respective class (BBB+ or BBB-). Another important aspect revealed by this analysis is the contribution of each FC. Based on the results shown in Figure 33, the physicochemical and structural descriptors extracted using Mordred package (FC-2) provided more information in predict B3PPs when compared with the descriptors approached in Dichiarà's work and that are associated with bioavailability of the molecules and with the BBB penetration (FC-1).

4.2.2 Independent test analysis in B3PPs prediction

This session covers the independent test performed by DPF-3BPPred and other ML models in predicting B3PPs. Also presented in this section is a performance comparison of the proposed framework with other tools developed and made available on web servers. The results presented here originated from the best model by FC according to the average accuracy obtained in cross-validation analysis.

Table 7 shows the outcomes obtained in independent tests for all algorithms evaluated. Regarding the DPF-3BPPred, the accuracy outcomes obtained for each feature composition indicate that the feature distribution between the training and test data in the three datasets, which were constructed using random sampling, may have differed. 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.

Regarding other metrics, the best DPF-3BPPred based on FC-3 also yielded high F1-score and Matthew's correlation coefficient (MCC) values, along with the maximum recall value for one of the three datasets. The area under the receiver operating characteristic curve (AUC) values between 0.74 and 0.84 also indicate that the proposed framework can distinguish between the two classes of peptides (BBB+ and BBB-) with better performance than the other analyzed ML tools, which obtained values between 0.45 and 0.81. These results suggest that the DPF-3BPPred can accurately predict which peptides can penetrate the BBB among all relevant instances using as much of the selected molecular descriptors grouped in FC-3 as the properties included in FC-4, which maintained similar performance. Appendix K provides the metric values obtained by DPF-3BPPred and their respective gamma values using the three datasets in the 10-fold cross-validation and independent tests, respectively.

Table 7 – Independent test analysis for the best DPF-3BPPred, ANN, SVM, GPC, and Vcf-3BPP models by each FC.

FC-1	Method	Accuracy	F1-score	MCC	Precision	Recall	AUC
	DPF-3BPPred	0.75	0.70	0.52	0.85	0.75	0.74
	Vcf-3BPP	0.60	0.60	0.20	0.60	0.60	0.65
	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
	GPC	0.50	0.16	0	0.50	0.10	0.50
FC-2							
	DPF-3BPPred	0.80	0.81	0.61	0.75	0.80	0.75
	Vcf-3BPP	0.55	0.57	0.10	0.54	0.60	0.47
	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
	GPC	0.55	0.57	0.10	0.54	0.60	0.47
FC-3							
	DPF-3BPPred	0.90	0.90	0.81	0.83	0.90	0.75
	Vcf-3BPP	0.70	0.62	0.43	0.83	0.50	0.74
	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
	GPC	0.70	0.62	0.43	0.83	0.50	0.75
FC-4							
	DPF-3BPPred	0.85	0.84	0.70	0.88	0.85	0.84
	Vcf-3BPP	0.70	0.66	0.40	0.75	0.60	0.81
	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
	GPC	0.50	0.16	0	0.50	0.10	0.50

The results of the independent test presented in Table 7 also demonstrate that, among the analyzed ML classifiers, the Vcf-3BPP achieved its best performance using the descriptors from FC-4, obtaining an accuracy, F1-score, and AUC of 70%, 66%, and 0.81, respectively. The ANN model achieved its highest accuracy for FC-3 (80%); however, it did not surpass the performance of the proposed ML-based framework for the same FC (90%). SVM also obtained its best results with the descriptors from FC-3, achieving an accuracy, F1-score, and AUC of 70%, 66%, and 0.75, respectively. Comparatively, GPC also reached its highest performance using FC-3, with values of 70%, 62%, and 0.75 for the same metrics. These results indicated that, among the analyzed ML classifiers, the ANN outperformed the Vcf-3BPP, but did not surpass the results of DPF-3BPPred.

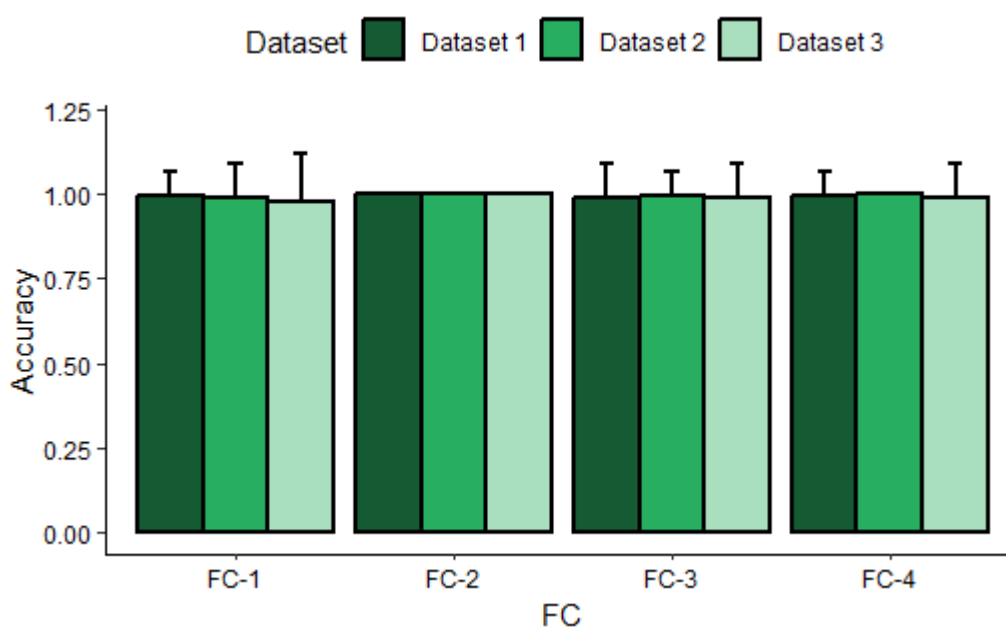
The findings from the two analyses demonstrate the efficacy of the proposed ML-based framework in accurately predicting B3PPs, with cross-validation accuracy values exceeding 90% and values between 75% and 90% for the external validation set. Furthermore, this study highlights the contribution of the 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 a 5% reduction in the accuracy of each model.

4.2.3 Leave-one-out cross-validation analysis

This study also employed leave-one-out cross-validation (LOOCV) as a complementary analysis to evaluate the proposed framework to determine the optimal DPF-3BPPred configuration. This analysis uses the three complete datasets (consisting of the training and testing subsets) for each FC. Figure 35 presents the average accuracy obtained by the framework for each dataset and FC.

Figure 35 – LOOCV analysis employed in DPF-3BPPred.



Source: Author's own.

The results demonstrated that FC-2 enabled the DPF-3BPPred 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. Comparing 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 accuracy values obtained by DPF-3BPPred using the three datasets in LOOCV analysis are shown in Appendix L.

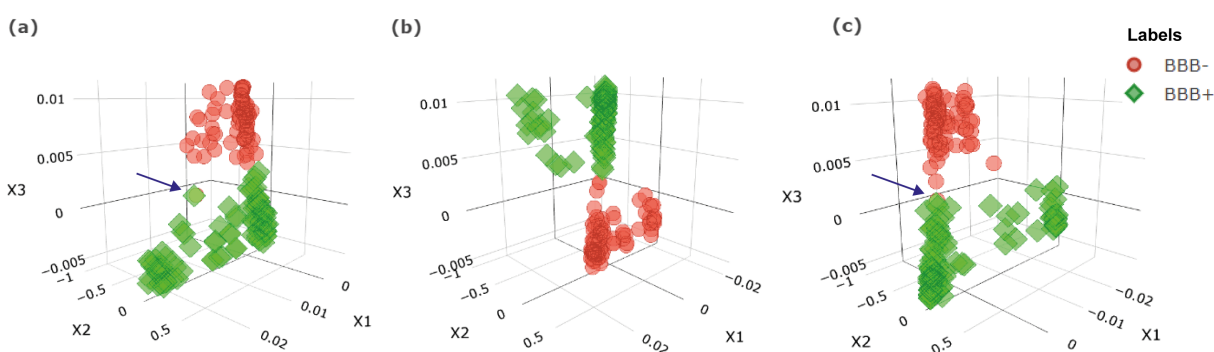
The outcomes from the LOOCV of the datasets belonging to FC-2 and FC-4 were analyzed through a pairwise comparison using an ANOVA test. The results indicate no statistically significant difference between the means of Datasets 1 and 3, with p -values of 0.318 and 0.157, respectively. Upon comparing the 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, DPF-3BPPred, 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 classify the three molecules correctly 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 performance analyses employed in the proposed 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 36 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.

The dimensionality reduction using sLE revealed that Dataset 1 exhibits an overlap between two peptides belonging to different classes (see the blue arrow in Figure 36a), whereas Dataset 3 displayed an overlap between at least three peptides from distinct classes (see the blue arrow in Figure 36c). This pattern is consistent with the results shown in Figure 35. 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 36 – Dimensionality reduction result of BrainPepPass in pattern learning stage for FC-4. (a) Dataset 1. (b) Dataset 2. (c) Dataset 3.



Source: Author's own.

4.2.4 Performance comparison with web servers in B3PPs prediction

The performance of the DPF-3BPPred is also compared with previously developed techniques for predicting B3PPs. While some ML-based tools, such as BBPpred, B3Pred, BBPpredict, and SCMB3PP have been developed to predict the BBB permeability of peptides using ML and other computational algorithms trained with properties extracted from the primary structure of natural peptides encoded in FASTA format, the proposed ML-based framework employs a distinct approach by incorporating the 3D structure of these molecules encoded in MOL format. Additionally, most peptides used for training and testing DPF-3BPPred contain chemical modifications, further distinguishing this tool from those focusing on natural peptides.

A comparative analysis was conducted between DPF-3BPPred and the BBPpred, BBPpredict, 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, the version based on FC-4 and trained with Dataset 2, which achieved the best performance in LOOCV analyses, was selected for this test. Seventeen natural BBB+ peptides extracted from Brainpeps were selected to compare the performance of the models since none of these molecules were utilized in any of the previously described training or independent testing steps for the selected DPF-3BPPred 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 DPF-3BPPred model with FC-4, named DPF-3BPPred-N, which was exclusively trained using natural peptides collected from the same dataset that was used to train the SCMB3PP model. Table 8 presents the values all algorithms achieved based on the key metrics. Appendix M lists the peptide sequences used in this analysis.

Table 8 – Analysis of independent test comparing DPF-3BPPred and DPF-3BPPred-N with BBPpred, BBPpredict, and SCMB3PP algorithms using natural peptides.

Algorithm	Accuracy	F1-score	MCC	Precision	Recall
DPF-3BPPred	0.52	0.55	0.06	0.55	0.55
DPF-3BPPred-N	0.97	1.0	0.94	1.0	1.0
BBPpredict	0.64	0.71	0.33	0.60	0.88
BBPpred	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 8, DPF-3BPPred-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 more accurately than the other tools. It is also important to highlight that the DPF-3BPPred 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 concerning natural peptide data, considering that it was predominantly trained on structures featuring chemical

modifications.

Therefore, based on the results presented in this analysis, the proposed framework exhibits exceptional performance in predicting peptide penetration across the BBB, surpassing existing ML classifiers in this area of research. DPF-3BPPred achieved average accuracy values exceeding 93% in the 10-fold cross-validation and between 75% and 90% in the independent test (see the Table on Appendix K). For the FC-4 model, which exhibited a positive relationship between efficiency and complexity, average accuracy values of 99.21% and 75% were achieved in cross-validation and independent testing, respectively, across all three datasets (see the Table on Appendix K). Based on the molecular descriptors examined, these outcomes demonstrate that the proposed tool has predictive capabilities in determining whether natural or chemically modified peptides can penetrate the BBB.

4.3 Conclusion

This chapter presented the results obtained by the proposed ML framework developed to predict the permeability of peptides across the cell membrane and blood-brain barrier according to cross-validation and an independent test for both biomembranes. The chapter also explored how the groups of molecular descriptors contribute to the prediction of peptides. The next chapter concludes the thesis by approaching the main points revealed by the results, besides proposing future tasks.

5 CONCLUSIONS AND FUTURE WORKS

5.1 Final remarks

The present thesis proposed a machine learning framework based on dimensionality reduction and dimensional pattern learning to predict the permeability of peptides across the cell membrane and blood-brain barrier and visualize the 3D distribution of the molecules according to each class. The importance behind the prediction of penetration of peptides across biomembranes, as well as the theory behind each topic of the proposed method was approached in **Chapter 1** and **Chapter 2**, respectively.

Chapter 3 addresses the proposed method in this thesis and its main contributions, detailing how the peptide samples were obtained, processed, and divided between the appropriate FCs for both case studies. Regarding the problem of B3PPs classification, this chapter approaches the proposal pre-classification of peptides from the Brainpeps database in terms of BBB permeability based on experimentally validated markers. The methodology also scrutinizes the contributions regarding developing a framework architecture based on machine learning for predicting CPPs and B3PPs. Different from other state-of-the-art tools, the proposed framework uses supervised manifold dimensionality reduction as one of its steps performed by sLE algorithm, which contributes to visualizing the distribution of peptide samples in 3D with a significant degree of separation and clustering by class, as well as to classifying these molecules. Finally, the chapter also covers the other proposal within this framework, which is the pattern learning of reduction data using XGBoost algorithm.

The results presented in **Chapter 4** of this thesis reveal some interesting points about the analyses carried out on the prediction of penetration of the peptides in the two biomembranes. Regarding permeability across the cell membrane, it was shown that DPF-CPPred achieved interesting results for as much in 10-fold cross-validation as in the independent test, where the framework proved to be superior to Vcf-CPP and ANN, GPC, and SVM for both data structures. Comparatively, the cross-validation analysis indicated that the DPF-CPPred reached average accuracies 11% and 13.1% greater than the Vcf-CPP, which was the second-best classifier, for PDB and FASTA, respectively. Furthermore, the proposed tool also achieved an accuracy of 1.4% greater than the Vcf-CPP in the independent test. The proposed framework also outperformed state-of-the-art tools such as MLCPP, CPPred-RF, SkipCPP-Pred, and Kelm-CPPred, obtaining accuracies of 2.2% and 3.4% greater than the best web server in the experiments involving FASTA and PDB format, respectively. These results show that this thesis's main contribution was developed and achieved satisfactory results in predicting cell-penetrating peptides with the best performance with 92% accuracy in an independent test. It is also essential to highlight the findings regarding the composition of features, where the comparison between

descriptors based on amino acid sequence (AAC, DPC, and PseAAC) and the structural and physicochemical properties, which related to the oral bioavailability of molecules, revealed that the use of joint and optimized of these two groups promote a superior performance when compared to isolated use. This finding is important because it highlights these descriptors' contribution to the cell membrane's pharmacokinetic properties, facilitating the development of new drugs and biotechnological applications.

Regarding the data structures used in the cell membrane problem (FASTA and PDB), this thesis also brought significant contributions to the use of these file formats. First, the proposed framework can receive the features extracted from both formats, which differs from previous published ML tools. Furthermore, the results of this thesis also revealed how much the complexity of information coding can contribute to more accurate analysis since the DPF-CPPred trained and tested with peptides in PDB format achieved results numerically superior to the model trained and tested with FASTA. However, there is a counterpoint to conclude if FASTA-based DPF-CPPred is better than PDB-based one. On its side, the PDB format can encode peptide structures with chemical modifications and non-natural amino acids. On the other hand, the FASTA file encodes primary peptide structures, encompassing only natural molecules without chemical modifications. Although in **Chapter 4** we see that the FASTA-based model performs computationally smaller than the PDB-based model, in practical terms, the primary structure of the peptides is simpler to obtain than the tertiary one, on the other hand, the PDB is more useful in applications where modifications in the structure of the molecule are necessary. It is also important to highlight that during the development of this work, an open access web server¹ was developed with Vcf-CPP, which the scientific community has used to predict CPPs using both data structures

Similar to the case of the cell membrane, the results for predicting peptide penetration in the blood-brain barrier also reveal essential points. First, one of the highlights contributions of this thesis is the classification of Brainpeps peptides as permeable or not across the BBB based on the values of experimental parameters. This contribution was essential for constructing a reliable and experimentally validated dataset to train and evaluate the framework for this prediction, differentiating from previous works that focused on using peptides without experimental validation. Second, the results also show the good performance achieved by the DPF-3BPPred, outperforming the predictive capacity of Vcf-3BPP and its classifiers when analyzed with 10-fold cross-validation, where the framework reached average values ranging from 93% to 100%, which are 30% greater than the result reached by the second-best classifier (ANN). In the independent test, DPF-3BPPred obtained accuracy between 75% and 90%, which was 10% superior to the value obtained by the ANN in the best scenario.

Still, concerning these results, it was evident that the FC based on the ten best molecular properties from Mordred's descriptors along with the descriptors evaluated by Dichiara and her

¹The access to the web server is available in <http://comptools.linc.ufpa.br/BChemRF-CPPred/>

collaborators (FC-4) provided higher performance for the framework than these groups of properties evaluated isolated (FC-1, FC-2, and FC-3). This finding is compelling and, at the same time, controversial because Mordred's descriptors, specifically JGI5, JGI6, JGI7, JGI9, EState-VSA5, GATS3d, nAcid, RotRatio, and GhoseFilter, have never been experimentally evaluated to have a high correlation with the ability of a molecule to penetrate the BBB. In contrast, Diachara's descriptors have this correlation evaluated for several classes of molecules.

Despite the encouraging results achieved so far for the BBB case study showing significant performance for the proposed tool, there are still difficulties that prevent a more accurate analysis of the performance of the DPF-3BPPred, mainly due to the input format of peptides. Most ML-based tools to predict B3PPs use the FASTA format, while the original proposed framework was trained with natural and chemically modified peptides using in MDL format. Evaluating only the prediction of natural peptides across the BBB, a version of the proposed framework trained with only natural peptides (DPF-3BPPred-N) demonstrated to overcome the performance of other published tools, achieving 97% of accuracy in predicting natural B3PPs.

In a general analysis of this work, it can be concluded that the main objectives of this thesis were achieved with good results, considering that the framework architecture planned and developed to predict CPPs not only performed well in predicting these molecules but is also capable of processing peptides of both PDB and FASTA formats. The results also demonstrated that the framework performed better than baseline ML models for B3PPs prediction, besides indicating new possibilities for molecular filters to predict penetration into this barrier. Therefore, the proposed framework can be used to help scientists in the process of virtual screening of peptides for their permeability in the cell membrane and the blood-brain barrier.

5.2 Future works

Some relevant points can still be addressed regarding the contributions this work can bring to the state of the art. These points are listed below.

The first point would be to obtain more samples of peptides with permeability experimentally validated for BBB and retrain the DPF-3BPPred. It could improve the framework performance based on data encompassing a wide variety of validated peptides with distinct characteristics.

Another point that could be interesting to investigate would be the performance evaluation of the DPF-CPPred for predicting the permeability of the remaining CPPs samples in an independent test. This evaluation can provide additional information regarding the capacity of the proposed framework to correctly predict CPPs, which could have different characteristics from the training dataset.

Develop a pipeline to sample the most balanced training and testing datasets for CPPs and B3PPs concerning the distribution of descriptors to prevent distribution shift using Tani-

moto similarity or Jaccard or Sørensen–Dice coefficients is another contribution to this research field. This process could split the peptide data between these two datasets with more balance, besides preventing overfitting of the ML models.

Evaluate the performance of DPF-CPPred and DPF-3BPPred by investigating underperforming subpopulations in their training datasets is another aspect that can contribute to the improvement of the frameworks. This investigation can identify the subgroup of peptide samples with molecular descriptors that contribute most significantly to accurately distinguishing between the two permeability classes.

Another point that can be investigated for predicting CPPs and B3PPs using the proposed frameworks is the extraction of Macromolecular Reactivity Descriptors provided by PRIMoRDiA package. These descriptors are useful for understanding molecular interaction mechanisms and designing new drugs.

Create a pipeline to generate synthetic peptides to evaluate their permeability across the BBB using a generative neural network is a relevant contribution to constructing a robust database that can refine the training of predictive ML models, besides providing synthetic samples that can be experimentally investigated.

Compare the performance of the DPF-CPPred and DPF-3BPPred with other ML models such as Logistic Regression, Deep Learning, Random Forest, Radial Basis Function Neural Networks, and compare the performance of the frameworks substituting the sLE algorithm for another supervised dimensionality reduction technique are two points of investigation that can highlight the effectiveness of the proposed frameworks for predicting CPPs and B3PPs in comparison with other state-of-the-art predictive models.

Develop a generalist framework capable of predicting, at the same time, if a given peptide is capable of crossing the cell membrane and the BBB is a contribution to this research field, which can optimize research regarding the pharmacokinetics of these two classes of peptides.

Develop a pipeline to predict CPPs and B3PPs using Graph Neural Networks can contribute to this research field since this kind of technique can incorporate information regarding the distribution of the atoms in a graph and associate these distribution patterns with the permeability of the peptides.

Select some predicted CPPs by DPF-CPPred and investigating the penetration mechanism using molecular dynamics is another contribution to this research field and can provide information regarding the penetration mechanism of these peptides into this biomembrane.

5.3 Publications

Journal Papers

Oliveira, E.C.L., Santana, K., Josino, L., Lima e Lima, A.H., de Souza de Sales Júnior, C., 2021. *Predicting cell-penetrating peptides using machine learning algorithms and navigating in their chemical space*. Sci. Rep. 11. <https://doi.org/10.1038/s41598-021-87134-w>.

(See Appendix M)

Oliveira, E. C. L.; da Costa, K. S.; Taube, P. S.; Lima, A. H.; Junior, C. S. S., 2022. *Biological Membrane-Penetrating Peptides: Computational Prediction and Applications*. Frontiers in Cellular and Infection Microbiology. 12. <https://doi.org/10.3389/fcimb.2022.838259>.

(See Appendix O)

Oliveira, E. C. L.; Hirmz, H.; Wynendaele, E.; Feio, J. A. S.; Moreira, I. M. S.; da Costa, K. S.; Lima, A. H.; Spiegeleer, B. D.; Junior, C. S. S., 2023. *BrainPepPass: A Framework Based on Supervised Dimensionality Reduction for Predicting Blood-Brain Barrier-Penetrating Peptides*. Journal of Chemical Information and Modeling. <https://doi.org/10.1021/acs.jcim.3c00951>.

(See Appendix P)

Conference Works

MATOS, W. L. N. ; LEAL, A. F. R. ; OLIVEIRA, E. C. L. ; CUNHA, E. J. S. *Aplicação de Rede Neural Artificial na Classificação de Distância de Ocorrência de Raios Nuvem-Solo Negativos*. In: XL Congresso Nacional de Matemática Aplicada e Computacional, 2021. Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, 2021. v. 8.

REFERENCES

- A. DOBCHEV, Dimitar et al. Prediction of Cell-Penetrating Peptides Using Artificial Neural Networks. **Current Computer Aided-Drug Design**, v. 6, n. 2, p. 79–89, June 2010. ISSN 15734099. DOI: 10.2174/157340910791202478. Available from: <http://link.springer.com/10.1007/978-1-4939-2239-0_7<http://www.eurekaselect.com/openurl/content.php?genre=article&issn=1573-4099&volume=6&issue=2&spage=79>>.
- ABRAHAM, Michael H. The factors that influence permeation across the blood–brain barrier. **European Journal of Medicinal Chemistry**, v. 39, n. 3, p. 235–240, Mar. 2004. ISSN 02235234. DOI: 10.1016/j.ejmech.2003.12.004. Available from: <<https://linkinghub.elsevier.com/retrieve/pii/S0223523403002356>>.
- AGRAWAL, Piyush et al. CPPsite 2.0: a repository of experimentally validated cell-penetrating peptides. **Nucleic Acids Research**, v. 44, n. D1, p. d1098–d1103, Jan. 2016. ISSN 0305-1048. DOI: 10.1093/nar/gkv1266. Available from: <<https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv1266>>.
- AHLAWAT, Jyoti et al. Nanocarriers as Potential Drug Delivery Candidates for Overcoming the Blood–Brain Barrier: Challenges and Possibilities. **ACS Omega**, v. 5, n. 22, p. 12583–12595, June 2020. ISSN 2470-1343. DOI: 10.1021/acsomega.0c01592. Available from: <<https://pubs.acs.org/doi/10.1021/acsomega.0c01592>>.
- ANASPEC. **Cell Permeable Peptides (CPP)/Drug Delivery Peptides**. 2010. In Anaspec’s Catalog Listing of Cell Permeable Peptides.
- ANNUNZIATO, Giannamaria; COSTANTINO, Gabriele. Antimicrobial peptides (AMPs): a patent review (2015–2020). **Expert Opinion on Therapeutic Patents**, v. 30, n. 12, p. 931–947, Dec. 2020. ISSN 1354-3776. DOI: 10.1080/13543776.2020.1851679. Available from: <<https://www.tandfonline.com/doi/full/10.1080/13543776.2020.1851679>>.
- ARIF ALI, Zeravan et al. eXtreme Gradient Boosting Algorithm with Machine Learning: a Review. **Academic Journal of Nawroz University**, v. 12, n. 2, p. 320–334, May 2023. ISSN 2520-789X. DOI: 10.25007/ajnu.v12n2a1612. Available from: <<https://journals.nawroz.edu.krd/index.php/ajnu/article/view/1612>>.
- ARNOTT, John A; PLANEY, Sonia Lobo. The influence of lipophilicity in drug discovery and design. **Expert Opinion on Drug Discovery**, v. 7, n. 10, p. 863–875, Oct. 2012. ISSN

1746-0441. DOI: 10.1517/17460441.2012.714363. Available from: <<http://www.tandfonline.com/doi/full/10.1517/17460441.2012.714363>>.

ARUNAN, Elangannan et al. Definition of the hydrogen bond (IUPAC Recommendations 2011). **Pure and Applied Chemistry**, v. 83, n. 8, p. 1637–1641, July 2011. ISSN 1365-3075. DOI: 10.1351/PAC-REC-10-01-02. Available from: <<https://www.degruyter.com/document/doi/10.1351/PAC-REC-10-01-02/html>>.

BAGCHI, Sounak et al. In-vitro blood-brain barrier models for drug screening and permeation studies: An overview. **Drug Design, Development and Therapy**, v. 13, p. 3591–3605, 2019. ISSN 11778881. DOI: 10.2147/DDDT.S218708.

BAIG, Mohammad Hassan et al. Peptide based therapeutics and their use for the treatment of neurodegenerative and other diseases. **Biomedicine and Pharmacotherapy**, Elsevier, v. 103, March, p. 574–581, 2018. ISSN 19506007. DOI: 10.1016/j.biopha.2018.04.025. Available from: <<https://doi.org/10.1016/j.biopha.2018.04.025>>.

BENET, Leslie Z. et al. BDDCS, the Rule of 5 and drugability. **Advanced Drug Delivery Reviews**, v. 101, p. 89–98, June 2016. ISSN 0169409X. DOI: 10.1016/j.addr.2016.05.007. Available from: <<https://linkinghub.elsevier.com/retrieve/pii/S0169409X16301491>>.

BHANDARI, Divya et al. A Review on Bioactive Peptides: Physiological Functions, Bioavailability and Safety. **International Journal of Peptide Research and Therapeutics**, Springer Netherlands, v. 0, n. 0, p. 0, 2019. ISSN 1573-3904. DOI: 10.1007/s10989-019-09823-5. Available from: <<http://dx.doi.org/10.1007/s10989-019-09823-5>>.

BOLHASSANI, Azam. Potential efficacy of cell-penetrating peptides for nucleic acid and drug delivery in cancer. **Biochimica et Biophysica Acta (BBA) - Reviews on Cancer**, v. 1816, n. 2, p. 232–246, Dec. 2011. ISSN 0304419X. DOI: 10.1016/j.bbcan.2011.07.006. Available from: <<https://linkinghub.elsevier.com/retrieve/pii/S0304419X11000448>>.

BRACKE, Nathalie et al. Blood-brain barrier transport kinetics of NOTA-modified proteins: the somatropin case. **The Quarterly Journal of Nuclear Medicine and Molecular Imaging**, v. 64, n. 1, Apr. 2020. ISSN 18244785. DOI: 10.23736/S1824-4785.18.03025-X. Available from: <<https://www.minervamedica.it/index2.php?show=R39Y2020N01A0105>>.

CHANTEMARGUE, Benjamin. **In silico investigation of xenobiotic interactions with lipid bilayers and ABC membrane transporters, the case of ABCC4/MRP4**. 2018. S. 210.

PhD thesis – Université de Limoges. Available from: <<https://www.unilim.fr/ippritt/events/event/benjamin-chantemargue/>>.

CHAO, Guoqing; LUO, Yuan; DING, Weiping. Recent Advances in Supervised Dimension Reduction: A Survey. **Machine Learning and Knowledge Extraction**, v. 1, n. 1, p. 341–358, 2019. DOI: 10.3390/make1010020.

CHAROENKWAN, Phasit et al. Improved prediction and characterization of blood-brain barrier penetrating peptides using estimated propensity scores of dipeptides. **Journal of Computer-Aided Molecular Design**, v. 36, n. 11, p. 781–796, Nov. 2022. ISSN 0920-654X. DOI: 10.1007/s10822-022-00476-z. Available from:

<<https://link.springer.com/10.1007/s10822-022-00476-z>>.

CHEN, Lei et al. Prediction and analysis of cell-penetrating peptides using pseudo-amino acid composition and random forest models. **Amino Acids**, Springer Vienna, v. 47, n. 7, p. 1485–1493, July 2015. ISSN 0939-4451. DOI: 10.1007/s00726-015-1974-5.

Available from: <<http://dx.doi.org/10.1007/s00726-015-1974-5>>
<<http://link.springer.com/10.1007/s00726-015-1974-5>>.

CHEN, Long et al. Blood–Brain Barrier- and Blood–Brain Tumor Barrier-Penetrating Peptide-Derived Targeted Therapeutics for Glioma and Malignant Tumor Brain Metastases. **ACS Applied Materials Interfaces**, v. 11, n. 45, p. 41889–41897, Nov. 2019. ISSN 1944-8244. DOI: 10.1021/acsami.9b14046. Available from:

<<https://pubs.acs.org/doi/10.1021/acsami.9b14046>>.

CHEN, Tianqi; GUESTRIN, Carlos. XGBoost. In: PROCEEDINGS of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM, Aug. 2016. P. 785–794. ISBN 9781450342322. DOI: 10.1145/2939672.2939785. Available from:

<<https://dl.acm.org/doi/10.1145/2939672.2939785>>.

CHEN, Xue et al. BBPpredict: A Web Service for Identifying Blood-Brain Barrier Penetrating Peptides. **Frontiers in Genetics**, v. 13, May, p. 1–10, May 2022. ISSN 1664-8021. DOI: 10.3389/fgene.2022.845747. Available from: <<https://www.frontiersin.org/articles/10.3389/fgene.2022.845747/full>>.

CHOU, Kuo-Chen. Prediction of protein cellular attributes using pseudo-amino acid composition. **Proteins: Structure, Function, and Genetics**, v. 43, n. 3, p. 246–255, May 2001. ISSN 0887-3585. DOI: 10.1002/prot.1035. Available from:

<<https://onlinelibrary.wiley.com/doi/10.1002/prot.1035>>.

CHOWDHARY, K. R. **Fundamentals of artificial intelligence**. 2020. P. 1–716. ISBN 9788132239727. DOI: 10.1007/978-81-322-3972-7.

COCK, P. J. A. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. **Bioinformatics**, v. 25, n. 11, p. 1422–1423, June 2009. ISSN 1367-4803. DOI: 10.1093/bioinformatics/btp163. Available from: <<https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp163>>.

CÖMERT, Zafer; KOCAMAZ, Adnan. A Study of Artificial Neural Network Training Algorithms for Classification of Cardiotocography Signals. **Bitlis Eren University Journal of Science and Technology**, v. 7, n. 2, p. 93–103, Dec. 2017. ISSN 2146-7706. DOI: 10.17678/beuscitech.338085. Available from: <<https://dergipark.org.tr/en/doi/10.17678/beuscitech.338085>>.

COPOLOVICI, Dana Maria et al. Cell-Penetrating Peptides: Design, Synthesis, and Applications. **ACS Nano**, v. 8, n. 3, p. 1972–1994, Mar. 2014. ISSN 1936-0851. DOI: 10.1021/nn4057269. Available from: <<https://pubs.acs.org/doi/10.1021/nn4057269>>.

COUSINEAU, Denis; CHARTIER, Sylvain. Outliers detection and treatment: a review. **International Journal of Psychological Research**, v. 3, n. 1, p. 58–67, June 2010. ISSN 2011-7922. DOI: 10.21500/20112084.844. Available from: <<https://revistas.usb.edu.co/index.php/IJPR/article/view/844>>.

DAI, Ruyu et al. BBPpred: Sequence-Based Prediction of Blood-Brain Barrier Peptides with Feature Representation Learning and Logistic Regression. **Journal of Chemical Information and Modeling**, v. 61, n. 1, p. 525–534, Jan. 2021. ISSN 1549-9596. DOI: 10.1021/acs.jcim.0c01115. Available from: <<https://pubs.acs.org/doi/10.1021/acs.jcim.0c01115>>.

DAINA, Antoine; ZOETE, Vincent. A BOILED-Egg To Predict Gastrointestinal Absorption and Brain Penetration of Small Molecules. **ChemMedChem**, v. 11, n. 11, p. 1117–1121, June 2016. ISSN 18607179. DOI: 10.1002/cmdc.201600182. Available from: <<http://doi.wiley.com/10.1002/cmdc.201600182>>.

DAMIATI, Safa A et al. Novel machine learning application for prediction of membrane insertion potential of cell-penetrating peptides. **International Journal of Pharmaceutics**, v. 567, p. 118453, Aug. 2019. ISSN 03785173. DOI: 10.1016/j.ijpharm.2019.118453. Available from: <<https://linkinghub.elsevier.com/retrieve/pii/S0378517319304879>>.

DANEMAN, Richard; PRAT, Alexandre. The Blood–Brain Barrier. **Cold Spring Harbor Perspectives in Biology**, v. 7, n. 1, a020412, Jan. 2015. ISSN 1943-0264. DOI: 10.1101/cshperspect.a020412. Available from: <<http://cshperspectives.cshlp.org/lookup/doi/10.1101/cshperspect.a020412>>.

DENG, Naiyang; TIAN, Yingjie; ZHANG, Chunhua. **Support Vector Machines Optimization Based Theory, Algorithms, and Extensions**. CRC Press, 2013. ISBN 9781439857939.

DERAKHSHANKHAH, Hossein; JAFARI, Samira. Cell penetrating peptides: A concise review with emphasis on biomedical applications. **Biomedicine Pharmacotherapy**, Elsevier, v. 108, June, p. 1090–1096, Dec. 2018. ISSN 07533322. DOI: 10.1016/j.biopha.2018.09.097. Available from: <<https://doi.org/10.1016/j.biopha.2018.09.097>><https://linkinghub.elsevier.com/retrieve/pii/S0753332218340186>>.

DI, Li et al. High throughput artificial membrane permeability assay for blood–brain barrier. **European Journal of Medicinal Chemistry**, v. 38, n. 3, p. 223–232, Mar. 2003. ISSN 02235234. DOI: 10.1016/S0223-5234(03)00012-6. Available from: <<https://linkinghub.elsevier.com/retrieve/pii/S0223523403000126>>.

DÍAZ-EUFRACIO, Bárbara I. et al. Exploring the chemical space of peptides for drug discovery: a focus on linear and cyclic penta-peptides. **Molecular Diversity**, Springer International Publishing, v. 22, n. 2, p. 259–267, May 2018. ISSN 1381-1991. DOI: 10.1007/s11030-018-9812-9. Available from: <<https://doi.org/10.1007/s11030-018-9812-9>><http://link.springer.com/10.1007/s11030-018-9812-9>>.

DÍAZ-PERLAS, Cristina et al. Branched BBB-shuttle peptides: chemoselective modification of proteins to enhance blood-brain barrier transport. **Chemical Science**, v. 9, n. 44, p. 8409–8415, 2018. ISSN 20416539. DOI: 10.1039/c8sc02415d.

DICHIARA, Maria et al. Tuning Properties for Blood–Brain Barrier Permeation: A Statistics-Based Analysis. **ACS Chemical Neuroscience**, v. 11, n. 1, p. 34–44, Jan. 2020. ISSN 1948-7193. DOI: 10.1021/acscchemneuro.9b00541.

DIENER, Christian et al. Effective Design of Multifunctional Peptides by Combining Compatible Functions. **PLoS Computational Biology**, v. 12, n. 4, p. 1–19, 2016. ISSN 15537358. DOI: 10.1371/journal.pcbi.1004786.

DOAK, Bradley Croy et al. Oral druggable space beyond the rule of 5: Insights from drugs and clinical candidates. **Chemistry and Biology**, v. 21, n. 9, p. 1115–1142, Sept. 2014. ISSN 10745521. DOI: 10.1016/j.chembiol.2014.08.013.

DONG, Jie et al. PyBioMed: a python library for various molecular representations of chemicals, proteins and DNAs and their interactions. **Journal of Cheminformatics**, v. 10, n. 1, p. 16, Dec. 2018. ISSN 1758-2946. DOI: 10.1186/s13321-018-0270-2.

Available from:

<<https://jcheminf.biomedcentral.com/articles/10.1186/s13321-018-0270-2>>.

DONIGER, Scott; HOFMANN, Thomas; YEH, Joanne. Predicting CNS Permeability of Drug Molecules: Comparison of Neural Network and Support Vector Machine Algorithms. **Journal of Computational Biology**, v. 9, n. 6, p. 849–864, Dec. 2002. ISSN 1066-5277. DOI:

10.1089/10665270260518317. Available from:

<<http://www.liebertpub.com/doi/10.1089/10665270260518317>>.

DOUGHERTY, Patrick G.; SAHNI, Ashweta; PEI, Dehua. Understanding Cell Penetration of Cyclic Peptides. **Chemical Reviews**, v. 119, n. 17, p. 10241–10287, Sept. 2019. ISSN

0009-2665. DOI: 10.1021/acs.chemrev.9b00008. Available from:

<<https://pubs.acs.org/doi/10.1021/acs.chemrev.9b00008>>.

EBDEN, Mark. Gaussian Processes: A Quick Introduction. July, 2015. arXiv: 1505.02965.

Available from: <<http://arxiv.org/abs/1505.02965>>.

EMMERT-STREIB, Frank et al. An Introductory Review of Deep Learning for Prediction Models With Big Data. **Frontiers in Artificial Intelligence**, v. 3, February, p. 1–23, Feb.

2020. ISSN 2624-8212. DOI: 10.3389/frai.2020.00004. Available from: <<https://www.frontiersin.org/article/10.3389/frai.2020.00004/full>>.

FEIGIN, Valery L et al. Global, regional, and national burden of neurological disorders, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. **The Lancet Neurology**, v. 18, n. 5, p. 459–480, May 2019. ISSN 14744422. DOI:

10.1016/S1474-4422(18)30499-X. Available from: <<https://linkinghub.elsevier.com/retrieve/pii/S147444221830499X>>.

FELEGYI-TÓTH, Csenge Anna et al. Membrane Permeability and Aqueous Stability Study of

Linear and Cyclic Diarylheptanoids from *Corylus maxima*. **Pharmaceutics**, v. 14, n. 6, p. 1250, June 2022. ISSN 1999-4923. DOI: 10.3390/pharmaceutics14061250.

Available from: <<https://www.mdpi.com/1999-4923/14/6/1250>>.

FLEXA, Caio et al. Mutual equidistant-scattering criterion: A new index for crisp clustering. **Expert Systems with Applications**, v. 128, p. 225–245, Aug. 2019. ISSN 09574174. DOI: 10.1016/j.eswa.2019.03.027. Available from: <<https://linkinghub.elsevier.com/retrieve/pii/S0957417419301897>>.

FORBES, Jessica; KRISHNAMURTHY, Karthik. Biochemistry, Peptide. In: STATPEARLS [Internet]. Treasure Island (FL): StatPearls Publishing, 2021. Available from: <<https://www.statpearls.com/articlelibrary/viewarticle/26914/>>.

FU, Xiangzheng et al. Improved Prediction of Cell-Penetrating Peptides via Effective Orchestrating Amino Acid Composition Feature Representation. **IEEE Access**, 2019. DOI: 10.1109/ACCESS.2019.2952738.

FU, Yun. **Manifold Learning Theory and Applications**. Ed. by Yunqian Ma and Yun Fu. CRC Press, Dec. 2011. ISBN 9781439871102. DOI: 10.1201/b11431. Available from: <<https://www.taylorfrancis.com/books/9781439871102>>.

GALVEZ, J. et al. Charge Indexes. New Topological Descriptors. **Journal of Chemical Information and Computer Sciences**, v. 34, n. 3, p. 520–525, May 1994. ISSN 0095-2338. DOI: 10.1021/ci00019a008. Available from: <<https://pubs.acs.org/doi/abs/10.1021/ci00019a008>>.

GARG, Arunim; MAGO, Vijay. Role of machine learning in medical research: A survey. **Computer Science Review**, v. 40, p. 100370, May 2021. ISSN 15740137. DOI: 10.1016/j.cosrev.2021.100370. Available from: <<https://linkinghub.elsevier.com/retrieve/pii/S1574013721000101>>.

GAUTAM, Ankur et al. In silico approaches for designing highly effective cell penetrating peptides. **Journal of Translational Medicine**, v. 11, n. 1, p. 74, 2013. ISSN 1479-5876. DOI: 10.1186/1479-5876-11-74. Available from: <<http://translational-medicine.biomedcentral.com/articles/10.1186/1479-5876-11-74>>.

GELDENHUYS, Werner J et al. Molecular determinants of blood–brain barrier permeation. **Therapeutic Delivery**, v. 6, n. 8, p. 961–971, Aug. 2015. ISSN 2041-5990. DOI: 10.4155/tde.15.32. Available from: <<https://www.future-science.com/doi/10.4155/tde.15.32>>.

GÉRON, Aurélien. **Hands-on Machine Learning whith Scikit-Learning, Keras and Tensorflow**. 2019. P. 510. ISBN 978-1-492-03264-9.

GHOJOGH, Benyamin et al. **Elements of Dimensionality Reduction and Manifold Learning**. Cham: Springer International Publishing, 2023. P. 1–606. ISBN

978-3-031-10601-9. DOI: 10.1007/978-3-031-10602-6. Available from:
<<https://link.springer.com/10.1007/978-3-031-10602-6>>.

GHOLAMI, Raoof; FAKHARI, Nikoo. Support Vector Machine: Principles, Parameters, and Applications. In: HANDBOOK of Neural Computation. 1. ed.: Elsevier, 2017. P. 515–535. DOI: 10.1016/B978-0-12-811318-9.00027-2.

GUIDOTTI, Giulia; BRAMBILLA, Liliana; ROSSI, Daniela. Cell-Penetrating Peptides: From Basic Research to Clinics. **Trends in Pharmacological Sciences**, Elsevier Ltd, v. 38, n. 4, p. 406–424, Apr. 2017. ISSN 01656147. DOI: 10.1016/j.tips.2017.01.003. Available from:

<<http://dx.doi.org/10.1016/j.tips.2017.01.003>>
<<https://linkinghub.elsevier.com/retrieve/pii/S0165614717300172>>.

HAGAN, Martin T et al. **Neural Network Design**. 2. ed.: Martin Hagan, 2014. ISBN 0971732116.

HÄLLBRINK, Mattias et al. Prediction of Cell-Penetrating Peptides. **International Journal of Peptide Research and Therapeutics**, v. 11, n. 4, p. 249–259, Dec. 2005. ISSN 1573-3149. DOI: 10.1007/s10989-005-9393-1. Available from:
<<https://link.springer.com/10.1007/s10989-005-9393-1>>.

HANNON, Christine L.; ANSLYN, Eric V. The Guanidinium Group: Its Biological Role and Synthetic Analogs. v. 3, p. 193–255, 1993. DOI: 10.1007/978-3-642-78110-0_6.

HANSEN, Mats; KILK, Kalle; LANGEL, Ülo. Predicting cell-penetrating peptides. **Advanced Drug Delivery Reviews**, v. 60, n. 4-5, p. 572–579, Mar. 2008. ISSN 0169409X. DOI: 10.1016/j.addr.2007.09.003. Available from: <<https://linkinghub.elsevier.com/retrieve/pii/S0169409X07002918>>.

HAYKIN, Simon. **Neural Networks: A Comprehensive Foundation**. 2. ed.: Prentice Hall, 1998. ISBN 9780132733502.

ICHIM, Gabriel; TAUSZIG-DELAMASURE, Servane; MEHLEN, Patrick. Neurotrophins and cell death. **Experimental Cell Research**, Elsevier Inc., v. 318, n. 11, p. 1221–1228, 2012. ISSN 0014-4827. DOI: 10.1016/j.yexcr.2012.03.006. Available from:
<<http://dx.doi.org/10.1016/j.yexcr.2012.03.006>>.

INSTITUTE, National Cancer. FASTA Format. In: DEFINITIONS. Qeios, Feb. 2020. DOI: 10.32388/2B8KEZ. Available from:
<<https://www.qeios.com/read/definition/59916>>.

JACKMAN, Joshua A. et al. Therapeutic treatment of Zika virus infection using a brain-penetrating antiviral peptide. **Nature Materials**, v. 17, n. 11, p. 971–977, Nov. 2018. ISSN 1476-1122. DOI: 10.1038/s41563-018-0194-2. Available from: <<https://www.nature.com/articles/s41563-018-0194-2>>.

JAMES, Gareth et al. **An Introduction to Statistical Learning**. New York, NY: Springer New York, 2013. v. 103. (Springer Texts in Statistics). ISBN 978-1-4614-7137-0. DOI: 10.1007/978-1-4614-7138-7. Available from: <<http://link.springer.com/10.1007/978-1-4614-7138-7>>.

JANSSENS, Yorick et al. PapRIV, a BV-2 microglial cell activating quorum sensing peptide. **Scientific Reports**, Nature Publishing Group UK, v. 11, n. 1, p. 1–14, 2021. ISSN 20452322. DOI: 10.1038/s41598-021-90030-y. Available from: <<https://doi.org/10.1038/s41598-021-90030-y>>.

JINURAJ, K. R. et al. Feature optimization in high dimensional chemical space: Statistical and data mining solutions. **BMC Research Notes**, BioMed Central, v. 11, n. 1, p. 1–7, 2018. ISSN 17560500. DOI: 10.1186/s13104-018-3535-y. Available from: <<https://doi.org/10.1186/s13104-018-3535-y>>.

KARDANI, Kimia et al. Cell penetrating peptides: the potent multi-cargo intracellular carriers. **Expert Opinion on Drug Delivery**, 2019. ISSN 1742-5247. DOI: 10.1080/17425247.2019.1676720.

KELLEHER, John; MAC NAMEE, Brian; D'ARCY, Aoife. **Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies**. 2020. v. 53, p. 853. ISBN 9780262044691. eprint: arXiv:1011.1669v3.

KINGMA, Diederik P.; BA, Jimmy Lei. Adam: A method for stochastic optimization. **3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings**, p. 1–15, 2015. arXiv: 1412.6980.

KUMAR, Vinod; AGRAWAL, Piyush, et al. Prediction of Cell-Penetrating Potential of Modified Peptides Containing Natural and Chemically Modified Residues. **Frontiers in Microbiology**, v. 9, APR, p. 1–10, Apr. 2018. ISSN 1664-302X. DOI: 10.3389/fmicb.2018.00725. Available from: <<http://journal.frontiersin.org/article/10.3389/fmicb.2018.00725/full>>.

KUMAR, Vinod; PATIYAL, Sumeet, et al. B3pred: A random-forest-based method for predicting and designing blood–brain barrier penetrating peptides. **Pharmaceutics**, v. 13, n. 8, 2021. ISSN 19994923. DOI: 10.3390/pharmaceutics13081237.

LALATSA, Aikaterini; SCHATZLEIN, Andreas G.; UCHEGBU, Ijeoma F. Strategies To Deliver Peptide Drugs to the Brain. **Molecular Pharmaceutics**, v. 11, n. 4, p. 1081–1093, Apr. 2014. ISSN 1543-8384. DOI: 10.1021/mp400680d. Available from: <<https://pubs.acs.org/doi/10.1021/mp400680d>>.

LAMIABLE, Alexis et al. PEP-FOLD3: faster de novo structure prediction for linear peptides in solution and in complex. **Nucleic Acids Research**, v. 44, W1, w449–w454, July 2016. ISSN 0305-1048. DOI: 10.1093/nar/gkw329. Available from: <<https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw329>>.

LANGEL, Ülo et al. **Introduction to Peptides and Proteins**. 2009. P. 1–425. ISBN 9781439882047. DOI: 10.1201/b15106.

LEE, Andy Chi-Lung; HARRIS, Janelle Louise, et al. A Comprehensive Review on Current Advances in Peptide Drug Development and Design. **International Journal of Molecular Sciences**, v. 20, n. 10, p. 2383, May 2019. ISSN 1422-0067. DOI: 10.3390/ijms20102383. Available from: <<https://www.mdpi.com/1422-0067/20/10/2383>>.

LEE, M. R.; JAYANT, R. D. Penetration of the blood-brain barrier by peripheral neuropeptides: new approaches to enhancing transport and endogenous expression. **Cell and Tissue Research**, Cell and Tissue Research, v. 375, n. 1, p. 287–293, Jan. 2019. ISSN 0302-766X. DOI: 10.1007/s00441-018-2959-y. Available from: <<http://link.springer.com/10.1007/s00441-018-2959-y>>.

LI, Hu et al. Effect of Selection of Molecular Descriptors on the Prediction of BloodBrain Barrier Penetrating and Nonpenetrating Agents by Statistical Learning Methods. **Journal of Chemical Information and Modeling**, v. 45, n. 5, p. 1376–1384, Sept. 2005. ISSN 1549-9596. DOI: 10.1021/ci050135u. Available from: <<https://pubs.acs.org/doi/10.1021/ci050135u>>.

LIANG, Heng et al. XGBoost: An Optimal Machine Learning Model with Just Structural Features to Discover MOF Adsorbents of Xe/Kr. **ACS Omega**, v. 6, n. 13, p. 9066–9076, Apr. 2021. ISSN 2470-1343. DOI: 10.1021/acsomega.1c00100. Available from: <<https://pubs.acs.org/doi/10.1021/acsomega.1c00100>>.

LIPINSKI, Christopher A. et al. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. **Advanced Drug Delivery Reviews**, v. 64, SUPPL., p. 4–17, Dec. 2012. ISSN 0169409X. DOI: 10.1016/j.addr.2012.09.019. Available from: <<https://linkinghub.elsevier.com/retrieve/pii/S0169409X12002797>>.

- LOVERING, Frank. Escape from Flatland 2: complexity and promiscuity. **MedChemComm**, v. 4, n. 3, p. 515, 2013. ISSN 2040-2503. DOI: 10.1039/c2md20347b. Available from: <<http://xlink.rsc.org/?DOI=c2md20347b>>.
- LOVRIĆ, Mario; MOLERO, José Manuel; KERN, Roman. PySpark and RDKit: Moving towards Big Data in Cheminformatics. **Molecular Informatics**, v. 38, n. 6, p. 1800082, June 2019. ISSN 1868-1743. DOI: 10.1002/minf.201800082. Available from: <<https://onlinelibrary.wiley.com/doi/10.1002/minf.201800082>>.
- LUNDERMAN, Spencer et al. Screening Fuels for Autoignition with Small-Volume Experiments and Gaussian Process Classification. **Energy and Fuels**, v. 32, n. 9, p. 9581–9591, 2018. ISSN 15205029. DOI: 10.1021/acs.energyfuels.8b02112.
- LUNGA, Dalton et al. Manifold-learning-based feature extraction for classification of hyperspectral data: A review of advances in manifold learning. **IEEE Signal Processing Magazine**, v. 31, n. 1, p. 55–66, 2014. ISSN 10535888. DOI: 10.1109/MSP.2013.2279894.
- MANAVALAN, Balachandran et al. Machine-Learning-Based Prediction of Cell-Penetrating Peptides and Their Uptake Efficiency with Improved Accuracy. **Journal of Proteome Research**, American Chemical Society, v. 17, n. 8, p. 2715–2726, Aug. 2018. ISSN 1535-3893. DOI: 10.1021/acs.jproteome.8b00148. Available from: <<https://pubs.acs.org/doi/10.1021/acs.jproteome.8b00148>>.
- MATSUOKA, Masaaki. Humanin Signal for Alzheimer's Disease. **Journal of Alzheimer's Disease**, v. 24, p. 27–32, 2011. DOI: 10.3233/JAD-2011-102076.
- MCCULLOCH, Warren S.; PITTS, Walter. A logical calculus of the ideas immanent in nervous activity. **The Bulletin of Mathematical Biophysics**, v. 5, n. 4, p. 115–133, Dec. 1943. ISSN 0007-4985. DOI: 10.1007/BF02478259. Available from: <<http://link.springer.com/10.1007/BF02478259>>.
- MELONI, Bruno P. et al. The Neuroprotective Efficacy of Cell-Penetrating Peptides TAT, Penetratin, Arg-9, and Pep-1 in Glutamic Acid, Kainic Acid, and In Vitro Ischemia Injury Models Using Primary Cortical Neuronal Cultures. **Cellular and Molecular Neurobiology**, v. 34, n. 2, p. 173–181, Mar. 2014. ISSN 0272-4340. DOI: 10.1007/s10571-013-9999-3. Available from: <<http://link.springer.com/10.1007/s10571-013-9999-3>>.
- MEMARIANI, Hamed; MEMARIANI, Mojtaba. Melittin : a venom-derived peptide with promising anti-viral properties. **European Journal of Clinical Microbiology & Infectious Diseases**, 2019.

MIKITSH, John L.; CHACKO, Ann-Marie. Pathways for Small Molecule Delivery to the Central Nervous System across the Blood-Brain Barrier. **Perspectives in Medicinal Chemistry**, v. 6, pmc.s13384, Jan. 2014. ISSN 1177-391X. DOI: 10.4137/PMC.S13384. Available from: <<http://journals.sagepub.com/doi/10.4137/PMC.S13384>>.

MOREIRA, Igor Matheus Souza. **On Reducing the Dimensionality of Small Molecule Data for Visual-Exploratory Analysis in Human Intestinal Absorption Prediction On Reducing the Dimensionality of Small Molecule Data**. 2022. PhD thesis – Federal University of Pará.

MORIWAKI, Hirotomo et al. Mordred: a molecular descriptor calculator. **Journal of Cheminformatics**, v. 10, n. 1, p. 4, Dec. 2018. ISSN 1758-2946. DOI: 10.1186/s13321-018-0258-y. Available from: <<https://jcheminf.biomedcentral.com/articles/10.1186/s13321-018-0258-y>>.

MOSTAFA, Aya A.; SALEM, Sameh A.; MOHAMED, Amr E. The Effect of Singular-Vectors Feature Selection (SVFS) Based Hyper Dimensionality Framework in Ligand-Based Virtual Screening. *IEEE*, p. 0234–0240, 2022. DOI: 10.1109/ccwc54503.2022.9720796.

MULLARD, Asher. Re-assessing the rule of 5, two decades on. **Nature Reviews Drug Discovery**, v. 17, n. 11, p. 777–777, Nov. 2018. ISSN 1474-1776. DOI: 10.1038/nrd.2018.197. Available from: <<https://www.nature.com/articles/nrd.2018.197>>.

MÜLLER, Andreas C.; GUIDO, Sarah. **Introduction to Machine Learning with Python**. Berkeley, CA: O'Reilly Media, Inc., 2016. ISBN 9781449369415. Available from: <https://link.springer.com/10.1007/978-1-4842-7921-2_5>.

MURRAY, Christopher J L et al. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. **The Lancet**, v. 399, n. 10325, p. 629–655, Feb. 2022. ISSN 01406736. DOI: 10.1016/S0140-6736(21)02724-0. Available from: <<https://linkinghub.elsevier.com/retrieve/pii/S0140673621027240>>.

NAGAOKA, Isao; TAMURA, Hiroshi; REICH, Johannes. Therapeutic Potential of Cathelicidin Peptide LL-37, an Antimicrobial Agent, in a Murine Sepsis Model, 2020. DOI: 10.3390/ijms21175973.

NASSER, Maged et al. Feature Reduction for Molecular Similarity Searching Based on Autoencoder Deep Learning. **Biomolecules**, v. 12, n. 4, p. 508, 2022. ISSN 2218273X. DOI: 10.3390/biom12040508.

NAYARISSERI, Anuraj et al. PDB explorer — A web based algorithm for protein annotation viewer and 3D visualization. **Interdisciplinary Sciences: Computational Life Sciences**, v. 6, n. 4, p. 279–284, Dec. 2014. ISSN 1913-2751. DOI: 10.1007/s12539-012-0044-x. Available from:

<<http://link.springer.com/10.1007/s12539-012-0044-x>>.

OLIVEIRA, André Maurício de. **Introdução à Modelagem Molecular para Química, Engenharia e Biomédicas: Fundamentos e Exercícios**. 1. ed.: Appris, 2018. ISBN 978-85-473-1294-7.

OLIVEIRA, Ewerton Cristhian Lima de et al. Predicting cell-penetrating peptides using machine learning algorithms and navigating in their chemical space. **Scientific Reports**, v. 11, 2021. DOI: s41598-021-87134-w.

OLLER-SALVIA, Benjamí et al. Blood–brain barrier shuttle peptides: an emerging paradigm for brain delivery. **Chemical Society Reviews**, v. 45, n. 17, p. 4690–4707, 2016. ISSN 0306-0012. DOI: 10.1039/C6CS00076B. Available from: <<http://xlink.rsc.org/?DOI=C6CS00076B>>.

OPPER, Manfred; WINTHER, Ole. Gaussian processes for classification: Mean-field algorithms. **Neural Computation**, v. 12, n. 11, p. 2655–2684, 2000. ISSN 08997667. DOI: 10.1162/089976600300014881.

OUELLETTE, Robert J.; RAWN, J. David. **Organic Chemistry**. Elsevier, 2018. ISBN 9780128128381. DOI: 10.1016/C2016-0-04004-4. Available from: <<https://linkinghub.elsevier.com/retrieve/pii/C20160040044>>.

OWOLABI, Mayowa O. et al. Global synergistic actions to improve brain health for human development. **Nature Reviews Neurology**, Springer US, v. 19, n. 6, p. 371–383, June 2023. ISSN 1759-4758. DOI: 10.1038/s41582-023-00808-z. Available from: <<https://www.nature.com/articles/s41582-023-00808-z>>.

PANDEY, Poonam et al. KELM-CPPpred: Kernel Extreme Learning Machine Based Prediction Model for Cell-Penetrating Peptides. **Journal of Proteome Research**, v. 17, n. 9, p. 3214–3222, Sept. 2018. ISSN 1535-3893. DOI: 10.1021/acs.jproteome.8b00322. Available from: <<https://pubs.acs.org/doi/10.1021/acs.jproteome.8b00322>>.

PATEL, Ankur A. **Hands-On Unsupervised Learning Using Python**. O'Reilly Media, 2019. P. 515. ISBN 9781492035640. Available from: <<https://www.oreilly.com/library/view/hands-on-unsupervised-learning/9781492035633/>>.

PIGNATELLO, Rosario. **Drug–biomembrane interaction studies: The application of calorimetric techniques**. Philadelphia: Woodhead Publishing Limited, 2013. ISBN 978-1-908818-34-8.

PLISSON, Fabien; PIGGOTT, Andrew. Predicting Blood–Brain Barrier Permeability of Marine-Derived Kinase Inhibitors Using Ensemble Classifiers Reveals Potential Hits for Neurodegenerative Disorders. **Marine Drugs**, v. 17, n. 2, p. 81, Jan. 2019. ISSN 1660-3397. DOI: 10.3390/md17020081. Available from: <<http://www.mdpi.com/1660-3397/17/2/81>>.

PONNAPPAN, Nisha; CHUGH, Archana. Cell-penetrating and cargo-delivery ability of a spider toxin-derived peptide in mammalian cells. **European Journal of Pharmaceutics and Biopharmaceutics**, v. 114, p. 145–153, May 2017. ISSN 09396411. DOI: 10.1016/j.ejpb.2017.01.012. Available from: <<https://linkinghub.elsevier.com/retrieve/pii/S0939641116308840>>.

POTH, Aaron G. et al. Effects of backbone cyclization on the pharmacokinetics and drug efficiency of the orally active analgesic conotoxin cVc1.1. **Medicine in Drug Discovery**, v. 10, p. 100087, June 2021. ISSN 25900986. DOI: 10.1016/j.medidd.2021.100087. Available from: <<https://linkinghub.elsevier.com/retrieve/pii/S2590098621000087>>.

QIANG, Xiaoli et al. CPPred-FL: a sequence-based predictor for large-scale identification of cell-penetrating peptides by feature representation learning. **Briefings in Bioinformatics**, v. 21, n. 1, p. 11–23, Sept. 2018. ISSN 1467-5463. DOI: 10.1093/bib/bby091. Available from: <<https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bby091/5096827>>.

RADUCANU, B.; DORNAIKA, F. A supervised non-linear dimensionality reduction approach for manifold learning. **Pattern Recognition**, Elsevier, v. 45, n. 6, p. 2432–2444, 2012. ISSN 00313203. DOI: 10.1016/j.patcog.2011.12.006. Available from: <<http://dx.doi.org/10.1016/j.patcog.2011.12.006>>.

RASMUSSEN, C. E.; WILLIAMS, K. I. **Gaussian Processes for Machine Learning**. MIT Press, 2006. v. 7, p. 32–46. ISBN 026218253X.

REDDY, A. Srinivas; KUMAR, Sunil; GARG, Rajni. Hybrid-genetic algorithm based descriptor optimization and QSAR models for predicting the biological activity of Tipranavir analogs for HIV protease inhibition. **Journal of Molecular Graphics and Modelling**, Elsevier Inc., v. 28, n. 8, p. 852–862, June 2010. ISSN 10933263. DOI: 10.1016/j.jmgm.2010.03.005. Available from:

<<http://dx.doi.org/10.1016/j.jmgm.2010.03.005>>
<https://linkinghub.elsevier.com/retrieve/pii/S109332631000046X>.

RÖCKENDORF, Niels; NEHLS, Christian; GUTSMANN, Thomas. Design of Membrane Active Peptides Considering Multi-Objective Optimization for Biomedical Application. **Membranes**, v. 12, n. 2, p. 180, Feb. 2022. ISSN 2077-0375. DOI: 10.3390/membranes12020180. Available from: <<https://www.mdpi.com/2077-0375/12/2/180>>.

ROOHI, Adil et al. Unsupervised Machine Learning in Pathology. **Surgical Pathology Clinics**, v. 13, n. 2, p. 349–358, June 2020. ISSN 18759181. DOI: 10.1016/j.path.2020.01.002. Available from: <<https://linkinghub.elsevier.com/retrieve/pii/S1875918120300040>>.

ROSSI, Michele et al. Sustainable Drug Discovery of Multi-Target-Directed Ligands for Alzheimer's Disease. **Journal of Medicinal Chemistry**, v. 64, n. 8, p. 4972–4990, Apr. 2021. ISSN 0022-2623. DOI: 10.1021/acs.jmedchem.1c00048. Available from: <<https://pubs.acs.org/doi/10.1021/acs.jmedchem.1c00048>>.

RUMELHART, DAVID E. HINTON, Geoffrey E.; WILLIAMS, Ronald J. Learning representations by back-propagating errors. **Nature**, v. 323, pages 533–536, 1986.

RUSSELL, Stuart J.; NORVIG, Peter. **Artificial Intelligence: A Modern Approach**. Pearson, 2021. ISBN 9780134610993.

SALAM, Md. Abdus et al. Antimicrobial Resistance: A Growing Serious Threat for Global Public Health. **Healthcare**, v. 11, n. 13, p. 1946, July 2023. ISSN 2227-9032. DOI: 10.3390/healthcare11131946. Available from: <<https://www.mdpi.com/2227-9032/11/13/1946>>.

SANDERS, William S. et al. Prediction of Cell Penetrating Peptides by Support Vector Machines. Ed. by Sergei L. Kosakovsky Pond. **PLoS Computational Biology**, v. 7, n. 7, e1002101, July 2011. ISSN 1553-7358. DOI: 10.1371/journal.pcbi.1002101. Available from: <<http://dx.plos.org/10.1371/journal.pcbi.1002101>>.

SANTOS, Gabriela B.; GANESAN, A.; EMERY, Flavio S. Oral Administration of Peptide-Based Drugs: Beyond Lipinski's Rule. **ChemMedChem**, v. 11, n. 20, p. 2245–2251, 2016. ISSN 18607187.

SAXENA, Deeksha et al. Blood Brain Barrier Permeability Prediction Using Machine Learning Techniques: An Update. **Current Pharmaceutical Biotechnology**, v. 20, n. 14, p. 1163–1171, Nov. 2019. ISSN 13892010. DOI:

10.2174/1389201020666190821145346. Available from:

<<http://www.eurekaselect.com/174378/article>>.

SIMON, Alexandra et al. Blood-brain barrier permeability study of ginger constituents.

Journal of Pharmaceutical and Biomedical Analysis, v. 177, 2020. ISSN 1873264X. DOI:

10.1016/j.jpba.2019.112820.

SONG, Chen; GROOT, Bert L De; SANSOM, Mark S P. Article Lipid Bilayer Composition

Influences the Activity of the Antimicrobial Peptide Dermcidin Channel. **Biophysj**,

Biophysical Society, v. 116, n. 9, p. 1658–1666, 2019. ISSN 0006-3495. DOI:

10.1016/j.bpj.2019.03.033. Available from:

<<https://doi.org/10.1016/j.bpj.2019.03.033>>.

STALMANS, Sofie; GEVAERT, Bert, et al. Classification of Peptides According to their

Blood-Brain Barrier Influx. **Protein Peptide Letters**, v. 22, n. 9, p. 768–775, Aug. 2015.

ISSN 09298665. DOI: 10.2174/0929866522666150622101223. Available from:

<<http://www.eurekaselect.com/openurl/content.php?genre=article&issn=0929-8665&volume=22&issue=9&spage=768>>.

STALMANS, Sofie; WYNENDAELE, Evelien, et al. Chemical-Functional Diversity in

Cell-Penetrating Peptides. **PLoS ONE**, v. 8, n. 8, 2013. ISSN 19326203. DOI:

10.1371/journal.pone.0071752.

SU, Yongchao et al. Roles of Arginine and Lysine Residues in the Translocation of a

Cell-Penetrating Peptide from 13 C, 31 P, and 19 F Solid-State NMR. **Biochemistry**, v. 48,

n. 21, p. 4587–4595, June 2009. ISSN 0006-2960. DOI: 10.1021/bi900080d. Available

from: <<https://pubs.acs.org/doi/10.1021/bi900080d>>.

SZLASA, Wojciech et al. Lipid composition of the cancer cell membrane. **Journal of**

Bioenergetics and Biomembranes, v. 52, n. 5, p. 321–342, Oct. 2020. ISSN 0145-479X.

DOI: 10.1007/s10863-020-09846-4. Available from:

<<https://link.springer.com/10.1007/s10863-020-09846-4>>.

T.K., Balaji; ANNAVARAPU, Chandra Sekhara Rao; BABLANI, Annushree. Machine

learning algorithms for social media analysis: A survey. **Computer Science Review**, v. 40,

p. 100395, May 2021. ISSN 15740137. DOI: 10.1016/j.cosrev.2021.100395.

Available from: <<https://linkinghub.elsevier.com/retrieve/pii/S1574013721000356>>.

TANG, Hua et al. Prediction of cell-penetrating peptides with feature selection techniques.

Biochemical and Biophysical Research Communications, Elsevier Ltd, v. 477, n. 1,

p. 150–154, Aug. 2016. ISSN 0006291X. DOI: 10.1016/j.bbrc.2016.06.035.

Available from:

<<http://dx.doi.org/10.1016/j.bbrc.2016.06.035>>
<<https://linkinghub.elsevier.com/retrieve/pii/S0006291X16309536>>.

TODESCHINI, Roberto; CONSONNI, Viviana. **Handbook of Molecular Descriptors**. Wiley, Sept. 2000. (Methods and Principles in Medicinal Chemistry). ISBN 9783527299133. DOI: 10.1002/9783527613106. Available from: <<https://onlinelibrary.wiley.com/doi/book/10.1002/9783527613106>>.

USAMA, Muhammad et al. Unsupervised Machine Learning for Networking: Techniques, Applications and Research Challenges. **IEEE Access**, v. 7, p. 65579–65615, 2019. ISSN 2169-3536. DOI: 10.1109/ACCESS.2019.2916648. Available from: <<https://ieeexplore.ieee.org/document/8713992/>>.

VAN DORPE, Sylvia et al. Brainpeps: the blood–brain barrier peptide database. **Brain Structure and Function**, v. 217, n. 3, p. 687–718, July 2012. ISSN 1863-2653. Available from: <<http://link.springer.com/10.1007/s00429-011-0375-0>>.

VEBER, Daniel F. et al. Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. **Journal of Medicinal Chemistry**, v. 45, n. 12, p. 2615–2623, June 2002. ISSN 0022-2623. DOI: 10.1021/jm020017n. Available from: <<https://pubs.acs.org/doi/10.1021/jm020017n>>.

VERLI, Hugo. **Bioinformática: da Biologia à Flexibilidade Molecular**. Ed. by Intergovernmental Panel on Climate Change. Cambridge: Cambridge University Press, Mar. 2014. v. 53, p. 1–30. ISBN 978-85-69288-00-8. DOI: 10.1017/CBO9781107415324.004. arXiv: arXiv:1011.1669v3. Available from: <https://www.cambridge.org/core/product/identifier/CBO9781107415324A009/type/book_part>.

WADE, Corey. **Hands-On Gradient Boosting with XGBoost and scikit-learn: Perform accessible Python machine learning and extreme gradient boosting with Python**. PACKT Publishing LTD, 2020. ISBN 9781839218354.

WEI, Leyi; TANG, Jijun; ZOU, Quan. SkipCPP-Pred: an improved and promising sequence-based predictor for predicting cell-penetrating peptides. **BMC Genomics**, v. 18, S7, p. 742, Oct. 2017. ISSN 1471-2164. DOI: 10.1186/s12864-017-4128-1. Available from: <<https://bmcbgenomics.biomedcentral.com/articles/10.1186/s12864-017-4128-1>>.

WEI, Leyi; XING, PengWei, et al. CPPred-RF: A Sequence-based Predictor for Identifying Cell-Penetrating Peptides and Their Uptake Efficiency. **Journal of Proteome Research**, v. 16, n. 5, p. 2044–2053, May 2017. ISSN 1535-3893. DOI:

10.1021/acs.jproteome.7b00019. Available from:

<<https://pubs.acs.org/doi/10.1021/acs.jproteome.7b00019>>.

YANG, Nicole J.; HINNER, Marlon J. Getting Across the Cell Membrane: An Overview for Small Molecules, Peptides, and Proteins. In: **SITE-SPECIFIC Protein Labeling: Methods and Protocols**. 2015. v. 1266. P. 29–53. ISBN 9781493922727. DOI:

10.1007/978-1-4939-2272-7_3. Available from:

<http://link.springer.com/10.1007/978-1-4939-2272-7_3>.

ZARAGOZÁ, Rosa. Transport of Amino Acids Across the Blood-Brain Barrier. **Frontiers in Physiology**, v. 11, September, p. 1–11, Sept. 2020. ISSN 1664-042X. DOI:

10.3389/fphys.2020.00973. Available from: <<https://www.frontiersin.org/article/10.3389/fphys.2020.00973/full>>.

ZHOU, Hongyu; WANG, Feng; TAO, Peng. T-Distributed Stochastic Neighbor Embedding

Method with the Least Information Loss for Macromolecular Simulations. **Journal of Chemical Theory and Computation**, v. 14, n. 11, p. 5499–5510, 2018. ISSN 15499626.

DOI: 10.1021/acs.jctc.8b00652.

ZHOU, Xue; SMITH, Quentin R.; LIU, Xinli. Brain penetrating peptides and peptide–drug conjugates to overcome the blood–brain barrier and target CNS diseases. **WIREs Nanomedicine and Nanobiotechnology**, v. 13, n. 4, p. 1–34, July 2021. ISSN 1939-5116.

DOI: 10.1002/wnan.1695. Available from:

<<https://onlinelibrary.wiley.com/doi/10.1002/wnan.1695>>.

ZHU, Xiaojin; GOLDBERG, Andrew B. Overview of Semi-Supervised Learning. In: **INTRODUCTION to Semi-Supervised Learning**. Cham: Springer International Publishing, 2009. P. 9–19. ISBN 978-3-031-01548-9. DOI: 10.1007/978-3-031-01548-9_2.

Available from: <https://doi.org/10.1007/978-3-031-01548-9_2>.

ZORBAZ, Tamara et al. Potent 3-Hydroxy-2-Pyridine Aldoxime Reactivators of Organophosphate-Inhibited Cholinesterases with Predicted Blood-Brain Barrier Penetration. **Chemistry - A European Journal**, v. 24, n. 38, p. 9675–9691, July 2018. ISSN 09476539.

DOI: 10.1002/chem.201801394. Available from:

<<https://onlinelibrary.wiley.com/doi/10.1002/chem.201801394>>.

ZOU, Hongliang. Identifying blood-brain barrier peptides by using amino acids physicochemical properties and features fusion method. **Peptide Science**, Oct. 2021. ISSN

2475-8817. DOI: 10.1002/pep2.24247. Available from:
<<https://onlinelibrary.wiley.com/doi/10.1002/pep2.24247>>.

A APPENDIX A

A.1 Training dataset of CPPs and non-CPPs

Sequences of CPPs and non-CPPs used in the training dataset with their origin (CPP-Site 2, C2Pred, or DB1*) and their modelling origin: experimental (EXPR) or computational (COMP).

Sequence (CPP)	Reference	Origin	Model	Sequence (non-CPP)	Reference	Origin	Model
KETWWTWTEWSQPKRRKV	pep-1	DB1	COMP	TERQIKIWFQNRMRK	pAntp41-55	DB1	COMP
TRSSRAGLQFPVGRVHRLRK	Buforin	DB1	COMP	AHALCLTERQIKSNRRMKWKEN	pAntpHD 48S	DB1	COMP
TAKTRYKARRAELIAERR	Phi21 N (12-29)	DB1	COMP	FTKALGISYGRKKRRQC	ptat7	DB1	COMP
KFTTFPQTGIVGAP	hCT18-32	DB1	COMP	ILRRRIRKQAHASHK	pVEC(4-18)	DB1	COMP
TRQARRNRWRERQR	HIV-1 rev	DB1	COMP	AGCKNFFWKTFTSC	Somatostatin 14	DB1	COMP
ALWKTLLKKVLKA	K4-S4(1-13)a	DB1	COMP	GWTLNSAGYLLGPHAI	Galanin (1-16)	DB1	COMP
LLILRRIRKQAHASHK	pVEC	DB1	COMP	DFDMLRCMLGRVYPCWQV	HCM	DB1	COMP
KLALKALKALKA	MAP	DB1	COMP	WSYGLRPG	[1]	DB1	COMP
AAVALLPAVLLALLAKNNLKEGLY	[1]	DB1	COMP	KKKQYTSIHGGVVEVD	[1]	DB1	COMP
KMTRAQRRAAARRNRWTAR	[1]	DB1	COMP	GWTLNSAGYLLGPPGFSFPR	[1]	DB1	COMP
LLILRRIRKQAHASHK	[1]	DB1	COMP	PVVHLTLRQAGDDFSR	[1]	DB1	COMP
LLILRRIRKQAHASHK	[1]	DB1	COMP	EILLPNYNAVESYKYPMFIALSK	[1]	DB1	COMP
SWLGRQLRIAGKRLEGRSK	[1]	DB1	COMP	QNLGNQWAVGHLM	[1]	DB1	COMP
GAARVTSWLGRLRIAGKRLEGRSK	[1]	DB1	COMP	VPLPAGGGTVLNQDVPARQPLGG	[3]	C2Pred	COMP
AAVALLPAVLLALLAPVQRKQKLMF	[1]	DB1	COMP	AATAATPATAATPATAARA	[3]	C2Pred	COMP
RQGAARVTSWLGRLRIAGKRLEGR	[1]	DB1	COMP	IIGAAIAALPHVINAIKNTFG	[3]	C2Pred	COMP
GYGNCRXFKQKPRRD	[2]	CPPsite 2	COMP (CPPSite ID: 1266)	PSCVCSGFETSGIHFC	[3]	C2Pred	COMP
IGCRX	[2]	CPPsite 2	COMP (CPPSite ID: 1642)	SCIKHGFDCFDGDDDCQCCRDNGF	[3]	C2Pred	COMP
YGRKKRRQRRRTALDWSWLQTE	[3]	C2Pred	COMP	YQLQJELCCQHL	[3]	C2Pred	COMP
GRKGKHKRKKLP	[3]	C2Pred	COMP	IVQQCTSGICSLYQENYCN	[3]	C2Pred	COMP
KFLNRFVHWWLQLKPGQPMY	[3]	C2Pred	COMP	GIACGESCVFLGCFIPGCSCKSKVCYFN	[3]	C2Pred	COMP
RRRRRRRRRGGPGVTWTPQAWFQWV	[3]	C2Pred	COMP	HGVSGHGQGHVGH	[3]	C2Pred	COMP
AEKVDPVKLNLTLSAAAEALTGLGDK	[3]	C2Pred	COMP	PKVVPNGVQEETSEGFPLEF	[3]	C2Pred	COMP
WIIFKIAASXKK	[2]	CPPsite 2	COMP (CPPSite ID: 1622)	NPRWEFRGKFGVGR	[3]	C2Pred	COMP
CXXRRRRXXC	[2]	CPPsite 2	COMP (CPPSite ID: 2125)	LYISRQGFRA	[3]	C2Pred	COMP
GLKKLARLPHKKLKLGC	[3]	C2Pred	COMP	GSSGMIPFRV	[3]	C2Pred	COMP
VVLGKLYGRKKRRQRRR	[2]	CPPsite 2	COMP	GWKSVFRKAKKVGKTVGGLADHYLG	[3]	C2Pred	COMP
TSPLNIHNGQKL	[3]	C2Pred	COMP	LQQGSFRPSQQN	[3]	C2Pred	COMP
PSKRLLXNNLRR	[2]	CPPsite 2	COMP (CPPSite ID: 1655)	WLSKTAKKLENSAKKRISGEIAIAIKGGSR	[3]	C2Pred	COMP
NYTTYKSHFQDR	[3]	C2Pred	COMP	GVLSNVIGYLKGLTGALNAVLKQ	[3]	C2Pred	COMP
RKKRRQR	Tat (49-55)	CPPsite 2	COMP (CPPSite ID: 1008)	SFHVPPPWCMCKSLKCC	[3]	C2Pred	COMP
LIIFAASXKK	[2]	CPPsite 2	COMP (CPPSite ID: 1631)	GLLSKVLGVGKKVLCVSGSLC	[3]	C2Pred	COMP
CGGKDCERRFSRSDQLKRXQRRXTGVKPFQ	b-WT1-pTJ	CPPsite 2	COMP (CPPSite ID: 2303)	CCSQDCLVCPCCPN	[3]	C2Pred	COMP
MIYRIAASHKK	[3]	C2Pred	COMP	QATVGDVNTDRPGLLDLK	[3]	C2Pred	COMP
RRQRRTSKLMKR	[3]	C2Pred	COMP	CGECTVGTCTYTPGCACDWPVCKRD	[3]	C2Pred	COMP
LILGRRRRRRRRCG	LILIR8 (Alexa)	CPPsite 2	COMP (CPPSite ID: 2691)	LKLKDILGKIKVILSHLNK	[3]	C2Pred	COMP
CRQIKWFPNRRMKWKCC	Reduced linear penetratin	CPPsite 2	COMP (CPPSite ID: 1131)	AFDHYGFTGGL	[3]	C2Pred	COMP
CRWRWSSKK	Crot (27-39) dervative	CPPsite 2	COMP (CPPSite ID: 1167)	CKSKGAKCSKLMYDCSCGSCGTVGRC	[3]	C2Pred	COMP
SWAQHLSPPVL	[3]	C2Pred	COMP	SENPSNSRNFRIL	[3]	C2Pred	COMP
GRQLRIAGRRRLRRSR	[3]	C2Pred	COMP	KPNPERFYAPM	[3]	C2Pred	COMP
LGYTYQDFNFKFTFPQTGIVGAP	EGFP-hcT(9-32)	CPPsite 2	COMP (CPPSite ID: 2226)	GSLTGLISMPRT	[3]	C2Pred	COMP
QWQRNMRRKVR	M6	CPPsite 2	COMP (CPPSite ID: 1413)	PDERRQLNKIFLWDFCNSDSI	[3]	C2Pred	COMP
APWXLSSQYSRT	CTP	CPPsite 2	COMP (CPPSite ID: 2588)	CKCVQCESCTPCC	[3]	C2Pred	COMP
GLLEALAELEGLRRLRKRFRNKIKEK	[3]	C2Pred	COMP	DLWNSIKDMAAAGRAALNAVGMVNQ	[3]	C2Pred	COMP
AAVALLPAVLLALLAK	MPS	CPPsite 2	COMP (CPPSite ID: 1791)	MPSPGLRLPLLLPLPWLLVLT	[3]	C2Pred	COMP
NYQWRCKNQ	ECP(32-41)R3Q	CPPsite 2	COMP (CPPSite ID: 2024)	FLKSLLGPLIDLSKG	[3]	C2Pred	COMP
KFTTFPQTGIVGAP	hCT(18-32)	CPPsite 2	COMP (CPPSite ID: 1461)	LKISQYQKFAWPQYL	[3]	C2Pred	COMP
XRLRXALAXLLXKLKXLLXALXRLRX	[2]	CPPsite 2	COMP	ISCQDVQKSLAPCLPYVTGRAPKA	[3]	C2Pred	COMP
KCRKKRRRQRRKKPVVHLTLRQAGDDFSR	[2]	C2Pred	COMP	RCRGEGGFCGLTYEERCSCGWCFFVCV	[3]	C2Pred	COMP
AGYLLGXINLXALXLLXILC	TH peptide	CPPsite 2	COMP (CPPSite ID: 2122)	GSSGLIPFGRT	[3]	C2Pred	COMP
KRIPNKKPGKTKTKPTKPTIKTKK	[2]	CPPsite 2	COMP	EEMGFAKCCAGCSTEDFRMVC	[3]	C2Pred	COMP
ANIXPLLPIC	[2]	CPPsite 2	COMP	RKYVMGHFRWRDRGRNRSSSSGSGAGQKR	[3]	C2Pred	COMP
LNSAGYLLGKLKALAAALAK	[2]	CPPsite 2	COMP	SPANAQITRRHKINSFVGLM	[3]	C2Pred	COMP
EEEEAAKKK	[2]	CPPsite 2	COMP	AKWKEDVIKLCRELVRTQIAICG	[3]	C2Pred	COMP
RIKAERKMRNRNIAASKSRKRKLRIARGC	[3]	C2Pred	COMP	QYPLQGGSFRPS	[3]	C2Pred	COMP
VLGQSGYLMPMR	[3]	C2Pred	COMP	FLPFLAKILTGVL	[3]	C2Pred	COMP
RRKLQSQKEKK	[3]	C2Pred	COMP	KPSPDRLFYGLM	[3]	C2Pred	COMP
GSRVQIRCFRNRSTR	[3]	C2Pred	COMP	SPFPFPPGICKRLKRC	[3]	C2Pred	COMP
YWLKLLKKWLKLWKKLLKLW	[2]	CPPsite 2	COMP	YGGFLRRQFKVTVTSQEDPNAYSGELFDA	[3]	C2Pred	COMP
KLPCRSNTFLNIFRRKKPG	[2]	C2Pred	COMP	GGKFLKAKKGIGAVLKVLTTLG	[3]	C2Pred	COMP
CGRKKRRQRRARPPO	[2]	CPPsite 2	COMP	HSEGTFSNDYSKYLETRRAQDFVQWLKNS	[3]	C2Pred	COMP
ACRGRGRCGRGRGRCG	[2]	CPPsite 2	COMP	AAEFDFDYDSEEQMGPHEA	[3]	C2Pred	COMP
RLWMRWYSPITTRAG	[2]	CPPsite 2	COMP	FVSRHLCSNLVETLYSVCQDDGFFYIPKD	[3]	C2Pred	COMP
GKKRRKLSNRESAKRSR	[3]	C2Pred	COMP	FLPMLAKLLSGFLGK	[3]	C2Pred	COMP
AAVALLPAVLLALLAPSGASGLDKRDYV	[3]	C2Pred	COMP	GKCGEINGSCECYGGSVTCDCY	[3]	C2Pred	COMP
GWTLNSAGYLLGKLNKAPAAALAKKIL	[2]	CPPsite 2	COMP	HLPPPVHLPPPV	[3]	C2Pred	COMP
ACRGRGRCGRGRGRCG	[2]	CPPsite 2	COMP	GLVSSIGRALGGLLADVVKSKGQPA	[3]	C2Pred	COMP
LIIFRIASXKK	[2]	CPPsite 2	COMP	QFPFPQPPQFPQSQ	[3]	C2Pred	COMP
NYTTYKSXFQDR	[2]	CPPsite 2	COMP	CCPGWELCCEWDDGW	[3]	C2Pred	COMP
RKKRRQAR	[2]	CPPsite 2	COMP	NCPCYCVVYCCPAYCEASGRPP	[3]	C2Pred	COMP
LNSAGYLLGKALAAALAKKIL	[2]	CPPsite 2	COMP	DYDPRTEAPRRLPADDDVEDGEDRV	[3]	C2Pred	COMP
PPKSAQCLRYKKPE	[3]	C2Pred	COMP	FLPKMSTKLVRYPYRRGTDYH	[3]	C2Pred	COMP
MIYRDKSX	[2]	CPPsite 2	COMP	FLPLLAGVNAVFLPQICKIARKC	[3]	C2Pred	COMP
FFLPKGRRRRRRRRCG	[3]	C2Pred	COMP	GLPVCGETCFGGRCNTPGCTCSYPICTRN	[3]	C2Pred	COMP
KXKLLXLLXLLALLWLXLLXLLKXK	[2]	CPPsite 2	COMP	FFGHLFLKATKIIPSLFO	[3]	C2Pred	COMP
VKRFFKFFRKLKLV	[2]	CPPsite 2	COMP	GFFGKMKEYFKKFGASFKRRFANLKKRL	[3]	C2Pred	COMP
RRRRRRRRRGD	[2]	CPPsite 2	COMP	TCTLTGTCYTAGSCSWPVCTRNGVPICGE	[3]	C2Pred	COMP
KALKKLALALLAKLKL	[3]	C2Pred	COMP	GLFLDTLKGLAGKLLQGLKCIKAGCKP	[3]	C2Pred	COMP
ISFDELLDYGESGS	[3]	C2Pred	COMP	ALCCYGYRFFCCPNFR	[3]	C2Pred	COMP
RGDFK	[2]	CPPsite 2	COMP	GLTPNMNSLFF	[3]	C2Pred	COMP
KLWMRWYSPWTRRYG	[2]	CPPsite 2	COMP	FLGALIKGAIHGGRFHGMQIHNHG	[3]	C2Pred	COMP

NAKTRRXERRRKLAIERGC	[2]	CPPsite 2	COMP	FLPVLAGIAAKVVPALFCKITKKC	[3]	C2Pred	COMP
KKDGKKRRRSRKESYSVYVYKVLKQ	[3]	C2Pred	COMP	GRRRKRWLRIRIGKVKIIGGAALDHL	[3]	C2Pred	COMP
ACRRSRRCGRRSRRCG	[2]	CPPsite 2	COMP	NGVYCTKNKCTVDWAKATTIAGMSIGGF	[3]	C2Pred	COMP
HIQLSPFSQSWR	[3]	C2Pred	COMP	GLFDIIKKVASVVGASQ	[3]	C2Pred	COMP
MIIFKIAASXKK	[2]	CPPsite 2	COMP	GHDFDQDDVNSSGEKDESLVRI	[3]	C2Pred	COMP
TARRITPKDVIDVRSVTTEINT	[2]	CPPsite 2	COMP	TPVVNPFFLQQT	[3]	C2Pred	COMP
GLGSLLKKAGKKLKQPKSKRKV	[3]	C2Pred	COMP	LNLKALLAVAKKIL	[3]	C2Pred	COMP
TRQARRNRRRRWREQR	Rev (34-50)	CPPsite 2	EXPR (PDB code: 1RPV)	EPTWNNLKGMW	[3]	C2Pred	COMP
WEARLARALARALARXALARALARA	[2]	CPPsite 2	COMP	AEPGADDAEEVEQKQLQ	[3]	C2Pred	COMP
ASMWERVKSIKSSLAASNI	[3]	C2Pred	COMP	GLGDILGLLGL	[3]	C2Pred	COMP
RIRMIQNLIKKT	[3]	C2Pred	COMP	IPPYCTIAPFI	[3]	C2Pred	COMP
QIISRDLISX	[2]	CPPsite 2	COMP	QQPFVQQQPFVQQ	[3]	C2Pred	COMP
QAASRVENYMR	[3]	C2Pred	COMP	GPYGGGGLVGALLG	[3]	C2Pred	COMP
GLKKLAELFXKLLKLG	[2]	CPPsite 2	COMP	GCCPFACHTHTICRCC	[3]	C2Pred	COMP
LNSAGYLLGKINLKALAALAKKIL	[2]	CPPsite 2	COMP	MSPRPLAWALVL	[3]	C2Pred	COMP
GGAYVTRSSAVRLRSSVPGVRLQ	[3]	C2Pred	COMP	EYDDMYTEKRPKYAFGL	[3]	C2Pred	COMP
KWCFRVCYRGICYRRRCRGK	[3]	C2Pred	COMP	GGCRIGPITWVCGG	[3]	C2Pred	COMP
RRIRPRP	[2]	CPPsite 2	COMP	GLFTLIKCAYLIAPTVACN	[3]	C2Pred	COMP
RRARRPRRLRPAPGR	[3]	C2Pred	COMP	FIGAVAGLLSKIF	[3]	C2Pred	COMP
RKRRRRRESRKKRRRES	[2]	CPPsite 2	COMP	ALWKTLLKGAGKVFGHVAKQFLGSQQQPES	[3]	C2Pred	COMP
YPRAARRARR	[2]	CPPsite 2	COMP	MAASPRNSVLLA	[3]	C2Pred	COMP
KKKEERADLIAYLKKA	[2]	CPPsite 2	COMP	NGTGPQHLGCSHLVDALYLVCGPTGFFYNP	[3]	C2Pred	COMP
ACRGRGRGCRGRGCGC	[2]	CPPsite 2	COMP	DCHNTQLPFIYKTCPEGCNL	[3]	C2Pred	COMP
RRRQRRKKR	[2]	CPPsite 2	COMP	FIPLVSGLESRL	[3]	C2Pred	COMP
LIRLWSXLIXIWFQNRRLKWKKGCGC	[2]	CPPsite 2	COMP	CAETCIYPCFTEAVGCKCKDKVCYKN	[3]	C2Pred	COMP
KLALKAALKAWKAAAKLA	[2]	CPPsite 2	COMP	KSDLLGALLSRNPSYGLPSRDMSTAY	[3]	C2Pred	COMP
NKPLVIFY	[2]	CPPsite 2	COMP	VYVPRYIANLY	[3]	C2Pred	COMP
RLXRLXRLXRLXRLX	[2]	CPPsite 2	COMP	FKVQNHQGVVVKIFHH	[3]	C2Pred	COMP
RRLXRLXXYRRRWXRFR	[2]	CPPsite 2	COMP	GLVPNLLNLLGL	[3]	C2Pred	COMP
ERKKRRRE	[2]	CPPsite 2	COMP	MTPPPLPARVDFSLAGALN	[3]	C2Pred	COMP
DRDRDRDR	[2]	CPPsite 2	COMP	KIHFAQTQSLVYP	[3]	C2Pred	COMP
KKLALXALXLLALLWLXALXALKK	[2]	CPPsite 2	COMP	SLADTQSGHRW	[3]	C2Pred	COMP
MAMPEPRRANVMAHKLEPASLQLRSCA	[3]	C2Pred	COMP	GLWNKIEAASKAAGKAALGFVNEMVG	[3]	C2Pred	COMP
ALIILRRIRKQAXAXSK	[2]	CPPsite 2	COMP	YIKVLRCRVVFQNEC	[3]	C2Pred	COMP
GSPWGLQXXPPRT	[2]	CPPsite 2	COMP	FFPNVAVPGQVLLKKIFCAISKCC	[3]	C2Pred	COMP
LLYWRRRRRXRRRRXRR	[2]	CPPsite 2	COMP	GCIGNESQKKDNVYKFKE	[3]	C2Pred	COMP
KKAAQIRSQVMTXLRI	[2]	CPPsite 2	COMP	MRKEFHNVLSSGQLLADKRPARDYNRK	[3]	C2Pred	COMP
GKYVSLTTPKNPTKRRIPTKDV	[3]	C2Pred	COMP	FMGGLIKAAKIVPAAYCAITKKC	[3]	C2Pred	COMP
ARCSGSGSGCGSGSGCGR	[2]	CPPsite 2	COMP	VIVKAIATLSKKLL	[3]	C2Pred	COMP
CSSLDEPGRGGFSSESKV	[3]	C2Pred	COMP	ILGLLKGISALLS	[3]	C2Pred	COMP
ACSSSPSKXCGGGRRRRRRRRR	[2]	CPPsite 2	COMP	PGSAICNMACRLEHGHLYPFCNCD	[3]	C2Pred	COMP
RRWFRRWRR	[2]	CPPsite 2	COMP	GIVEQCCTSICSLYQLENYCN	[3]	C2Pred	COMP
RRVWRRYRRQRWCRR	[3]	C2Pred	COMP	AIIEWEGIESGSVEQA	[3]	C2Pred	COMP
VKLPPP	[2]	CPPsite 2	COMP	RCCKFPCPDSCRYLCC	[3]	C2Pred	COMP
NNNAAGRRKKRT	[3]	C2Pred	COMP	FIGPLISALASLFKG	[3]	C2Pred	COMP
TPKTMQTQYDFS	[3]	C2Pred	COMP	MQKLQISVYIYLFMLIVAGPVDLNENSEQK	[3]	C2Pred	COMP
XRRRRRRR	[2]	CPPsite 2	COMP	QGTINIVCECCMKPCTLSLRQYCP	[3]	C2Pred	COMP
RILQQLLFIKFRIGCRXSRI	[2]	CPPsite 2	COMP	SCIALTLVLVANSAPTSSSTKETQQOLE	[3]	C2Pred	COMP
PRPRLPFRPG	[2]	CPPsite 2	COMP	MHSSALLCCLVLT	[3]	C2Pred	COMP
ROIKIWFQNMWKWK	[2]	CPPsite 2	COMP	FIGALLGPLLNLK	[3]	C2Pred	COMP
GLKKLARLAXKLLKLG	[2]	CPPsite 2	COMP	SAGATANPLRS	[3]	C2Pred	COMP
GKRRRRATAKYRSAH	[2]	C2Pred	COMP	VQQQPLGQQQP	[3]	C2Pred	COMP
SRWRWKSSKK	[2]	CPPsite 2	COMP	VVCNRYRDVRFESIRLPGGPRGVNPVSY	[3]	C2Pred	COMP
RKLTTIFPLNWKYRKALSLG	[3]	C2Pred	COMP	QGRLTQWAVGHLM	[3]	C2Pred	COMP
RRRRNRTRNRNRVRGC	[3]	C2Pred	COMP	PACGGFWISGRPG	[3]	C2Pred	COMP
AGYLLGKLKALAALAKKIL	[2]	CPPsite 2	COMP	CLGSGEQCVRDTSCCSMCTNNICF	[3]	C2Pred	COMP
RXVYXVLLSQ	[2]	CPPsite 2	COMP	SYCGSTTRICCGYCAFGKCKIDYPSN	[3]	C2Pred	COMP
LLIALRRIRKQAXAXSK	[2]	CPPsite 2	COMP	DCCPAKLLCCNP	[3]	C2Pred	COMP
ROIKIWFQNMWKWKLRKKKKKH	[3]	C2Pred	COMP	LPYPVNCKTECECVMCGLGIICKQCYYYQ	[3]	C2Pred	COMP
DTWAGVEAIRILQQLLFIKFR	[2]	CPPsite 2	COMP	IVAVLFLTACQFNAADSRVRRNAEH	[3]	C2Pred	COMP
VSRRRRRRGGRRRRK	[2]	CPPsite 2	COMP	GFMDTAKNAVKNVAVTLIDKLCKVTGGC	[3]	C2Pred	COMP
LKLAELAXKLLKLG	[2]	CPPsite 2	COMP	SDLTWYQSPGDPNTSNK	[3]	C2Pred	COMP
GRGDSPPRRSPRR	[3]	C2Pred	COMP	GWMSKIASGIGTFLSGIGQQG	[3]	C2Pred	COMP
WLRRKAWLRRKALNRQLGVAA	[2]	CPPsite 2	COMP	APESPKRAPSGFLGVR	[3]	C2Pred	COMP
YGRRRRRRRR	[2]	CPPsite 2	COMP	KKAVRRQEAVDAL	[3]	C2Pred	COMP
YPYDANHTRSP	[3]	C2Pred	COMP	GWKDWLNKGKEWLKKKGPGIMKAALKAATQ	[3]	C2Pred	COMP
GALFLGFLGAAGSTMGAWSQPKKKRKV	[2]	CPPsite 2	COMP	MKLNNGKSLDPTGLY	[3]	C2Pred	COMP
CGGKDCERRFSRSDQLKRHRRTGVKPFQ	[3]	C2Pred	COMP	DDASDRAKKFYGLM	[3]	C2Pred	COMP
GWTLNSAGYLLGPXAVGNXRSFSDKNGLTS	[2]	CPPsite 2	COMP	SSPETLISDLLMRESTENVPRTREDPAMW	[3]	C2Pred	COMP
SWLPYPWXPSS	[2]	CPPsite 2	COMP	YGGFMKPYTKQSHKPLITLLKHTLKNEQ	[3]	C2Pred	COMP
ACRGRGRRCGSGSRSCG	[2]	CPPsite 2	COMP	TCCKFOFLNFCNE	[3]	C2Pred	COMP
WIIFRIAYXKK	[2]	CPPsite 2	COMP	SAATANVHRCCLTGTCTQDQLGLCPH	[3]	C2Pred	COMP
ARRRRCSDRFRNCPADEALCGRRRR	[2]	CPPsite 2	COMP	VLSHNNESSYSDTSSCTSQ	[3]	C2Pred	COMP
GRQLRIAGKRLEGRSK	[2]	CPPsite 2	COMP	VMMVEAGFGTHGCPLLQGTCDSHCRGMDA	[3]	C2Pred	COMP
RLLRLLRLWRLLRLLR	[3]	C2Pred	COMP	AADHDVGSELPPGVLGALLRV	[3]	C2Pred	COMP
KLGV	[2]	CPPsite 2	COMP	RMTLSEKCCQVGCIRKDIARLC	[3]	C2Pred	COMP
KKTTTKPTKK	[2]	CPPsite 2	COMP	KKDGYPVEYDRAY	[3]	C2Pred	COMP
CTWLKYY	[2]	CPPsite 2	COMP	FLKPLFNAALKLLP	[3]	C2Pred	COMP
KKHLLHLLHLLALLWLHLLHLLKHK	[3]	C2Pred	COMP	GIGAILKVLATGLPTLISWIKNRKQ	[3]	C2Pred	COMP
PSKRLHNNLR	[3]	C2Pred	COMP	ARHPHPLSFM	[3]	C2Pred	COMP
MAARLCCQLDPADV	[2]	CPPsite 2	COMP	GRGARRYCGRVLADTLAYLCPEMEEVE	[3]	C2Pred	COMP
RGERLERRELRLERELRC	[2]	CPPsite 2	COMP	CKGKGAPECTRLMYDCCHGSCSSSKGRC	[3]	C2Pred	COMP
KRIIQRILSRNS	[3]	C2Pred	COMP	EEESRPRKLCGRHLLIEVIKLCGQSDWS	[3]	C2Pred	COMP
RLLRLLRLL	[2]	CPPsite 2	COMP	PPPPGGPQPRPPQG	[3]	C2Pred	COMP
RSVTTEINTLQTLTISIAEKVDP	[3]	C2Pred	COMP	RCTCTIISSSSTF	[3]	C2Pred	COMP
SMLKRNXTSNR	[2]	CPPsite 2	COMP	KPKCGLCRYRCCSGGSGKCVNGACDCS	[3]	C2Pred	COMP
RQPKIWFPPNRRMPWK	[3]	C2Pred	COMP	GIMDSVKGAKNLAGKLLDSLCKKITGC	[3]	C2Pred	COMP
ROIKIWAQNRRMKWK	[2]	CPPsite 2	COMP	GLFLDTLKKFAKAGMEAVINPK	[3]	C2Pred	COMP
ARCSDRFRNCPADEALCGR	[2]	CPPsite 2	COMP	DEATVFLWPLCSYRMLPF	[3]	C2Pred	COMP
VELPPPVELPPPVELPPP	[3]	C2Pred	COMP	RSNKGFFNMVMDIQAISK	[3]	C2Pred	COMP
MDAQTRRRERRAEKQAWKAANGC	[3]	C2Pred	COMP	GLISGLLVGKMLVCGLSGLC	[3]	C2Pred	COMP
HPGSPFPEHRP	[3]	C2Pred	COMP	LETAPQVQPARILLP	[3]	C2Pred	COMP

PPXNRIQRRLLNM	[2]	CPPsite 2	COMP	LLGMIPLAISALSLSKL	[3]	C2Pred	COMP
GPFXFYQFLFPPV	[2]	CPPsite 2	COMP	DGKLYKMTFRWSEGS	[3]	C2Pred	COMP
GDXLPLXKLC	[2]	CPPsite 2	COMP	AKKELCTCQPKHLKYIEKGLQKAKDYAT	[3]	C2Pred	COMP
RXNXRFNFRFFNFNFRNTRTN	[2]	CPPsite 2	COMP	DSLFSYNNFEEDD	[3]	C2Pred	COMP
LIRLWSXLIXIWFQNRRLKWKK	[2]	CPPsite 2	COMP	LAKRADICQPGKTSQRACET	[3]	C2Pred	COMP
KKKKKKKKLQQRGD	[3]	C2Pred	COMP	GFFTLIKAANKLINKTVNKEAGKGLEIMA	[3]	C2Pred	COMP
LIIFAILISXXX	[2]	CPPsite 2	COMP	KCCMRPICMPCPCIGAG	[3]	C2Pred	COMP
INLKKLAKLXKKIL	[2]	CPPsite 2	COMP	QPSYDRDIMSFG	[3]	C2Pred	COMP
GIGKFLXSAKKWGKAFVQGIMNC	[2]	CPPsite 2	COMP	GLFDIVKKVVGITAGLG	[3]	C2Pred	COMP
GKINLKALAAKAKIL	[2]	CPPsite 2	COMP	IRDECCSNPACRVNNPHVC	[3]	C2Pred	COMP
RRRRRRRGGIYALAKWALKQ	[2]	CPPsite 2	COMP	FFPMLAGVAARVVPKVICLITKKC	[3]	C2Pred	COMP
IPSRWKDQFWKRWXY	[2]	CPPsite 2	COMP	FLPFIAGMAANFLPKIFCAISKCC	[3]	C2Pred	COMP
FITKALGISYGRKKRR	[2]	CPPsite 2	COMP	WEAKLAKALAKAKHLAKALAKALCEA	[3]	C2Pred	COMP
MAIYRDLIS	[2]	CPPsite 2	COMP	GFFDRIKALTKNVTELELNTITCKLPVTPP	[3]	C2Pred	COMP
EEEEA	[2]	CPPsite 2	COMP	AVLDFIKAAGKGLVTNMEKVG	[3]	C2Pred	COMP
DRRRRRSRPSGAERRRRRAAAA	[3]	C2Pred	COMP	GMATKAGTALGKAVAVIGAAL	[3]	C2Pred	COMP
RKKARQRRR	[2]	CPPsite 2	COMP	WQIPEQSQCCAI	[3]	C2Pred	COMP
KMDSRWRWKSSKK	[2]	CPPsite 2	COMP	YLRNPRKNLLKNILADVLRQLQKK	[3]	C2Pred	COMP
SHAFTWPTYLQL	[3]	C2Pred	COMP	YVSQRLCGSQLVDITLVSVCRRHGFYRPN	[3]	C2Pred	COMP
RQIKIWFQNRMMKWAK	[2]	CPPsite 2	COMP	AGETHTVMINHAGRGAPKLVVGGKKLS	[3]	C2Pred	COMP
EKGKKIIFMK	[2]	CPPsite 2	COMP	AFDSLGSFGDFNGFN	[3]	C2Pred	COMP
MDCRWRWKCKCK	[2]	CPPsite 2	COMP	LEEELEELEGCE	[3]	C2Pred	COMP
GRKKRRQPPQC	[2]	CPPsite 2	COMP	YCNGKRVVCVRG	[3]	C2Pred	COMP
VPMIK	[2]	CPPsite 2	COMP	NSELINSLGIPKVMTDA	[3]	C2Pred	COMP
RQIKIAFQNRMMKWKK	[2]	CPPsite 2	COMP	CFKKDMHKVETYL	[3]	C2Pred	COMP
EEEEEEEEPLAGRRRRRRRRN	[3]	C2Pred	COMP	FLGVVFKLASKVFPVAFGKV	[3]	C2Pred	COMP
FAPWDTASFMLG	[3]	C2Pred	COMP	INLKAIAALARNY	[3]	C2Pred	COMP
XYRIKPTFRRLKWYKGFKA	[2]	CPPsite 2	COMP	RGPDRHFAFGL	[3]	C2Pred	COMP
PQNRLQIRRXSK	[2]	CPPsite 2	COMP	HADGRYTSDISSYLEGQAACEFIWLVN	[3]	C2Pred	COMP
MYKSKGSWILVLFVAMWSDVGLCKRKP	[3]	C2Pred	COMP	FFPLLFGALSSHLPKLF	[3]	C2Pred	COMP
CRKKRRQRRR	[2]	CPPsite 2	COMP	GRVRDQIMLSLGG	[3]	C2Pred	COMP
GWTLNPPGYLLGKINLKALAAKAKIL	[2]	CPPsite 2	COMP	HIGPNPVYSAVSNTD	[3]	C2Pred	COMP
KCFQWQRNMKVR	[2]	CPPsite 2	COMP	DTNFPICLFCCKCKNSSCGLCIT	[3]	C2Pred	COMP
RSRGLRRGAIRLQRG	[2]	CPPsite 2	COMP	FPPPGESAVDMSFFYALSNP	[3]	C2Pred	COMP
KGKKIFIMK	[2]	CPPsite 2	COMP	YGGFIGIRKSARKWNNQ	[3]	C2Pred	COMP
GRKKRRQRRRP	[2]	CPPsite 2	COMP	DVDFNSESTRRKNKQKEIVDLHNSLKKT	[3]	C2Pred	COMP
LLKTTELLKTTELLKTTE	[3]	C2Pred	COMP	ALWKDMLSGIGKLAGQAALGAVKTLV	[3]	C2Pred	COMP
VRLPPP	[2]	CPPsite 2	COMP	RKFHEKHSHSRGYRSNYLYDN	[3]	C2Pred	COMP
RRRRRRRRXXX	[2]	CPPsite 2	COMP	YSSQHLGCSNLVEALYMTGCRSGFYRPHD	[3]	C2Pred	COMP
RKKRRQRRRGGKLLKLLKLLKLLK	[3]	C2Pred	COMP	GIFSTVFKAGKGIVCGLTGLC	[3]	C2Pred	COMP
MIYRDLISKK	[3]	C2Pred	COMP	RRKMCGEALIQALDVICVNGFT	[3]	C2Pred	COMP
KKWKMRRGAGRRRRRRRRR	[2]	CPPsite 2	COMP	RPWCHPINAILAVEKVCTYRDVRFESIRL	[3]	C2Pred	COMP
KALAKALAKLWALAKAA	[2]	CPPsite 2	COMP	ACAAHCLLRGNRGYCNKGK	[3]	C2Pred	COMP
GWTLNSKINLKALAAKAKIL	[2]	CPPsite 2	COMP	ALNSVAYERSVMQDYE	[3]	C2Pred	COMP
RQIKIWFQNRRAKWK	[2]	CPPsite 2	COMP	ILGTILGLLKL	[3]	C2Pred	COMP
ACSGSGSGSGSGSGSGRRRRRRRR	[2]	CPPsite 2	COMP	SPVDYDRPIMAFG	[3]	C2Pred	COMP
ARRRCSGSGSGSGSGSGCRRR	[2]	CPPsite 2	COMP	GLLGLGLLGPLLGGGGGGGGGLL	[3]	C2Pred	COMP
RRGC	Oligoarginine R2	CPPsite 2	EXPR (PDB code: 3C88)	FLPLGLIAGIAKML	[3]	C2Pred	COMP
MXKRPTPSRK	[2]	CPPsite 2	COMP	ITCQQVTSSELGPCVPYLTGGQIP	[3]	C2Pred	COMP
LLRILRRSIRARRAIRR	[3]	C2Pred	COMP	RVCFAIPLPICH	[3]	C2Pred	COMP
KRIPNKKPGKKT	[2]	CPPsite 2	COMP	MGMRLPNIIFL	[3]	C2Pred	COMP
GSRXPSLIIPRQ	[2]	CPPsite 2	COMP	GGTYSCHFGLTWVCKPQGG	[3]	C2Pred	COMP
GRKKRRQARAPPQC	[3]	C2Pred	COMP	VWPLGLVICKALKIC	[3]	C2Pred	COMP
WELYGRKKRRRQRRR	[2]	CPPsite 2	COMP	WLNALLHHLNCAKGVLA	[3]	C2Pred	COMP
LAQLLAQLLAQLGGGRRRRRRRRR	[2]	CPPsite 2	COMP	FLPLAIGLLGLKLF	[3]	C2Pred	COMP
YSHIATLPFTPT	[3]	C2Pred	COMP	QWGYGGMPYGGYGGMGYGMGGYGMGY	[3]	C2Pred	COMP
AAVALPAVLLALLAPRRRRRR	[2]	CPPsite 2	COMP	FLPLFLPKIICVITKKC	[3]	C2Pred	COMP
VGALAVVWLWLWLAGSGPKKKRKVC	[2]	CPPsite 2	COMP	QGVNDNEEGFASAR	[3]	C2Pred	COMP
RQIKIWFQNRAMKWK	[2]	CPPsite 2	COMP	NGARVSDMFRPSGDDFGDYSANWGDF	[3]	C2Pred	COMP
GCGGGYGRKKRRRQRRR	[3]	C2Pred	COMP	GFLDKLKKGASDFANALVNSIKGT	[3]	C2Pred	COMP
RFTFHFREFFTFHFEGGRRRRRRR	[3]	C2Pred	COMP	KRGGAQYAPYWQETYLRSRK	[3]	C2Pred	COMP
GKKALKLAAKLLKCC	[3]	C2Pred	COMP	QWAQWRPRTPIPP	[3]	C2Pred	COMP
RKKRRQRRR	[2]	CPPsite 2	COMP	SLSRFLSFLKIVYPPAF	[3]	C2Pred	COMP
KPRSKNPPKPK	[3]	C2Pred	COMP	VVNTPGHAVSYHVV	[3]	C2Pred	COMP
LLIILRRIRKQAAAXSK	[2]	CPPsite 2	COMP	IKIMDILAKGLKVLAVHG	[3]	C2Pred	COMP
LTMPSLDQPVW	[3]	C2Pred	COMP	DVLKKIGTVALHAGKALGAVADTISQ	[3]	C2Pred	COMP
RRRRRRRRRRXXX	[2]	CPPsite 2	COMP	FLSAITSLGKLL	[3]	C2Pred	COMP
QIKIWFQNRMMKWKK	[2]	CPPsite 2	COMP	SDESDGDRPQASPLGPGP	[3]	C2Pred	COMP
KLALKALAKALKA	[2]	CPPsite 2	COMP	GIVEQCDDTPCSLYDPENYCN	[3]	C2Pred	COMP
RHNFRFFNFRTNR	[3]	C2Pred	COMP	SGTGLSATLPQRF	[3]	C2Pred	COMP
MLLTTRRRST	[2]	CPPsite 2	COMP	GTLPCESCVMIPCISSVVGCSCKSKVCYKN	[3]	C2Pred	COMP
WIIFRIASXKK	[2]	CPPsite 2	COMP	GCCSTPPCAVLYC	[3]	C2Pred	COMP
IRQRRR	[2]	CPPsite 2	COMP	VCIADDMPCGFLGGPLCCSGWCLFVCL	[3]	C2Pred	COMP
RQIKIWFQNRMMKWKA	[2]	CPPsite 2	COMP	KGAAGKLLLEVASCCKLSKSC	[3]	C2Pred	COMP
CSKSSDYQC	[2]	CPPsite 2	COMP	MRTWACLILLGCGYLAFAALAV	[3]	C2Pred	COMP
CGNKKTR	[2]	CPPsite 2	COMP	SVLTPSLSSLGESLESIS	[3]	C2Pred	COMP
SWLPYPWHVPSS	[3]	C2Pred	COMP	LNENLLRFFVAPFPEVFG	[3]	C2Pred	COMP
RQIKAWFQNRMMKWKK	[2]	CPPsite 2	COMP	ALQTLPAMCNVY	[3]	C2Pred	COMP
GRRERNKMAAAKCRNRRR	[2]	CPPsite 2	COMP	RDSLQRGQKILEKAERIGDRIKDIFRG	[3]	C2Pred	COMP
RVTSWLGRLRIAGKRLGRSK	[2]	CPPsite 2	COMP	TAEALRCQENYLPSPCQ	[3]	C2Pred	COMP
SWWTPWVXVSES	[2]	CPPsite 2	COMP	FLPAVIRVAANVLPTAFCAISKCC	[3]	C2Pred	COMP
YARAARRAARR	[2]	CPPsite 2	COMP	QCCITIECCRI	[3]	C2Pred	COMP
RWRRWRRW	[2]	CPPsite 2	COMP	GCCSDPRCRYRC	[3]	C2Pred	COMP
GGRRARRRRRR	[2]	CPPsite 2	COMP	QIDPLGFSGGI	[3]	C2Pred	COMP
XATKSQINF	[2]	CPPsite 2	COMP	NGGTSGLFAFPRV	[3]	C2Pred	COMP
WLKLWKKWLKLW	[2]	CPPsite 2	COMP	GLFDIINKIVSTL	[3]	C2Pred	COMP
RLRPRRPRLPFRPG	[2]	CPPsite 2	COMP	FLAGLIGGLAKML	[3]	C2Pred	COMP
KRIPNKKPGKKTITKPTKPTIKTKKDLK	[3]	C2Pred	COMP	INWLKLGAHIDAL	[3]	C2Pred	COMP
RKKNPNCRXX	[2]	CPPsite 2	COMP	FLPAVLRAVAVGPAVFAITQKC	[3]	C2Pred	COMP
WEAKLAKALAKALAKHLAKALAKALACEA	[3]	C2Pred	COMP	ASEDALFGTMR	[3]	C2Pred	COMP
SWAQXLSLPPVL	[2]	CPPsite 2	COMP	SPMQRSSMVRF	[3]	C2Pred	COMP
GALFLGFLGAAGSTMGAWSQPKSKRKV	[3]	C2Pred	COMP	PFSILPHAIIGGLISAIK	[3]	C2Pred	COMP
RLRLRLRLRLRLRLRL	[3]	C2Pred	COMP	SEAAALPRASAAAAMRAAWPSPSVERV	[3]	C2Pred	COMP
RIFIRIGC	[2]	CPPsite 2	COMP	VPSAGDMMVRF	[3]	C2Pred	COMP
RRRRRRRQIKILFQNRMMKWKKGGC	[3]	C2Pred	COMP	TQRLANFLHSSNNFGAIFSPN	[3]	C2Pred	COMP
CRGDC	[2]	CPPsite 2	COMP	GKLAFLAKMKEIAAQT	[3]	C2Pred	COMP
XXRIRQSLIML	[2]	CPPsite 2	COMP	KLSPSLGPVSKGLLAGQR	[3]	C2Pred	COMP

FQNRMRKWK	[2]	CPPsite 2	COMP	EWKLPDLIINHITLRRNCNKYRCG	[3]	C2Pred	COMP
RKKRRQRRRG	[2]	CPPsite 2	COMP	GFGMLFKFLAKKVAKKLVSHVAQKQLE	[3]	C2Pred	COMP
PIRRRKLRRLK	[3]	C2Pred	COMP	CTCFYTKDKECVYYCHLDIHWINTP	[3]	C2Pred	COMP
VKRFFKKFRKLKKV	[2]	CPPsite 2	COMP	FLPFLIPALTSILSL	[3]	C2Pred	COMP
RILQQLLFI	[2]	CPPsite 2	COMP	SMAMGRLGLRPG	[3]	C2Pred	COMP
GSPWGLQHHPRT	[3]	C2Pred	COMP	KLFNGNEVCLDPKEKWVQKVQIFLK	[3]	C2Pred	COMP
MIYRIAASXKK	[2]	CPPsite 2	COMP	DCLPGWSVYEGRCYKVFNQKTWKAEEKFC	[3]	C2Pred	COMP
TRRSKRSSHRRK	[3]	C2Pred	COMP	MMRDSGCFGRRLDRIGSLGSLGCVNLRRY	[3]	C2Pred	COMP
AAVACRICMRNFSTRQARRNHRHRHRR	[3]	C2Pred	COMP	QEADPSSSLEADSTLKDEPRELSNM	[3]	C2Pred	COMP
CGGRRRRRRRRRLLL	[2]	CPPsite 2	COMP	AGNLSECFWKYCV	[3]	C2Pred	COMP
WELVVLGKYGRKKRRQRRR	[2]	CPPsite 2	COMP	GGAGEPLAFSPDMLSLRF	[3]	C2Pred	COMP
KWFKIQMQIRRWKNR	[3]	C2Pred	COMP	GLEESPGHPGQPGPPGAPGP	[3]	C2Pred	COMP
RWRCKNQ	[2]	CPPsite 2	COMP	MYKIQLLSIALTLALVANGAPTSSSTGNT	[3]	C2Pred	COMP
LNVPSPWFLSQR	[3]	C2Pred	COMP	FKAPYNHWHCKPGLLC	[3]	C2Pred	COMP
YGRKKRRQRRSVYDFVWL	[3]	C2Pred	COMP	PAETPNSLDLTFNRRIMDTI	[3]	C2Pred	COMP
QRIRKSKISRTL	[3]	C2Pred	COMP	HSDAVFTDNYTRLRKQMAVKKYLSILN	[3]	C2Pred	COMP
LCLR	[2]	CPPsite 2	COMP	FSETIPAPTSKNEAQKS	[3]	C2Pred	COMP
VPALR	[2]	CPPsite 2	COMP	GPRPPGFSFPRGKFHSQS	[3]	C2Pred	COMP
GYGRKKRRQRRG	[2]	CPPsite 2	COMP	SDPSVPVEPEDDDMVQ	[3]	C2Pred	COMP
PLSSIFSRIQDP	[3]	C2Pred	COMP	MIASHLAFELSKLGSKHTML	[3]	C2Pred	COMP
RQIRIWFQNRMRWRRC	[3]	C2Pred	COMP	RSLDASPSAFSGNHSLS	[3]	C2Pred	COMP
RQIKIWFQNRMRMAWK	[2]	CPPsite 2	COMP	SEAAALPRASAAAMSCVAEPECREG	[3]	C2Pred	COMP
LCL	[2]	CPPsite 2	COMP	SSSMYDRDIMSFG	[3]	C2Pred	COMP
KMDCRWKWKSSKK	[2]	CPPsite 2	COMP	MKVFFLFAVLFCVLRRNSVHISHQEARGP	[3]	C2Pred	COMP
SRXXCRSKAARSXX	[2]	CPPsite 2	COMP	SVNTKNDFMRF	[3]	C2Pred	COMP
RTLNEYKNTLKF	[3]	C2Pred	COMP	MQFITDLIKKAVDFKGLFGNK	[3]	C2Pred	COMP
YARKARRAAR	[2]	CPPsite 2	COMP	SWPVCTRNLPGVCGETCVGGTCNTPGCTC	[3]	C2Pred	COMP
LLILRRRARKQAXASK	[2]	CPPsite 2	COMP	MSNRGASLKLFLAVLLVNTLLTKEGVT	[3]	C2Pred	COMP
CIGAVLKVLTTGLPALISWIKRKRQQ	[3]	C2Pred	COMP	GWFDVVKHIAVAV	[3]	C2Pred	COMP
WCKRRQCFRVLXXWN	[2]	CPPsite 2	COMP	STDCGGPKTQPLACDHPPLPDILFL	[3]	C2Pred	COMP

A.2 Test dataset of CPPs and non-CPPs

Sequences of CPPs and non-CPPs used in the independent test dataset with their origin (CPPSite 2, C2Pred, DB1*, or DB2**), and their model: Experimental (EXPR) or Computational (COMP).

Sequence	Reference	Origin	Model	Sequence	Reference	Origin	Model
KWRRKLKLRPKKKRV	LDP-NLS	DB2	COMP	KKLSECLKRIGDELS	Bax BH3	DB1	COMP
KALKKLLAKWAAKALL	MAP 8	DB1	COMP	RPPGFSFPR	Bradykinin	PDB	EXPR (PDB code: 6F3V)
RRLSSYSRRRF	SynB3	DB1	COMP	IAARIKLSRQHIKRLHL	scr pVEC	DB1	COMP
GRKKRRQRRPPC	ptat4	DB1	COMP	CYFQNCPRG	Vasopressin	DB1	COMP
CNGRCG	Aminopeptase	DB1	COMP	FVPIFTHSELQKIREKERNKGQ	Motolin	DB1	COMP
LIRLWSHLIHFQNRRLKWK	EB1	DB1	COMP	AWRRKLKALAPAKKAV	Mut-LDP-NLS	DB2	COMP
AHALCPPERQIKIWFQNRMRKWKKEN	pAntpHD 40P2	DB1	COMP	KIWFQNRMRK	pAntp(4-13)	DB1	COMP
AAVALLPAVLALLAKNNLKDCLF	[1]	DB1	COMP	DSSNLPPNQKQIVD	[3]	C2Pred	COMP
CNGRCGKLAKLAKLAK	[1]	DB1	COMP	VKRCCDEECSSACWPCCWG	[3]	C2Pred	COMP
GGQIKIWFQNRMRKWK	[1]	DB1	COMP	LLKELWTKMKGAGKAVLGKIKGLL	[3]	C2Pred	COMP
LLILARIRKQAHASHK	[1]	DB1	COMP	TTTTVNVKNSYTVWPGALPGGGVLD	[3]	C2Pred	COMP
MDAQTRRRERRAEKQAQWKAAN	[1]	DB1	COMP	YKVEDLQAGGQISRGYFFFRPN	[3]	C2Pred	COMP
MGLGLHLVLAALQGAKKKRV	[1]	DB1	COMP	DHLPHDVYSPRL	[3]	C2Pred	COMP
NAKTRRHERRRKLAIR	[1]	DB1	COMP	APGDRIYVHPF	[3]	C2Pred	COMP
CGRKKRWQRQRRPPQ	[2]	CPPsite 2	COMP (CPPSite ID: 2623)	HADGLFTSGYSKLLGQLSARRYLESLI	[3]	C2Pred	COMP
MIYRDL	TCTP (1-7)	CPPsite 2	COMP (CPPSite ID: 1586)	EGGGPQWAVGHFM	[3]	C2Pred	COMP
LLRARWRRRRSRRFR	[3]	C2Pred	COMP	MHVERRECAVCLTINTTICAGYCMTR	[3]	C2Pred	COMP
GGRRRRRRYGRKKRRQRR	[3]	C2Pred	COMP	AALKGCVTKSPKPCSGKR	[3]	C2Pred	COMP
GRQLRAGRRLRGRSR	[3]	C2Pred	COMP	GVVTDLLKTAGKLLGNLVGSLSG	[3]	C2Pred	COMP
YRAARRAARA	CTP503	CPPsite 2	COMP (CPPSite ID: 1724)	GWASSIGSILGFAKGGAQAFQPK	[3]	C2Pred	COMP
GWTLNSAGYLLGKINLAKAALAKLL	TP2	CPPsite 2	COMP (CPPSite ID: 1045)	SYGWAEGDITDNEYLR	[3]	C2Pred	COMP
GSVRRRRRRRGGRRRR	[3]	C2Pred	COMP	GFGSFLGKALKALKIGANVLGGAPQ	[3]	C2Pred	COMP
IYRDLIX	[2]	CPPsite 2	COMP	AIFIFIRWLLKLGHHGRAPP	[3]	C2Pred	COMP
CGYGRKKRRQRRRG	Tat	CPPsite 2	COMP (CPPSite ID: 2491)	GCSSHPACNVNPHICG	[3]	C2Pred	COMP
RXXRLXRLXL	F3	CPPsite 2	COMP (CPPSite ID: 2924)	GNNRPVYIOPRPPHPRI	[3]	C2Pred	COMP
VCVR	[2]	CPPsite 2	COMP	DTHISEKIICNDIG	[3]	C2Pred	COMP
RKSSKPIEMKRRRAR	[3]	C2Pred	COMP	MWITNGGVANWYFVLAR	[3]	C2Pred	COMP
KFFKFFKFFK	CPP-PNA	CPPsite 2	COMP (CPPSite ID: 2072)	QGLPPGPIPR	[3]	C2Pred	COMP
KETWFETWTEWSOPKKKRV	[2]	CPPsite 2	COMP	NPKVAHCASQIGRSTAWGAVSGA	[3]	C2Pred	COMP
KALAALLKLAALLAALK	[2]	CPPsite 2	COMP	RCCQTFYWCVCQ	[3]	C2Pred	COMP
WELVYGRKKRRQRRR	[2]	CPPsite 2	COMP	ADSDPVGGEFLAEGGGVR	[3]	C2Pred	COMP
LCLK	[2]	CPPsite 2	COMP	DNTVTSKPLNCMNYPWKSRTAC	[3]	C2Pred	COMP
GLFKALLKLLKSLWKLKLLKA	[2]	CPPsite 2	COMP	DGCSNAGACFGIHPGLCCSEICIVWCT	[3]	C2Pred	COMP
KMIFVGKKEERA	[2]	CPPsite 2	COMP	MKVSAAALAVILIALCA	[3]	C2Pred	COMP
KRIPNKPQK	[2]	CPPsite 2	COMP	SDPSVPLRPEDELIDQ	[3]	C2Pred	COMP
GYGRKKRRGRRTHRLPRRRRRR	[3]	C2Pred	COMP	CKGKGQSCSKLMYDCCTGSCSRRGKC	[3]	C2Pred	COMP
MVRRLVTLRIRACGPPRVV	[3]	C2Pred	COMP	GVSHFPLRKEKDDNSGSRKSNPK	[3]	C2Pred	COMP
KRIHPRLTSIR	[3]	C2Pred	COMP	IIPPLPGYFAKT	[3]	C2Pred	COMP
NTCTWLKYX	[2]	CPPsite 2	COMP	QADPNKFYGLM	[3]	C2Pred	COMP
RLWMRWYSPRTRAYG	[2]	CPPsite 2	COMP	FLGRVLPFTRATASTHRSRL	[3]	C2Pred	COMP
ACSSSPSKXG	[2]	CPPsite 2	COMP	FWGHIWNAVKRVGANALHGAVTGALS	[3]	C2Pred	COMP
SARXXCRSKAKRSXX	[2]	CPPsite 2	COMP	GNTKKAVPGFYGR	[3]	C2Pred	COMP
WEYGRKKRRQRRR	[2]	CPPsite 2	COMP	FLPLVTMLLGKLF	[3]	C2Pred	COMP
FQWQRNMRKVRGPPVS	[2]	CPPsite 2	COMP	VTMVEAGFGCFSPSPRSDSHCRGMGR	[3]	C2Pred	COMP

KRIXPRLTRSIR	[2]	CPPsite 2	COMP	YGGFLRRIRFARKLANQ	[3]	C2Pred	COMP
RLVMRVYSPTTRYG	[2]	CPPsite 2	COMP	GNGVVLTLTHECNLATWTKKLKCC	[3]	C2Pred	COMP
KLALKLALKALKAA	[2]	CPPsite 2	COMP	TMKLCGRKLPETLSKLCVY	[3]	C2Pred	COMP
RWRWRWRWRWR	[3]	C2Pred	COMP	GLRSKIWLWVLLMIWQESNKFKKM	[3]	C2Pred	COMP
ACSDRFRCNCPADEALCGRRRRRRRR	[2]	CPPsite 2	COMP	KFDMVAVYSEEDS	[3]	C2Pred	COMP
LKTLTETLKELTKLTTEL	[3]	C2Pred	COMP	GLWSKIKDVAAAAGKAALGAVNEALGEQ	[3]	C2Pred	COMP
LGLLLRXLRXXSNLLANI	[2]	CPPsite 2	COMP	SDRPTRAMDSPLIRF	[3]	C2Pred	COMP
KSHAHQAQRIRRLIILL	[3]	C2Pred	COMP	QRFSQPTFKLPQGRLTLRKRF	[3]	C2Pred	COMP
GRKKRRQRARPPQC	[2]	CPPsite 2	COMP	NGVCCGYKLCHPC	[3]	C2Pred	COMP
MRRIRPRPRLPRPRPLPFPRPGGCYPG	[3]	C2Pred	COMP	LFAKINGLKVGPLKIQIV	[3]	C2Pred	COMP
PPRLPRPRPLPFPRPG	[2]	CPPsite 2	COMP	NPELYQMNHFRWGQPPTHFKQ	[3]	C2Pred	COMP
RLYMRYYSPTTRYG	[3]	C2Pred	COMP	GCCGSFACRFGCVPCCV	[3]	C2Pred	COMP
RRHLRRHLRHLRRLRHLRHL	[3]	C2Pred	COMP	FIITGLVRGLTKLF	[3]	C2Pred	COMP
CKYGRKKRRQRRR	[2]	CPPsite 2	COMP	GSSFLSPEFKKIQQNDPTKTTAKIH	[3]	C2Pred	COMP
CXAIYPRX	[2]	CPPsite 2	COMP	IIDYYDEGEDRDVGVDAR	[3]	C2Pred	COMP
SRXXCRAKAKRSRXX	[2]	CPPsite 2	COMP	PLVQQQFLGQQQFPFPQ	[3]	C2Pred	COMP
GNYAHRVGAGAPVWL	[3]	C2Pred	COMP	NPFKELERAGQVRDAIIS	[3]	C2Pred	COMP
YTFGLKTSFNVQYTFGLKTSFNVQ	[3]	C2Pred	COMP	CHRRDSHKIDNYFKVLKRLIHDSNC	[3]	C2Pred	COMP
KRPAATKAGQAKKKKL	[3]	C2Pred	COMP	ALFEESTVSAEPR	[3]	C2Pred	COMP
GRKKKKRT	[2]	CPPsite 2	COMP	GIGGKPVQTAfVDNDGIYD	[3]	C2Pred	COMP
MIYRDLI	[2]	CPPsite 2	COMP	GAFGDLKGVAKAEAGLKLNNMAQCKLSGNC	[3]	C2Pred	COMP
GRRHHCRSAKRSRHH	[3]	C2Pred	COMP	FFGHLYRGITSVVKHVGLLSG	[3]	C2Pred	COMP
RKKRRRESWVXLPPVXLPPPGGXXXXXX	[2]	CPPsite 2	COMP	LRTLLELARTQSQRERAEQNRIIFDSVGK	[3]	C2Pred	COMP
RQARRNRRRALWKTLLKKVLKA	[3]	C2Pred	COMP	GLLSGVLGVGKKVDCGLSGLC	[3]	C2Pred	COMP
WELVVLYGRKKRRQRRR	[2]	CPPsite 2	COMP	SLSYEDKMFNDNVEFTPLR	[3]	C2Pred	COMP
RRRRRRRGIIYLATALAKWALKQGF	[3]	C2Pred	COMP	INWKAIIIEAAKQAL	[3]	C2Pred	COMP
KLALKLALKWAKLAKAA	[3]	C2Pred	COMP	FLGALFKVASKVLPSVFCAITKKC	[3]	C2Pred	COMP
TKRRITPKDVIDV	[2]	CPPsite 2	COMP	DGSVDFKKNWIQYKEGFGHLSPTG	[3]	C2Pred	COMP
RVRVFVVIHPRLT	[3]	C2Pred	COMP	NWTPQAMLYLKGAQ	[3]	C2Pred	COMP
WRFKWRFKWRFK	[3]	C2Pred	COMP	RELEELNVPGEIVESLSSEESITRINK	[3]	C2Pred	COMP
RRGRRG	[2]	CPPsite 2	COMP	MTDMWSLKICAWLGFLLLFKP	[3]	C2Pred	COMP

* DB1 means the database from Sanders et al (2011) [1].

** DB2 means the database from Ponnappan, N. Chugh, A. (2017) [4].

References

- [1] – Sanders, W. S., Johnston, C. I., Bridges, S. M., Burgess, S. C. Willeford, K. O. Prediction of Cell Penetrating Peptides by Support Vector Machines. PLoS Comput. Biol. 7, e1002101 (2011).
- [2] – Agrawal, P. et al. CPPsite 2.0: a repository of experimentally validated cell-penetrating peptides. Nucleic Acids Res. 44, D1098–D1103 (2016).
- [3] – Tang, H., Su, Z.-D., Wei, H.-H., Chen, W. Lin, H. Prediction of cell-penetrating peptides with feature selection techniques. Biochem. Biophys. Res. Commun. 477, 150–154 (2016).
- [4] - Ponnappan, N. Chugh, A. Cell-penetrating and cargo-delivery ability of a spider toxin-derived peptide in mammalian cells. European Journal of Pharmaceutics and Biopharmaceutics 114, 145–153 (2017)

B APPENDIX B

Categorization of BBB permeability of peptides according to five classes of influx level.

	K_{in}	P
Class 1 Very low influx	$[0; 1.83 \times 10^{-4}]$	$[0; 1.74 \times 10^{-5}]$
Class 2 Low influx	$]1.83 \times 10^{-4}; 3.78 \times 10^{-4}]$	$]1.74 \times 10^{-5}; 3.44 \times 10^{-5}]$
Class 3 Medium influx	$]3.78 \times 10^{-4}; 9.37 \times 10^{-4}]$	$]3.44 \times 10^{-5}; 8.21 \times 10^{-5}]$
Class 4 High influx	$]9.37 \times 10^{-4}; 2.10 \times 10^{-3}]$	$]8.21 \times 10^{-5}; 1.53 \times 10^{-4}]$
Class 5 Very high influx	$]2.10 \times 10^{-3}; \infty]$	$]1.3 \times 10^{-4}; \infty]$

Source: Adapted from Stalmans et al. 2015.

C APPENDIX C

List of peptides' name and Brainpeps's identification (PID) used in cross-validation analysis. Peptides that can cross the BBB are labeled as BBB+ and those that can not are labeled as BBB-.

PID	Name	Label
1	Vasopressin	BBB-
2	Oxytocin	BBB+
3	Thyrotropin releasing hormone	BBB+
6	Vapreotide	BBB-
7	Corticotropin-releasing hormone	BBB-
8	Pituitary adenylate cyclase-activating polypeptide-27	BBB+
9	Pituitary adenylate cyclase-activating polypeptide-38	BBB+
10	Vasoactive intestinal peptide	BBB+
11	Neuropeptide Y	BBB-
14	Orexin A	BBB-
15	Orexin B	BBB-
16	Exendin-4	BBB+
17	Phe ¹³ , Tyr ¹⁹ -Melanin concentrating hormone	BBB-
18	Deltorphan I, [D-Ala ²]deltorphan I, Deltorphan C	BBB+
19	Deltorphan II, [D-Ala ²]deltorphan II	BBB+
21	Adrenomedullin	BBB+
22	Urocortin-I	BBB-
23	Insulin	BBB+
27	Amylin	BBB+
28	[Tyr ¹⁰] Secretin-27	BBB+
29	[Met ⁵] Enkephalin	BBB+
30	[Leu ⁵] Enkephalin	BBB+
31	Endomorphin-1	BBB+
32	Endomorphin-2, Endomorphin II	BBB+
36	Tyr-Trp-Melanocyt-stimulating hormone (MSH) release-inhibiting factor	BBB+
37	Biphalin	BBB+
38	D-Penicillamine-D-Penicillamine Enkephalin	BBB-
39	[Glu ⁴] Deltorphan	BBB+
40	SAM 995	BBB-
41	SAM 1095	BBB-
43	Ebiratide	BBB+
47	Pancreatic Polypeptide	BBB+
55	Epidermal growth factor	BBB+
57	[Met(O) ⁶⁷]- (cocaine and amphetamine regulated transcript)-(55—102)	BBB+
58	Mahogany (1377-1428)	BBB+
59	Urocortin II	BBB+

60	Urocortin III	BBB+
61	Luteinizing hormone-releasing hormone	BBB+
64	Des-octanoyl ghrelin	BBB+
66	Peptide YY (3-36)	BBB+
67	Melanocyte stimulating hormone	BBB+
69	p-[Cl-Phe ^{4'4'}] Biphalin	BBB+
72	[p-Cl-Phe ⁴]-[(D)Pen ² , (D)Pen ⁵] Enkephalin	BBB+
74	Connective tissue-activating peptide	BBB-
75	LSZ 916	BBB-
76	LSZ 1025	BBB-
77	LSZ 62	BBB-
78	SAM 1025	BBB-
79	SAM 1040	BBB-
80	Dynorphin (1-13)	BBB-
83	Cyclo (His-Pro)	BBB-
85	Glycylsarcosine	BBB+
90	[(D)Ala ² , (D)Leu ⁵] Enkephalin	BBB-
91	DALDA	BBB-
93	[(D)Ala ² , (N-Me)Phe ⁴ , Gly-ol] Enkephalin	BBB+
98	Desglycinamide-arginine-vasopressin	BBB+
100	TAPA	BBB-
101	Substance P	BBB+
104	TAPS	BBB-
105	TAPP	BBB+
106	CTOP	BBB-
110	Dermorphin	BBB+
113	mouse Obestatin	BBB+
114	P41	BBB-
115	P42	BBB-
116	P43	BBB-
117	Sb-Aba	BBB+
121	Api88	BBB-
122	Apidaecin Api137	BBB+
130	Agouti-related protein (83-132)	BBB-
131	RC-121	BBB-
132	RC-161	BBB-
134	(3-methyl-His ²) Thyrotropin-releasing Hormone	BBB+
135	SKB P5 (12)	BBB-
136	AN110 (14)	BBB-
137	Dmt ¹ -Endomorphin 2	BBB+
138	E-2078	BBB+
139	ADAB	BBB-
140	ADAMB	BBB-
141	cationic AVP ₄₋₉	BBB-
142	Acyloxyalkoxy-based cyclic DADLE	BBB-
143	Coumarinic acid-based cyclic DADLE	BBB-
144	Oxymethyl-modified coumarinic acid-based cyclic DADLE	BBB-
145	Dalargin-SS-SynB1	BBB+
146	Dalargin-SS-SynB3	BBB+
147	[(1S,2R)-Acpc] ² -Endomorphin 2	BBB+
148	[(1S,2R)-Achc] ² -Endomorphin 2	BBB+
149	Angiopep-1	BBB+
150	Angiopep-2	BBB+

151	Angiopep-5	BBB+
152	Angiopep-7	BBB+
153	Exendin-4-NH ₂	BBB+
154	Oncocin	BBB+
155	Drosocin (unglycosylated)	BBB+
156	Drosocin Pro5Hyp (unglycosylated)	BBB-
157	LinS	BBB+
158	LinNMe	BBB-
159	CycS	BBB-
160	CycNMe	BBB-
161	mouse Ghrelin	BBB-
162	ANG1005	BBB+
163	Benzylpenicillin-SynB1	BBB+
164	Doxorubicin-SynB1	BBB+
165	Doxorubicin-(D)-Penetratin	BBB+
166	Doxorubicin-SynB3	BBB+
167	Doxorubicin-(D)-SynB3	BBB+
168	SynB3	BBB+
169	Tat 47-57	BBB+
170	pVEC	BBB+
171	Transportan 10	BBB-
172	TP10-2	BBB-
173	D-[Ala ¹]-peptide T-amide	BBB+
174	NT1, Neurotensin ⁸⁻¹³ analog	BBB+
175	N-Tyr-Delta Sleep Inducing Peptide	BBB+
176	Gastrin-releasing peptide	BBB+
177	Arginine vasopressin	BBB+
178	Glucagon	BBB+
179	N-Tyr-Corticotropin Releasing Factor	BBB+
180	N-Tyr-Bovine Adrenal Medulla Dodecapeptide	BBB+
181	N-Tyr--Endorphin	BBB+
182	N-Tyr-Somatostatin	BBB+
183	N-Tyr-Growth Hormone Releasing Factor	BBB+
184	N-Tyr-Molluscan Cardioexcitatory Neuropeptide	BBB+
185	N-Tyr-MIF-1	BBB+
186	MIF-1	BBB+
187	-Amyloid ₁₋₂₈	BBB+
188	-Amyloid ₁₋₄₀	BBB+
189	[D-Ala ² ,Ser ⁴ ,D-Val ⁵]deltorphan	BBB-
190	[D-Ala ² ,Ser ⁴ ,D-Ala ⁵]deltorphan	BBB-
191	[D-Ala ² ,Gln ⁴ ,D-Val ⁵]deltorphan	BBB-
192	[D-Ala ² ,Gln ⁴ ,D-Ala ⁵]deltorphan	BBB-
193	[Arg ⁻¹ ,Arg ⁰ ,D-Ala ²]deltorphan II	BBB-
194	[Arg ⁰ ,D-Ala ²]deltorphan II	BBB-
195	[Lys ⁻¹ ,Lys ⁰ ,D-Ala ²]deltorphan II	BBB+
196	[Lys ⁰ ,D-Ala ²]deltorphan II	BBB-
197	[Ala ⁻¹ ,Pro ⁰ ,D-Ala ²]deltorphan II	BBB-
198	[Pro ⁻¹ ,Pro ⁰ ,D-Ala ²]deltorphan II	BBB+
199	[Abu ⁻¹ ,Abu ⁰ ,D-Ala ²]deltorphan II	BBB+
200	PEG-D-Penicillamine-D-Penicillamine Enkephalin	BBB+

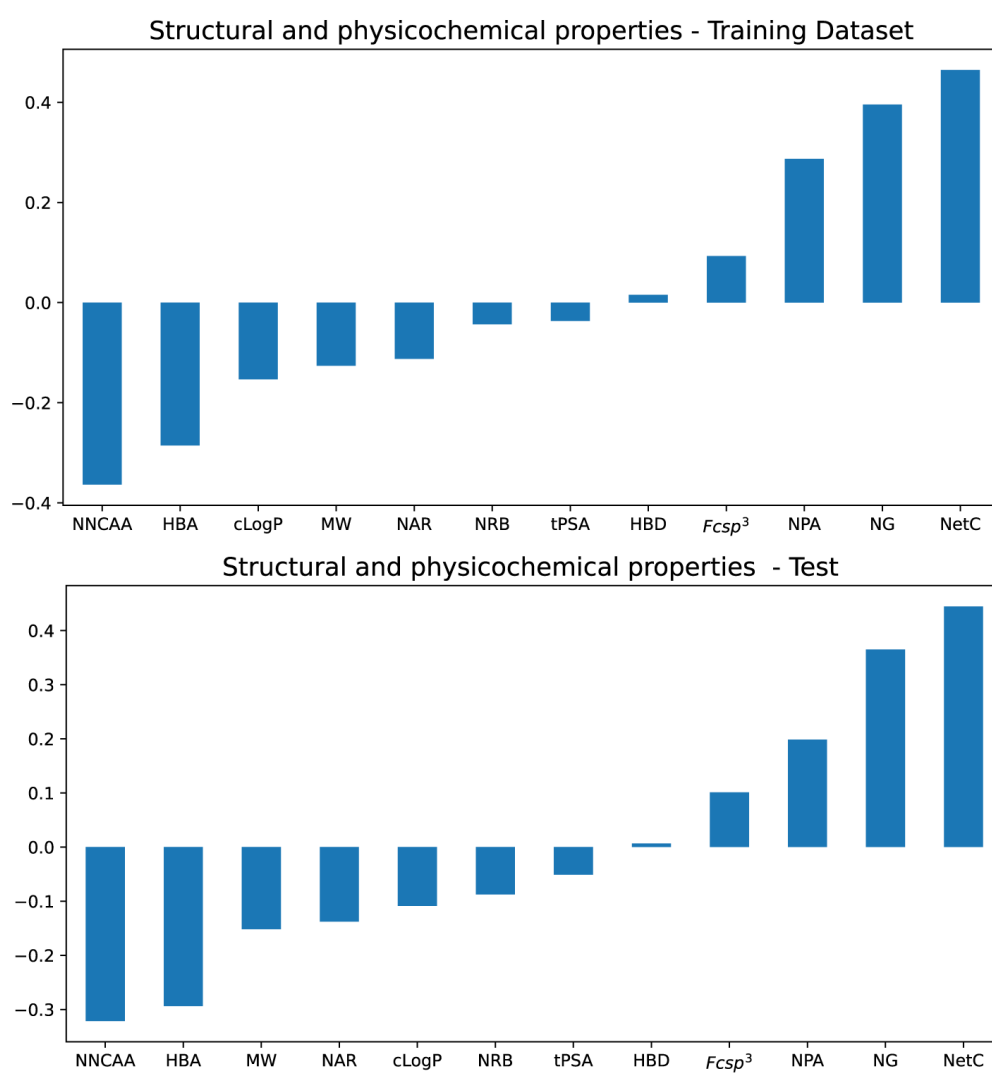
201	DPLPE-Phe-NH ₂	BBB-
202	DPLPE-Phe-OH	BBB-
203	p-[Cl-Phe ⁴]DPLPE-Phe	BBB+
204	p-[Br-Phe ⁴]DPLPE-Phe	BBB+
205	p-[F-Phe ⁴]DPLPE-Phe	BBB-
206	p-[I-Phe ⁴]DPLPE-Phe	BBB-
207	Guanidino-Endomorphin II	BBB+
208	Morphiceptin	BBB+
209	Guanidino-Morphiceptin	BBB+
210	Phe ⁰ -DPDPE	BBB+
211	DPDPE-Phe	BBB+
212	DPDPE-Arg-Gly	BBB+
213	DPDPE-Phe-Ala-NH-(CH ₂) ₅ -CONH ₂	BBB+
214	DPLCE	BBB+
215	DPLCE-Arg-Pro-Ala	BBB+
216	Insulin detemir, Levemir®	BBB-
217	Ziconotide, SNX-111, Prialt®	BBB+
218	SNX-194	BBB-
219	SNX-231	BBB-
220	SNX-185	BBB-
221	THRPPMWSPVWP-NH ₂	BBB+
222	T(NMe)H(NMe)RPPM(NMe)WSPVWP-NH ₂	BBB+
223	thrppmwspvwp-NH ₂	BBB+
224	pwvpswmprrht-NH ₂	BBB+
226	Opiorphin	BBB+
227	des-Tyrosine ¹ -D-Phenylalanine ³ --casomorphin	BBB+
228	Melanotan-II	BBB-
229	PhrANTH2	BBB-
230	BIP-2	BBB-
231	PhrCACET1	BBB+
232	Neuromedin U	BBB+
233	Neurotensin	BBB+
234	Neuromedin N	BBB+
235	Neuromedin B	BBB-
236	Arginine vasopressin 1-7	BBB+
237	Arginine vasopressin 1-6	BBB+
238	[Ser7-O-Glc]dermorphin	BBB+
239	[Ser7-O-Glc(Ac) ₄]dermorphin	BBB+
240	[Ala7-C-Gal]dermorphin	BBB+
241	[Ala7-C-Gal(Ac) ₄]dermorphin	BBB+
242	gH625	BBB+
243	D1	BBB-
244	D3	BBB+
245	Beauvericin	BBB+
246	PepH1	BBB+
247	PepH2	BBB+
248	PepH3	BBB+
249	PepH4	BBB+
250	GPE	BBB+
251	MEL-N1606	BBB+
252	MEL-N1608	BBB+
253	MEL1201	BBB+

254	MEL1209	BBB+
255	MEL1214	BBB+
256	MEL1224	BBB+
257	PepNeg	BBB+
258	C-36	BBB-
259	F-81	BBB-
260	LLVV-H4	BBB+
261	LVV-H4	BBB+
262	VV-H4	BBB+
263	VV-H7	BBB+
264	H7	BBB+
265	L57	BBB+
266	D3D3	BBB-
267	RD2D3	BBB-
268	cRD2D3	BBB+
269	AH-D	BBB+
270	NT2	BBB+
271	NT4	BBB+
272	Macrocyclic inhibitory peptide 2e	BBB+
273	Gly-Pro	BBB+
274	Tyr-Pro	BBB+
275	mouse Des-octanoyl Ghrelin	BBB-
276	ANK6	BBB-
277	tANK6	BBB+
278	cANK6r	BBB-
279	PEP inhibitory peptide	BBB+
280	OP5	BBB+
281	SLSHSPQ	BBB+
282	NTGSPYE	BBB+
287	Amyloid precursors protein derived peptide	BBB+
288	P1	BBB+
289	P2*	BBB+
290	P3	BBB+
291	P4*	BBB+
292	P5	BBB+
293	P6*	BBB+
301	THRre_2f	BBB+
304	p1	BBB-
305	p2	BBB+
306	p3	BBB+
307	p4	BBB+
308	p5	BBB+
309	p6	BBB+
310	p7	BBB-
311	p8	BBB+
312	p9	BBB+
313	p10	BBB-
314	p11	BBB+
315	p12	BBB+
316	p13	BBB+

316	p13	BBB+
317	p14	BBB+
318	p15	BBB+
319	p16	BBB+
320	AGBBB015F	BBB+
322	Monocyclic 5Br	BBB+
323	Monocyclic 6Br	BBB+
324	Monocyclic 7Br	BBB-
325	Head-to-tail cyclic	BBB-
326	Linear	BBB-
327	Bicyclic 5,5	BBB-
328	Bicyclic 6,6	BBB-
329	Bicyclic 7,7	BBB+
330		BBB-
331		BBB+
332		BBB+
333		BBB-
334	dynantin	BBB+
335	AE344	BBB+
336	Scramble	BBB-
337	Che-ADAV	BBB+
338	Che-AD(OMe)AV	BBB+
339		BBB-
340		BBB+
341		BBB+
342		BBB+
343		BBB-
344		BBB+
345		BBB+
346		BBB+
347		BBB+
348		BBB+
349		BBB+
350	GM6	BBB+
351	OFP006	BBB+
352		BBB+
353		BBB-
354		BBB-

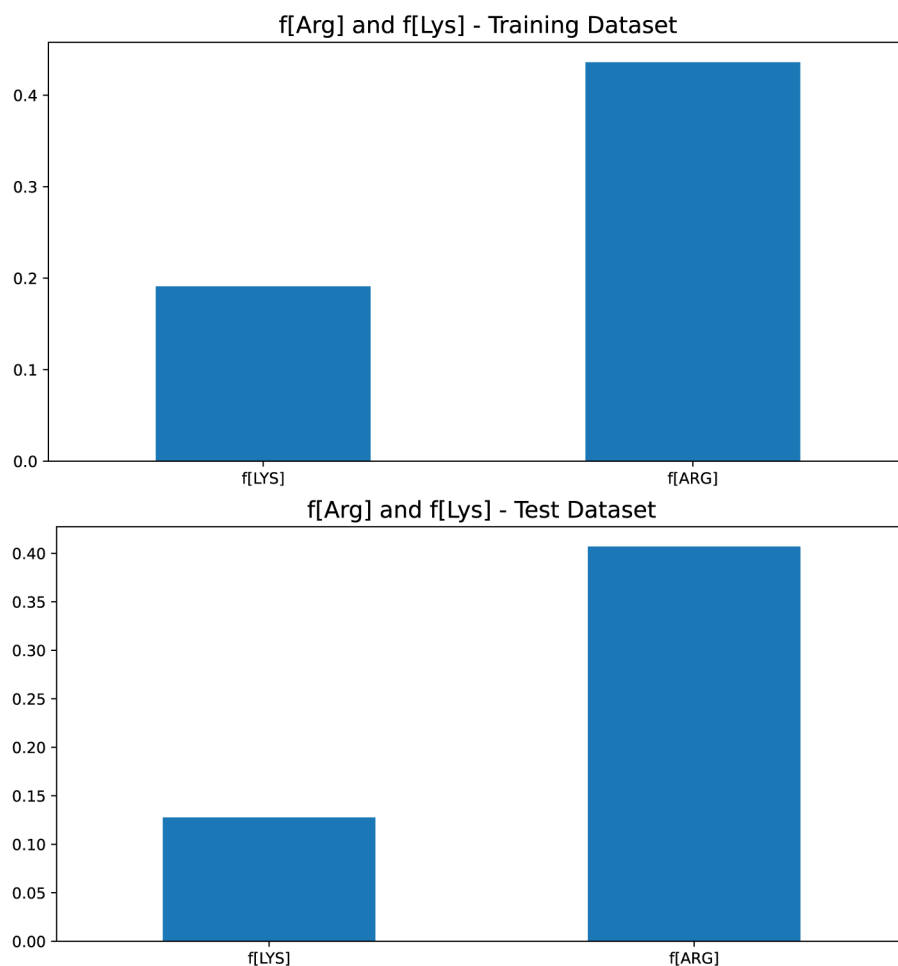
D APPENDIX D

Kendall's correlation of the twelve analyzed structural and physicochemical descriptors demonstrating the relevance to CPPs' prediction for training and independent test dataset.



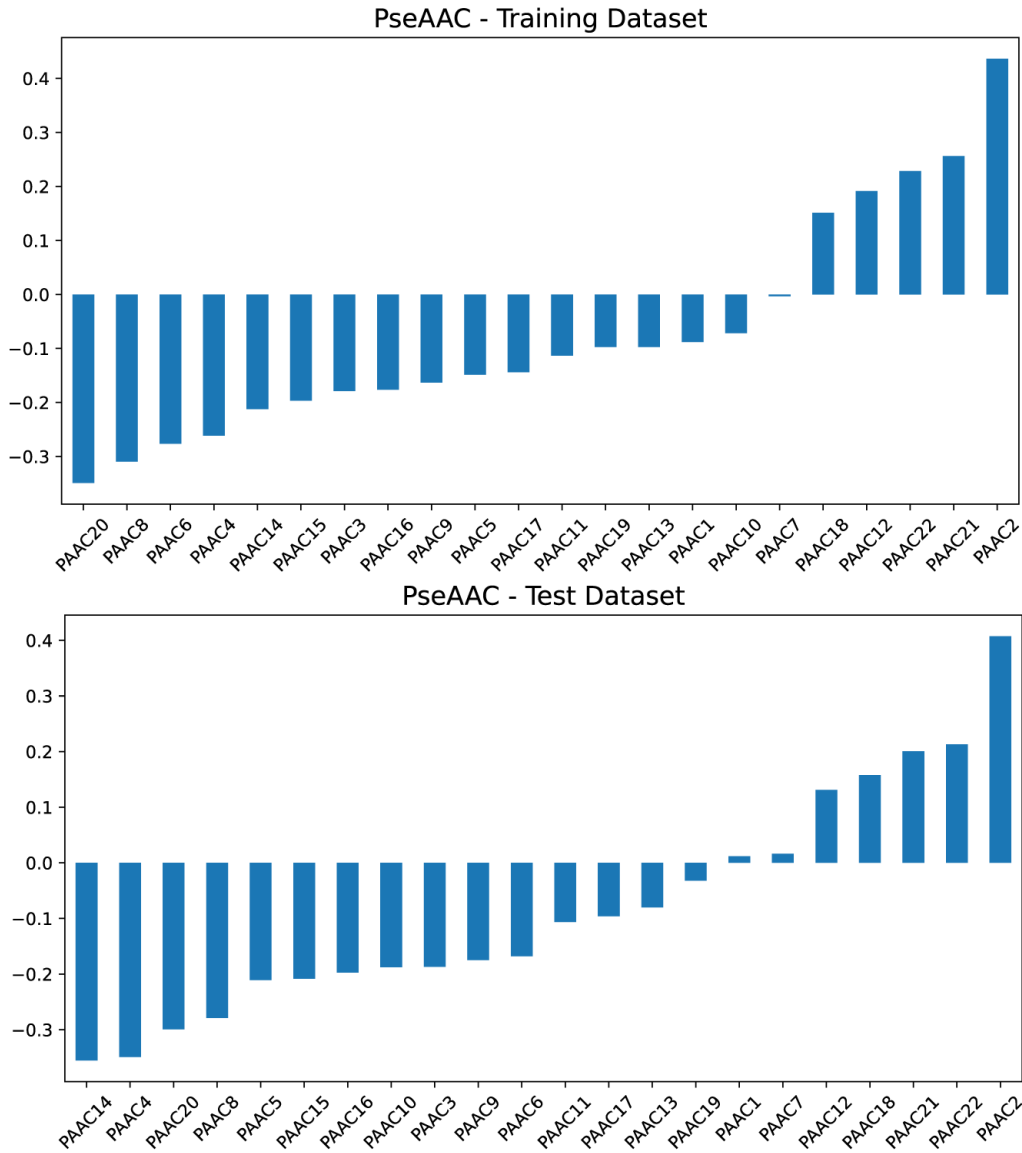
Source: Author's own.

Kendall's correlation of the analyzed sequence-based descriptors $f[\text{Arg}]$ and $f[\text{Lys}]$, demonstrating the relevance of these properties to CPPs' prediction for training and independent test dataset.



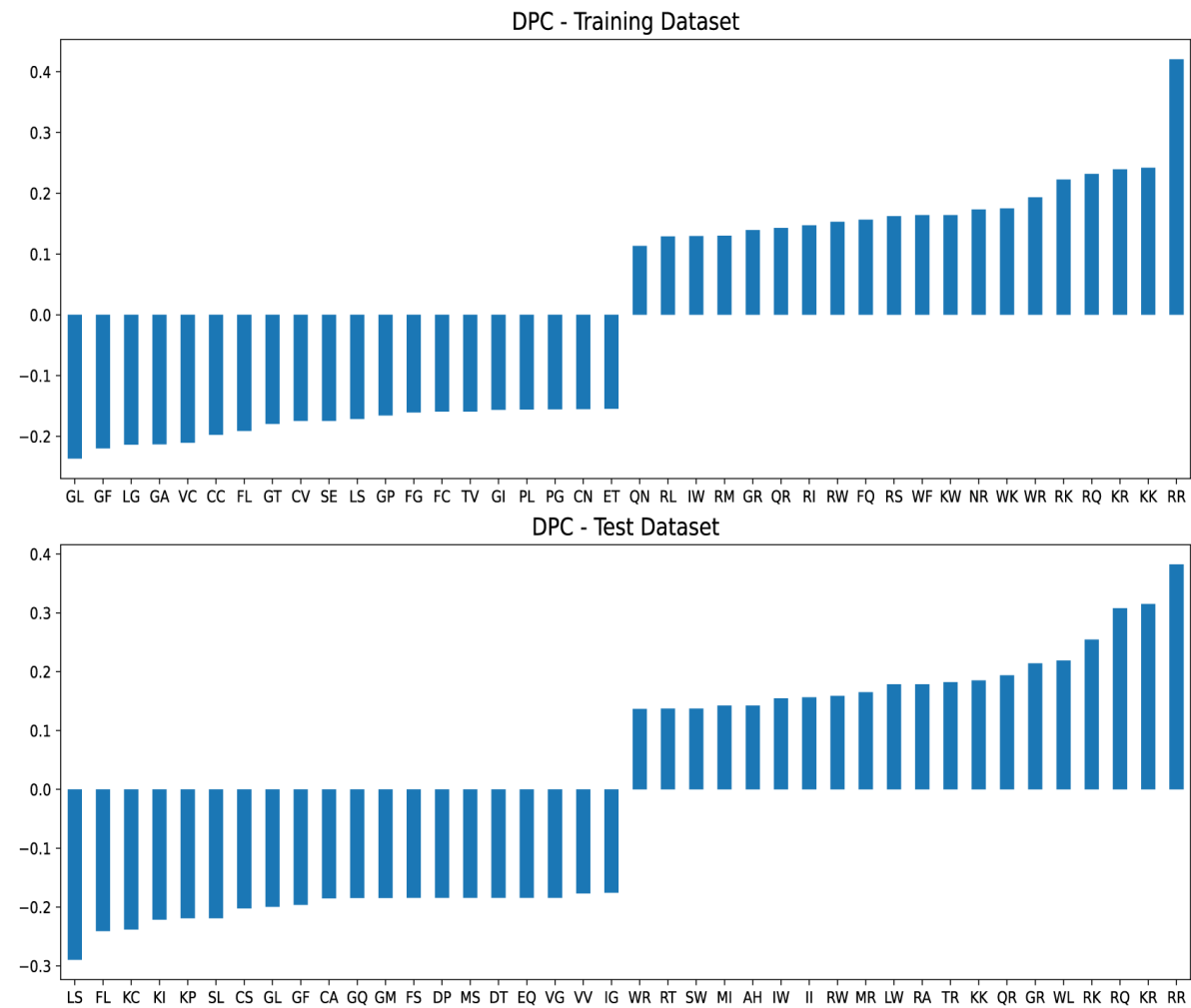
Source: Author's own.

Kendall’s correlation of the analyzed pseudo amino acid composition (PseAAC) descriptors, demonstrating their relevance to CPPs’ prediction for training and independent test dataset.



Source: Author’s own.

Kendall’s correlation of the forty most-well correlated dipeptide composition (DPC) descriptors, demonstrating their relevance to CPPs’ prediction for training and independent test dataset.



Source: Author’s own.

E APPENDIX E

List of modules of Mordred's descriptors.

Modules of Mordred's descriptors	
ABCIndex	InformationContent
AcidBase	KappaShapeIndex
AdjacencyMatrix	Lipinski
Aromatic	McGowanVolume
AtomCount	MoeType
Autocorrelation	MolecularDistanceEdge
BalabanJ	MolecularId
BaryszMatrix	PathCount
BCUT	Polarizability
BertzCT	RingCount
BondCount	RotatableBond
CarbonTypes	SLogP
Chi	TopologicalCharge
Constitutional	TopologicalIndex
DetourMatrix	TopoPSA
DistanceMatrix	VdwVolumeABC
EccentricConnectivityIndex	VertexAdjacencyInformation
EState	WalkCount
ExtendedTopochemicalAtom	Weight
FragmentComplexity	WienerIndex
Framework	ZagrebIndex
HydrogenBond	

The complete list of Mordred's descriptors is available in <https://jcheminf.biomedcentral.com/articles/10.1186/s13321-018-0258-y>.

F APPENDIX F

F.1 Appendix F1

Table with the range of the searching hyperparameters for framework models XGBr, XGBc, and γ_s of sLE for both biomembranes.

Hyperparameter	XGBr	XGBc	sLE
N° of estimators	[100; 200; 500; 1000]	[100; 200; 500; 1000]	-
Max depth	[2; 4; 8; 10; 15]	[2; 4; 8; 10; 15]	-
Learning rate	[0.01; 0.05; 0.1]	[0.01; 0.05; 0.1]	-
γ_s	-	-	[0.01; 0.02; 0.05; 0.1; 0.2; 0.5]

F.2 Appendix F2

Table with the range of the searching hyperparameters for ANN, GPC, and SVM for both biomembranes.

ANN		GPC		SVM	
Activation function	[Relu; Logistic]	Kernel	[*RBF; *Rational Quadratic]	Kernel	RBF Sigmoid
N° of hidden layers	[1; 2]		[1; 2]	C	[500; 1000; 2000; 5000]
N° of neurons by layer	[30; 40; 50; 60; 70; 80]	Length scale (RBF)	[0.01; 0.02; 0.05]	γ	[0.005; 0.01; 0.1; 0.2]
Number of iterations*	1500	Length scale (RQ)	[0.01; 0.05]	Number of iterations*	150000
Training algorithm*	Adam	Number of iterations*	100	-	-
-	-	Training algorithm*	L-BFGS-B	-	-

*Prefixed parameters.

G APPENDIX G

G.1 BChemRF-CPPred web server

The BChemRF-CPPred web server is a free web tool to predict whether peptides can cross the cell membrane. The ML algorithm behind this tool is the best voting classifier model described in the results of this thesis. Figure 37 shows the initial screen of the web server, which can be accessed at the link <http://comptools.linc.ufpa.br/BChemRF-CPPred/>.

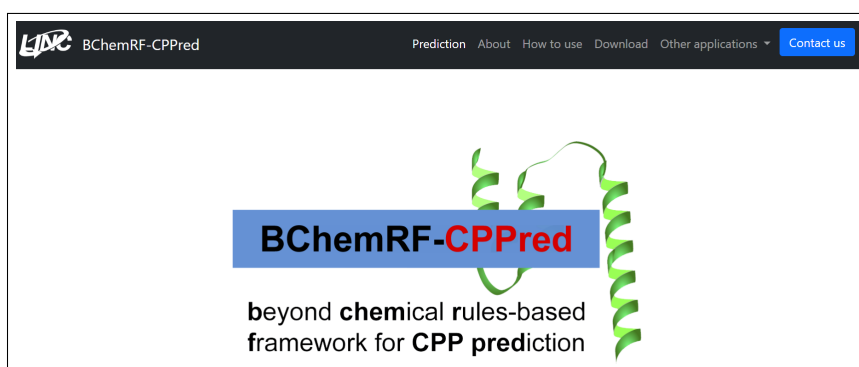


Figure 37 – Home screen of the BChemRF-CPPred web server.

On the initial screen shown above, the user can scroll down the web browser and reach the part of the tool where the user can upload and push a button to predict the permeability of peptides across the cell membrane, as shown in Figure 38. On this screen, the user can upload both FASTA or PDB files of the peptides, and select the FC and the version¹ of the ML model desired.

Upload your peptides to begin.

You can enable 'Demonstration mode' instead to use sample peptides as input.

Peptides

☐ Demonstration mode

Please type your peptides here.

Input type

☒ FASTA
 ☐ PDB

Feature composition

☐ FC-1
 ☐ FC-2
 ☐ FC-3
 ☐ FC-4

Model

☐ Version 1.0
 ☐ Version 2.0

Figure 38 – Screen for uploading peptide files for permeability prediction.

After uploading the peptide files, the user can press the button **Submit** to perform prediction or the button **Reset** to clean up the uploaded files before the prediction. The permeability prediction is shown on another screen.

The computational technologies involved in developing this web application are Python language, Flask framework, HTML, Bootstrap, JavaScript, and Docker. More information about this application can be accessed in the **About** tab located at the top of the home screen.

G.2 BrainPepPass web application

The BrainPepPass web application is a free web tool to predict whether peptides can cross the blood-brain barrier. The ML algorithm behind this tool is the framework model based on FC-4 as described in the results of this thesis. Figure 39 shows the initial screen of the GitHub page where the web application and its files are hosted, which can be accessed at the link <https://github.com/ewerton-cristhian/BrainPepPass/tree/master>.

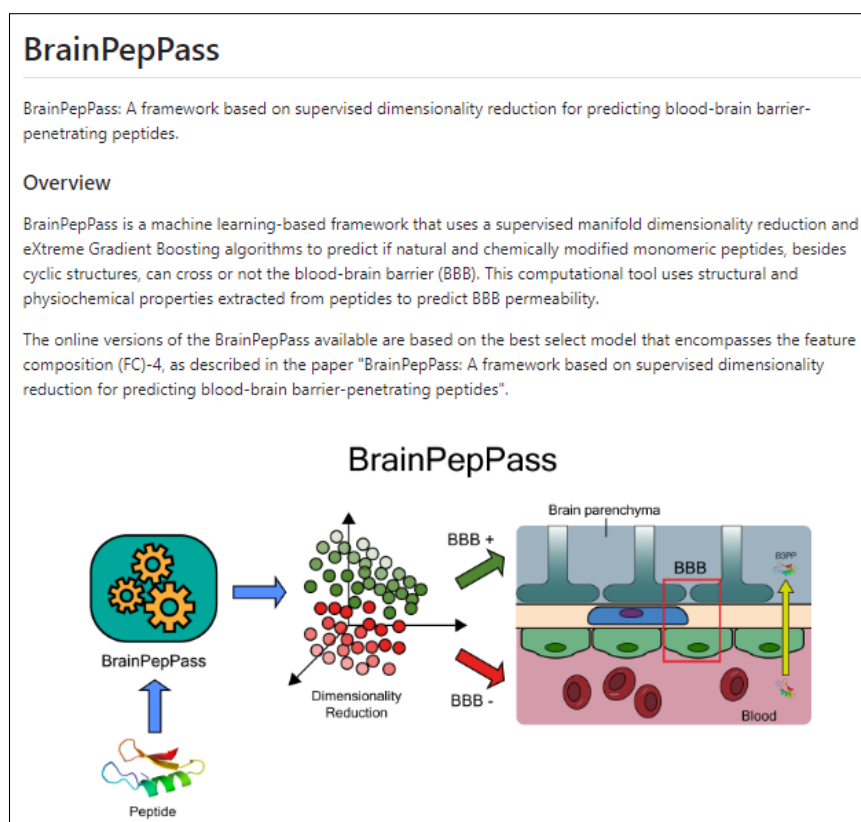


Figure 39 – Home screen of the BrainPepPass web application in GitHub.

The application where the user can access, upload, and execute the prediction of the peptide files was developed using Python language and is currently running in a notebook in Google Colab. Figure 40 shows the home screen of the notebook with the BrainPepPass application, which can be accessed at the link

<https://colab.research.google.com/drive/1O-obGm1mN7RdyevRzs3h0uQ0ZtIsNC?usp=sharing>.

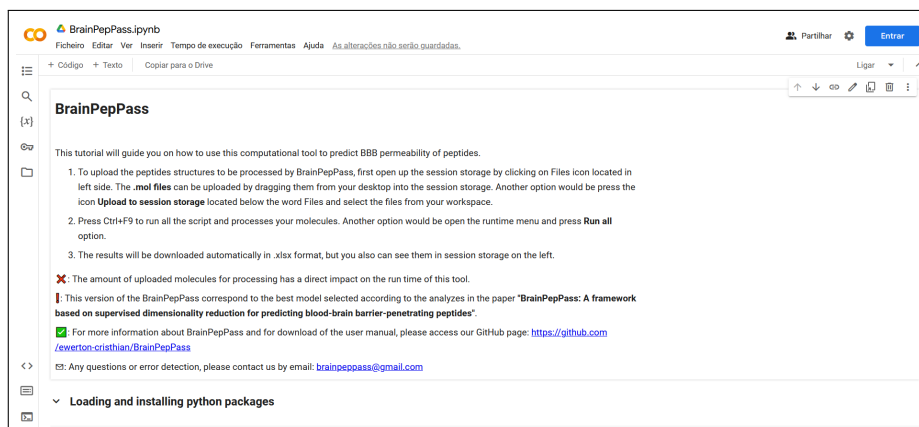


Figure 40 – Home screen of the BrainPepPass web application in Google Colab.

To upload the peptide structures to be processed by BrainPepPass, first, open up the session storage by clicking on the **Files** icon located on the left side. The .mol files can be uploaded by dragging them from your desktop into the session storage. Another option would be to press the icon Upload to session storage located below the word Files and select the files from your workspace. Press **Ctrl+F9** to run all the script and process your molecules. Another option would be to open the runtime menu and press **Run all** option. The results will be downloaded automatically in .xlsx format, but you also can see them in session storage on the left.

H APPENDIX H

Summary of the best hyper-parameters reached by gridsearch applied in ML models that compose the Vcf-CPP and the proposed DPF-CPPred.

Tables of hyper-parameters for ANN, GPC, and SVM achieved in gridsearch by FC in prediction of CPPs with Vcf-CPP using PDB encoding.

ANN		GPC		SVM	
Activation function	FC-1: Relu FC-2: Relu FC-3: Relu FC-4: Relu	Kernel	FC-1: Sigmoid FC-2: RBF FC-3: RBF FC-4: RBF	Kernel	FC-1: RBF FC-2: RBF FC-3: RBF FC-4: RBF
N° of hidden layers	FC-1: 2 FC-2: 2 FC-3: 1 FC-4: 2	α	FC-1: 1 FC-2: 1 FC-3: 1 FC-4: 1	C	FC-1: 500 FC-2: 1000 FC-3: 2000 FC-4: 5000
N° of neurons by layer	FC-1: 70 FC-2: 70 FC-3: 60 FC-4: 80	Length scale (RBF)	FC-1: 0.05 FC-2: 0.01 FC-3: 0.05 FC-4: 0.05	γ	FC-1: 0.005 FC-2: 0.1 FC-3: 0.01 FC-4: 0.005

XGBr		XGBc		sLE	
N° of estimators	FC-1: 2000 FC-2: 2000 FC-3: 1000 FC-4: 2000	N° of estimators	FC-1: 1000 FC-2: 1000 FC-3: 2000 FC-4: 1000	γ_s	FC-1: 0.02 FC-2: 0.1 FC-3: 0.02 FC-4: 0.5
Max depth	FC-1: 10 FC-2: 10 FC-3: 10 FC-4: 10	Max depth	FC-1: 4 FC-2: 4 FC-3: 8 FC-4: 4	-	-
Learning rate	FC-1: 0.05 FC-2: 0.05 FC-3: 0.05 FC-4: 0.05	Learning rate	FC-1: 0.1 FC-2: 0.1 FC-3: 0.05 FC-4: 0.1	-	-

Tables of hyper-parameters for ANN, GPC, and SVM achieved in gridsearch by FC in prediction of CPPs with Vcf-CPP using FASTA encoding.

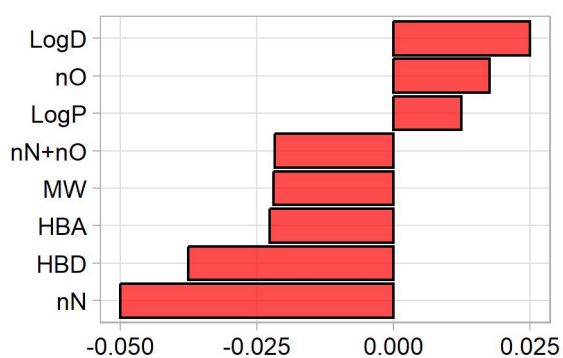
ANN		GPC		SVM	
Activation function	FC-1: Relu FC-2: Relu FC-3: Relu FC-4: Relu	Kernel	FC-1: RBF FC-2: RBF FC-3: RBF FC-4: RBF	Kernel	FC-1: RBF FC-2: RBF FC-3: RBF FC-4: RBF
N° of hidden layers	FC-1: 2 FC-2: 2 FC-3: 2 FC-4: 1	α	FC-1: 1 FC-2: 1 FC-3: 1 FC-4: 1	C	FC-1: 500 FC-2: 5000 FC-3: 500 FC-4: 500
N° of neurons by layer	FC-1: 80 FC-2: 80 FC-3: 80 FC-4: 60	Length scale (RBF)	FC-1: 0.05 FC-2: 0.02 FC-3: 0.05 FC-4: 0.05	γ	FC-1: 0.005 FC-2: 0.01 FC-3: 0.005 FC-4: 0.01

XGBr		XGBc		sLE	
N° of estimators	FC-1: 1000 FC-2: 2000 FC-3: 1000 FC-4: 2000	N° of estimators	FC-1: 1000 FC-2: 1000 FC-3: 2000 FC-4: 1000	γ_s	FC-1: 0.02 FC-2: 0.1 FC-3: 0.02 FC-4: 0.5
Max depth	FC-1: 10 FC-2: 10 FC-3: 10 FC-4: 10	Max depth	FC-1: 4 FC-2: 4 FC-3: 8 FC-4: 4	-	-
Learning rate	FC-1: 0.05 FC-2: 0.05 FC-3: 0.05 FC-4: 0.05	Learning rate	FC-1: 0.1 FC-2: 0.1 FC-3: 0.05 FC-4: 0.1	-	-

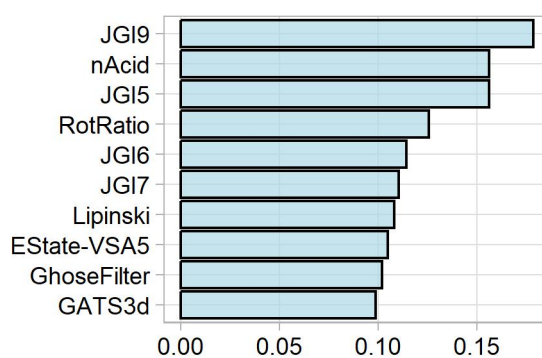
I APPENDIX I

Kendall's 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).

(a)



(b)



Source: Author's own.

J APPENDIX J

Summary of the best hyper-parameters reached by gridsearch applied in ML models (ANN, GPC, and SVM) that compose the Vcf-3BPP and the proposed DPF-3BPPred.

ANN		GPC		SVM	
Activation function	FC-1: Relu FC-2: Relu FC-3: Relu	Kernel	FC-1: RQ FC-2: RBF FC-3: RQ	Kernel	FC-1: Sigmoid FC-2: Sigmoid FC-3: Sigmoid
N° of hidden layers	FC-1: 2 FC-2: 1 FC-3: 1	α	FC-1: 0.5 FC-2: 1 FC-3: 1	C	FC-1: 500 FC-2: 2000 FC-3: 500
N° of neurons by layer	FC-1: 60 FC-2: 50 FC-3: 70	Length scale	FC-1: 0.5 FC-2: 0.01 FC-3: 0.5	γ	FC-1: 0.005 FC-2: 0.01 FC-3: 0.005

XGBr		XGBc		sLE	
N° of estimators	FC-1: 1000 FC-2: 1000 FC-3: 1000	N° of estimators	FC-1: 1000 FC-2: 1000 FC-3: 1000	γ_s	FC-1: 0.01 FC-2: 0.02 FC-3: 0.02
Max depth	FC-1: 10 FC-2: 10 FC-3: 10	Max depth	FC-1: 4 FC-2: 4 FC-3: 4	-	-
Learning rate	FC-1: 0.05 FC-2: 0.05 FC-3: 0.05	Learning rate	FC-1: 0.05 FC-2: 0.05 FC-3: 0.05	-	-

K APPENDIX K

Results reached in 10-fold cross-validation and independent test by DPF-3BPPred for all FCs and datasets. CV Avg: average accuracy in cross-validation, CV Std: standard deviation of the accuracy in cross-validation.

FC	Dataset	Gamma	CV Avg	CV Std	Accuracy	F1 Score	MCC	Precision	Recall	AUC
FC-1	Dataset 1	0.01	0.95	0.04	0.70	0.67	0.41	0.75	0.60	0.74
FC-1	Dataset 2	0.10	0.96	0.04	0.45	0.35	-0.10	0.43	0.30	0.29
FC-1	Dataset 3	0.01	0.94	0.04	0.75	0.71	0.52	0.86	0.60	0.64
FC-2	Dataset 1	0.01	0.99	0.02	0.70	0.73	0.41	0.67	0.80	0.65
FC-2	Dataset 2	0.50	0.99	0.02	0.60	0.50	0.22	0.67	0.40	0.70
FC-2	Dataset 3	0.05	0.99	0.02	0.80	0.82	0.61	0.75	0.90	0.75
FC-3	Dataset 1	0.05	0.98	0.03	0.90	0.91	0.82	0.83	1.00	0.75
FC-3	Dataset 2	0.05	0.99	0.02	0.80	0.82	0.61	0.75	0.90	0.73
FC-3	Dataset 3	0.20	0.98	0.04	0.80	0.78	0.61	0.88	0.70	0.82
FC-4	Dataset 1	0.02	0.99	0.02	0.75	0.74	0.50	0.78	0.70	0.70
FC-4	Dataset 2	0.20	1.00	0.00	0.65	0.63	0.30	0.67	0.60	0.68
FC-4	Dataset 3	0.05	0.99	0.02	0.85	0.84	0.70	0.89	0.80	0.85

Source: Author's own.

L APPENDIX L

Results reached in LOOCV DPF-3BPPred for all FCs and datasets. CV Avg: average accuracy in LOOCV, Std: standard deviation of the accuracy in LOOCV.

FC	Dataset	Gamma	Avg	Std
FC-1	Dataset 1	0.1	0.99	0.07
FC-1	Dataset 2	0.05	0.99	0.1
FC-1	Dataset 3	0.2	0.98	0.14
FC-2	Dataset 1	0.2	1	0
FC-2	Dataset 2	0.01	1	0
FC-2	Dataset 3	0.2	1	0
FC-3	Dataset 1	0.1	0.99	0.1
FC-3	Dataset 2	0.05	0.99	0.07
FC-3	Dataset 3	0.05	0.99	0.1
FC-4	Dataset 1	0.01	0.99	0.07
FC-4	Dataset 2	0.05	1	0
FC-4	Dataset 3	0.01	0.99	0.1

Source: Author's own.

M APPENDIX M

List of natural peptide sequences used to compare DPF-3BPPred with other online tools.

Table 9

ID	Peptide Name	Sequence	Label
2	Oxytocin	CYIQNCPLG	BBB+
8	Pituitary adenylate cyclase-activating polypeptide-27	HSDGIFTDSYSRYRKQMAVKKYLA AVL	BBB+
9	Pituitary adenylate cyclase-activating polypeptide-38	HSDGIFTDSYSRYRKQMAVKKYLA AVL GKRYKQRVKNK	BBB+
10	Vasoactive intestinal peptide	HSDAVFTDNYTRLRKQMAVKKYLSILN	BBB+
18	Deltorphan I	YAFDVVG	BBB+
19	Deltorphan II	YAFEVVG	BBB+
21	Adrenomedullin	YRQSMNQGSRSSTGCRFGTCTMQKLAHQYQFTDKDKDGMAPRNKISPQGY	BBB+
27	Amylin	KCNTATCATQRLANFLVRSSNGLPVLPTNVGSNTY	BBB+
44	Human Beta Defensin-1	DHYNCSVSSGGQCLYSACPIFTKIQTCTYRGKAKCK	BBB+
45	Human Beta Defensin-2	GIGDPVTCCLKSGAICHVPFCPRRYKQIGTCGLPGTKCKKP	BBB+
46	Delta sleep inducing peptide	WAGGDASGE	BBB+
47	Pancreatic Polypeptide	APLEPEYPGDNATPEQMAQYAAELRRYINMLTRPY	BBB+
59	Urocortin II	VILSLDVPIGLLRILLEQARYKAARNQAATNAQILAHV	BBB+
60	Urocortin III	FTLSLDVPTNIMNLLFNIAKAKNLRAQAAANAHLMAQI	BBB+
73	Enterostatin	VPDPR	BBB+
101	Substance P	RPKPQQFFGLM	BBB+
233	Neurotensin	ELYENLPRRPYIL	BBB+
N_C2	-	KLTRAQRRAAARKNKRNTRGC	BBB-
N_C3	-	GGAYVTRSSAVRLRSSVPGVRLQ	BBB-
N_C4	-	ARRRCSGSGSGCGSGSGSGRRR	BBB-
N_C8	-	YGRKKRRQRRTALDASALQTE	BBB-
N_C11	-	KSTGKANKITITNDKGRLSK	BBB-
N_C14	-	TVDNPASTTNKDKLFAVRK	BBB-
N_C17	-	LLHILRRSIRKQAHAIK	BBB-
N_C19	-	LNSAGYLLGKALAALAKKIL	BBB-
N_C20	-	CGRKKRWWRQRWWRWWRPPQ	BBB-
N_C26	-	MIIFRAAASHKK	BBB-
N_C33	-	RILQQLFIHFRIGCRHSRI	BBB-
N_C41	-	RQIKIWFQNR	BBB-
N_C43	-	CSSLDEPGRGGFSSESKEV	BBB-
N_C44	-	KLIKGRTPIKFGKADCDRPPKHSGK	BBB-
N_C45	-	ALWKTLLKKVLKAPKKRKV	BBB-
N_C46	-	KNAWKHSSCHHRHQI	BBB-
N_C54	-	LIRLWSHLIHWFNRRRLKWKKKC	BBB-

scientific reports



OPEN

Predicting cell-penetrating peptides using machine learning algorithms and navigating in their chemical space

Ewerton Cristhian Lima de Oliveira¹, Kauê Santana², Luiz Josino³,
Anderson Henrique Lima e Lima³ & Claudomiro de Souza de Sales Júnior¹

Cell-penetrating peptides (CPPs) are naturally able to cross the lipid bilayer membrane that protects cells. These peptides share common structural and physicochemical properties and show different pharmaceutical applications, among which drug delivery is the most important. Due to their ability to cross the membranes by pulling high-molecular-weight polar molecules, they are termed Trojan horses. In this study, we proposed a machine learning (ML)-based framework named BChemRF-CPPred (*beyond chemical rules-based framework for CPP prediction*) that uses an artificial neural network, a support vector machine, and a Gaussian process classifier to differentiate CPPs from non-CPPs, using structure- and sequence-based descriptors extracted from PDB and FASTA formats. The performance of our algorithm was evaluated by tenfold cross-validation and compared with those of previously reported prediction tools using an independent dataset. The BChemRF-CPPred satisfactorily identified CPP-like structures using natural and synthetic modified peptide libraries and also obtained better performance than those of previously reported ML-based algorithms, reaching the independent test accuracy of 90.66% (AUC = 0.9365) for PDB, and an accuracy of 86.5% (AUC = 0.9216) for FASTA input. Moreover, our analyses of the CPP chemical space demonstrated that these peptides break some molecular rules related to the prediction of permeability of therapeutic molecules in cell membranes. This is the first comprehensive analysis to predict synthetic and natural CPP structures and to evaluate their chemical space using an ML-based framework. Our algorithm is freely available for academic use at <http://comptools.linc.ufpa.br/BChemRF-CPPred>.

Peptides are a structurally diverse class of bioactive molecules with several physicochemical and structural properties^{1,2}. Naturally derived peptides have numerous pharmaceutical applications, such as acting selectively against pathogens^{3,4}, and human targets^{5,6}; and as cargo and delivery vehicles of covalently bound bioactive molecules, such as drugs, small-interfering RNAs (siRNAs), plasmids, and nanoparticles^{7–10}. Additionally, the recent advances in peptide synthesis have led to increased use in the pharmaceutical industry, because of their improved potency, specificity against molecular targets, and permeability to cell membranes^{11,12}.

The cell membrane is considered the main obstacle for therapeutic molecules to reach their active sites in cells. The selective control of the permeability of molecules through the cell membrane regulates passive diffusion and active transport to the intracellular medium impairing the entrance of some therapeutic compounds¹³. Cell-penetrating peptides (CPPs) can naturally cross the lipid bilayer membrane that protects the cells. These peptides share common structural and physicochemical features: they contain a sequence length between 5 and 42 amino acids, (2) they are soluble in water and partially hydrophobic, (3) they are often cationic (positive charge at physiological pH) or amphipathic, and (4) they are rich in the arginine and lysine residues^{14,15}. CPPs possess a wide range of biological activities, such as antiviral^{16,17}, antifungal¹⁸, and antibacterial activities^{19,20}, thus showing potential in pharmaceutical applications, but the main category has being drug delivery systems^{21–23}, and because they can cross the membranes pulling high molecular weight polar molecules, they are termed Trojan

¹Institute of Technology, Federal University of Pará, Belém, Pará 66075-110, Brazil. ²Institute of Biodiversity, Federal University of Western Pará, Vera Paz street, s/n Salé, Santarém, Pará 68040-255, Brazil. ³Laboratório de Planejamento e Desenvolvimento de Fármacos, Instituto de Ciências Exatas e Naturais, Universidade Federal do Pará, Belém, Pará 66075-110, Brazil. ✉email: kaue.costa@ufopa.edu.br; anderson@ufpa.br; claudomiro.sales@gmail.com

O APPENDIX N



Biological Membrane-Penetrating Peptides: Computational Prediction and Applications

OPEN ACCESS

Edited by:

Luciana Scotti,
Federal University of Paraíba, Brazil

Reviewed by:

Shaun Lee,
University of Notre Dame,
United States
Stéphanie Andrade,
University of Porto, Portugal

*Correspondence:

Kauê Santana da Costa
kauê.costa@ufpa.edu.br
Ewerton Cristhian Lima de Oliveira
ewerton.o43@gmail.com

*ORCID:

Ewerton Cristhian Lima de Oliveira
orcid.org/0000-0002-2338-7178
Kauê Santana da Costa
orcid.org/0000-0002-2735-8016
Paulo Sérgio Taube
orcid.org/0000-0001-5786-7615
Anderson H. Lima
orcid.org/0000-0002-8451-9912
Claudio de Souza de Sales Junior
orcid.org/0000-0002-2735-1383

[†]These authors have contributed
equally to this work

Specialty section:

This article was submitted to
Clinical Microbiology,
a section of the journal
Frontiers in Cellular and
Infection Microbiology

Received: 17 December 2021

Accepted: 21 February 2022

Published: 25 March 2022

Citation:

de Oliveira ECL, da Costa KS,
Taube PS, Lima AH and Junior CdSdS
(2022) Biological Membrane-
Penetrating Peptides: Computational
Prediction and Applications.
Front. Cell. Infect. Microbiol. 12:838259.
doi: 10.3389/fcimb.2022.838259

Ewerton Cristhian Lima de Oliveira^{1†}, Kauê Santana da Costa^{2††},
Paulo Sérgio Taube^{2†}, Anderson H. Lima^{3†} and Claudomiro de Souza de Sales Junior^{1†}

¹ Institute of Technology, Federal University of Pará, Belém, Brazil, ² Laboratory of Computational Simulation, Institute of Biodiversity, Federal University of Western Pará, Santarém, Brazil, ³ Laboratório de Planejamento e Desenvolvimento de Fármacos, Instituto de Ciências Exatas e Naturais, Universidade Federal do Pará, Belém, Brazil

Peptides comprise a versatile class of biomolecules that present a unique chemical space with diverse physicochemical and structural properties. Some classes of peptides are able to naturally cross the biological membranes, such as cell membrane and blood-brain barrier (BBB). Cell-penetrating peptides (CPPs) and blood-brain barrier-penetrating peptides (B3PPs) have been explored by the biotechnological and pharmaceutical industries to develop new therapeutic molecules and carrier systems. The computational prediction of peptides' penetration into biological membranes has been emerged as an interesting strategy due to their high throughput and low-cost screening of large chemical libraries. Structure- and sequence-based information of peptides, as well as atomistic biophysical models, have been explored in computer-assisted discovery strategies to classify and identify new structures with pharmacokinetic properties related to the translocation through biomembranes. Computational strategies to predict the permeability into biomembranes include cheminformatic filters, molecular dynamics simulations, artificial intelligence algorithms, and statistical models, and the choice of the most adequate method depends on the purposes of the computational investigation. Here, we exhibit and discuss some principles and applications of these computational methods widely used to predict the permeability of peptides into biomembranes, exhibiting some of their pharmaceutical and biotechnological applications.

Keywords: pharmacokinetics, machine learning, cell membrane, peptides, blood-brain barrier, structure activity, cell-penetrating peptides, drug system carriers

GETTING ACROSS THE BIOLOGICAL BARRIERS: AN OVERVIEW ON THE SCIENTIFIC SIGNIFICANCE AND CURRENT KNOWLEDGE

Penetration into biological membranes is a desired characteristic for bioactive molecules to reach their target site related to the molecular mode of action (Doak et al., 2014; Daina and Zoete, 2016). Molecules that naturally cross these biomembranes have been investigated aiming at different biotechnological and pharmaceutical applications (Rossi Sebastiano et al., 2018; Derakhshankhah

P APPENDIX O



pubs.acs.org/jcim

Article

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

Ewerton Cristhian Lima de Oliveira, Hannah Hirmz, Evelien Wynendaele, Juliana Auzier Seixas Feio, Igor Matheus Souza Moreira, Kauê Santana da Costa,* Anderson H. Lima, Bart De Spiegeleer,* and Claudomiro de Souza de Sales Júnior*

Cite This: <https://doi.org/10.1021/acs.jcim.3c00951>

Read Online

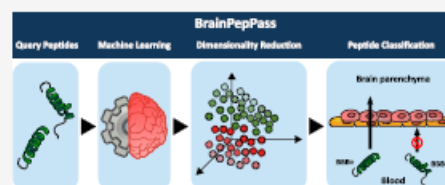
ACCESS |

Metrics & More

Article Recommendations

Supporting Information

ABSTRACT: Peptides that pass through the blood-brain barrier (BBB) not only are implicated in brain-related pathologies but also are 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 nonselective 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 ¹²⁵I-Met-enkephalin and ³H- 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 bioactive 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, coculture models, and triple-cell coculture models. These are

Special Issue: Machine Learning in Bio-cheminformatics

Received: June 23, 2023

Revised: November 6, 2023

Accepted: November 13, 2023



© XXXX American Chemical Society

A

<https://doi.org/10.1021/acs.jcim.3c00951>
J. Chem. Inf. Model. XXXX, XXX, XXX–XXX

Access in <https://pubs.acs.org/doi/10.1021/acs.jcim.3c00951>