論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Development of Prediction Models for Membrane Permeability and Plasma Protein Binding of Cyclic Peptides with Multi-Level Molecular Features by Deep Learning
著者(和文)	李佳男
Author(English)	Jianan Li
出典(和文)	学位:博士(工学), 学位授与機関:東京科学大学, 報告番号:甲第20号, 授与年月日:2024年12月31日, 学位の種別:課程博士, 審査員:秋山 泰,石田 貴士,岡﨑 直観,村田 剛志,関嶋 政和
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Institute of Science Tokyo, Report number:甲第20号, Conferred date:2024/12/31, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

Development of Prediction Models for Membrane Permeability and Plasma Protein Binding of Cyclic Peptides with Multi-Level Molecular Features by Deep Learning

Jianan Li

Graduate Major in Artificial Intelligence
Department of Computer Science
School of Computing
Institute of Science Tokyo

Supervisor: Yutaka Akiyama

A Thesis Submitted for the Degree of *Doctor of Engineering*

25 November 2024

Abstract

Recently, cyclic peptides have been considered breakthrough therapeutic agents that boast high binding affinity, minimal toxicity, and the potential to engage "undruggable" targets such as intracellular protein–protein interactions. However, the pharmaceutical utility of cyclic peptides is limited by their low membrane permeability and plasma protein binding (PPB) rate, which are essential indicators of oral bioavailability and intracellular targeting. Current machine learning-based prediction methods of cyclic peptide permeability and PPB rate show variable performance owing to the limitations of experimental data. Furthermore, these methods use features derived from the whole molecule that have traditionally been used to predict small molecules and ignore the unique structural properties of cyclic peptides.

In this dissertation, we proposed a novel multi-level molecular features representation method and model architectures to concurrently capture the local sequence variations and global structural changes in cyclic peptides. By incorporating hierarchical peptide-, monomer-, and atom-level information, the fusion model effectively captures the complex structure of cyclic peptides. Additionally, due to the inherent data scarcity in biological datasets, especially for cyclic peptides, we also proposed data augmentation strategies for cyclic peptides to improve model training efficiency.

For the membrane permeability prediction of cyclic peptides, we first curated a comprehensive dataset (CycPeptMPDB) containing 7,334 data points from over 40 published papers and pharmaceutical company patent documents. Then, we constructed a permeability prediction model, CycPeptMP, based on the proposed fusion model. CycPeptMP (MAE = 0.355, R = 0.883) outperformed several baseline models, including traditional ML-based cyclic peptide permeability prediction approaches (MAE = 0.418–0.488, R = 0.781–0.834), as well as state-of-the-art DL-based small molecule property prediction methods (MAE = 0.443–0.591, R = 0.660–0.825). Ablation studies demonstrated that all feature levels contributed and were relatively essential for prediction, consistent with physicochemical findings that both global and local structural changes in cyclic peptides significantly affect membrane permeability.

For the PPB rate prediction, we collected experimental data from collaborations with pharmaceutical companies and published literature. The fusion model performed best on the internal test set (MAE = 2.44%, R = 0.973), while the monomer model (CycPeptPPB) showed strong generalization on the external DrugBank dataset (MAE =

4.40%, R = 0.947), consistent with physicochemical findings that the substructure significantly affects the PPB rate because it forms a specific bond with plasma proteins. We also confirmed that our CycPeptPPB model accurately identified key monomers that greatly influenced PPB rates.

Acknowledgements

I would like to express my sincere gratitude to Professor Yutaka Akiyama for providing me with an excellent research environment and continuous guidance and encouragement throughout my research work consistently for seven years. I would also like to express my sincerest appreciation to Assistant Professor Keisuke Yanagisawa for providing kind guidance and discussing my research. Professor Takashi Ishida and Associate Professor Masahito Ohue also provided valuable suggestions and support during joint seminars held by Akiyama, Ishida, and Ohue laboratories, among others. I also thank Dr. Masatake Sugita, Mr. Takuya Fujie, and Mr. Rin Sato, as well as everyone in Akiyama, Ishida, and Ohue laboratories, who gave me considerable suggestions for my research work. I would also like to thank Ms. Kanako Ozeki and Ms. Masami Fujii, the secretaries of Akiyama laboratory, for their much-appreciated support during my seven years of laboratory life. Finally, I would like to thank my classmates in Akiyama laboratory, Satoshi Sugiyama and Shunpei Matsuno, who supported and encouraged me for three years.

I would like to thank the Japan Science and Technology Agency (JST) SPRING Program Fellow, the Japan Society for the Promotion of Science (JSPS) Research Fellow (DC2), the Program for Building Regional Innovation Ecosystems "Program to Industrialize an Innovative Middle Molecule Drug Discovery Flow through Fusion of Computational Drug Design and Chemical Synthesis Technology" from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), and the Platform Project for Supporting Drug Discovery and Life Science Research (Basis for Supporting Innovative Drug Discovery and Life Science Research (BINDS)) from the Japan Agency for Medical Research and Development (AMED) for financial support throughout my research work. Moreover, this study was supported in part by a Grant-in-Aid for Scientific Research (B) 17H01814, a Grant-in-Aid for Scientific Research (B) 20H04280, a Grant-in-Aid for Early-Career Scientists 20K19917, a Grant-in-Aid for Scientific Research (B) 22H03684, a Grant-in-Aid for Scientific Research (B) 23H03495, a Grant for JST SPRING program Fellow JPMJSP2106, and a Grant-in-Aid for JSPS Fellow

23KJ0891.

Last but certainly not least, I would like to extend a special thanks to my family, Ning Li and Lin Zhou, and my friend, Shan Wang. Without their kind support and constant encouragement, this work would not have been possible. Finally, I dedicate this work to my family.

Contents

Abstract **ii**

Acknowledgements **v**

1 General Introduction of Cyclic Peptide Drug Discovery **1**

1.1 Drug Development Process 1
1.2 Major Drug Modalities 2
1.2.1 Small molecule drugs 2
1.2.2 Antibody drugs 4
1.2.3 Conventional middle molecule drugs 5
1.3 Cyclic Peptide Drugs 6
1.3.1 Overview of cyclic peptide drugs 6
1.4 Remaining Two Major Challenges for Cyclic Peptide Drug Discovery . 8
1.4.1 Cell membrane permeability 10
1.4.2 Plasma protein binding (PPB) 14
1.4.3 Limitation of existing computational prediction methods of cyclic peptide membrane permeability and PPB rate 17
1.5 Overview of Major Deep Learning Architecture 18
1.6 Purpose of Study 20
1.7 Summary of Contributions 20
1.8 Thesis Organization 21

2 Overview of Cyclic Peptide Membrane Permeability and PPB Rate Prediction Methods **23**

2.1 Introduction 23
2.2 Cyclic Peptide Membrane Permeability Prediction Methods 24
2.2.1 MD-based methods for cyclic peptide membrane permeability prediction 24

2.2.2 ML-based methods for cyclic peptide membrane permeability prediction . . . 25
2.3 Cyclic Peptide PPB Rate Prediction Methods . . . 27
2.3.1 Summary of small molecule PPB rate prediction methods . . . 27
2.3.2 Cyclic peptide PPB rate prediction based on small molecule data 27
2.4 Overview of DL-based Small Molecule Properties Prediction Methods and Potential Application to Cyclic Peptides . . . 28
2.5 Summary . . . 29

3 Multi-Level Molecular Features Design and Data Augmentation for Cyclic Peptides 31
3.1 Introduction . . . 31
3.2 Overall of the Proposed Fusion Model . . . 32
3.3 Peptide-Level Feature Design . . . 34
3.3.1 Conformation generation of peptides . . . 34
3.3.2 Descriptor calculation of peptides . . . 35
3.3.3 Preprocessing and selection of peptide descriptors . . . 35
3.3.4 Architecture of the peptide model . . . 37
3.4 Monomer-Level Feature Design . . . 37
3.4.1 Division of the main chain of cyclic peptide into monomers . . . 39
3.4.2 Conformation generation and descriptor calculation of monomers 39
3.4.3 Architecture of the monomer model . . . 39
3.5 Atom-Level Feature Design . . . 42
3.5.1 Atom-level features calculated from molecular graph representation 42
3.5.2 Architecture of the atom model . . . 42
3.6 Architecture of the Fusion Model . . . 45
3.7 Data Augmentation for Cyclic Peptides . . . 46
3.8 Summary . . . 50

4 Development of a Membrane Permeability Prediction Model of Cyclic Peptides (CycPeptMP) 51
4.1 Introduction . . . 51
4.2 Development of a Comprehensive Database of Membrane Permeability of Cyclic Peptides (CycPeptMPDB) . . . 52
4.2.1 Data collection . . . 52
4.2.2 Sequence representation of cyclic peptides . . . 55

4.2.3 Monomer definition in CycPeptMPDB 56
4.2.4 Future update schedule for CycPeptMPDB 58
4.3 Materials and Methods 59
4.3.1 Experimental dataset 59
4.3.2 Evaluation metrics 62
4.3.3 Descriptors selection 63
4.3.4 Hyperparameter search 67
4.3.5 Baseline methods 71
4.4 Results and Discussion 75
4.4.1 Performance comparison for the test set 75
4.4.2 Application of trained PAMPA model to other assay data 80
4.4.3 Performance comparison for 10-fold cross-validation 80
4.4.4 Comparison to DL-based methods based on CycPeptMPDB . . 81
4.4.5 Ablation study of atom and monomer models 84
4.4.6 Ablation study of the fusion model 85
4.4.7 Performance using regenerated conformations 87
4.4.8 Comparison with MD-based method 87
4.5 Summary 90

5 Development of a PPB Rate Prediction Model of Cyclic Peptides (CycPeptPPB) 91
5.1 Introduction 91
5.2 Materials and Methods 92
5.2.1 Experimental dataset 92
5.2.2 Descriptors selection 97
5.2.3 Hyperparameter search of proposed models 100
5.2.4 Baseline methods 103
5.3 Results and Discussion 105
5.3.1 Performance comparison 105
5.3.2 Ablation study for the fusion and monomer models 110
5.3.3 Analysis of prediction results of peptide pairs in the DrugBank set with similar main structures 113
5.4 PPB Rate Prediction of Cyclic Peptides by Docking Simulation 117
5.4.1 3D structures of HSAs and cyclic peptides 117
5.4.2 Overview of docking simulation 119
5.4.3 Results of docking simulation 121

5.4.4 Prediction based on glide descriptors 124
5.5 Summary 125

6 Overall Discussion **127**
6.1 Descriptor Selection 127
6.1.1 Common important descriptors in membrane permeability and PPB rate prediction 127
6.1.2 Integrated selection of 2D and 3D descriptors 131
6.1.3 Alternative algorithms for descriptor selection 134
6.2 Conformation Generation of Cyclic Peptides by MD Simulation 139
6.2.1 Simulation details 139
6.2.2 Analysis of descriptor differences between RDKit and MD conformations 140
6.2.3 Comparison of RDKit and MD conformations in PCA space . . 141
6.3 Conformation Generation of Monomers 145
6.4 Effect of Atom-Level Features 149
6.5 Effect of Single-Level Data Augmentation 152

7 Conclusion **155**
7.1 Conclusion 155
7.1.1 Contributions 155
7.2 Future Works 158
7.2.1 Technical improvements of proposed methods 158
7.2.2 Challenges in cyclic peptide drug discovery 159

A Full List of FDA-Approved Macrocyclic Drugs **163**

B Full List of Calculated $1,857$ Descriptors **177**

C Development of a Comprehensive Database of Membrane Permeability of Cyclic Peptides (CycPeptMPDB) **183**
C.1 Overview of CycPeptMPDB Framework 183
C.2 3D Structure Generation of Cyclic Peptides 184
C.3 Introduction of Web Page Functions 185
C.3.1 Peptide browsing function 185
C.3.2 Peptide search function for quick data retrieval 185
C.3.3 Visualization functions on peptide detail page 187

C.3.4 Browsing and visualization functions of monomers 189
C.4 Data and Software Availability 189

References **194**

List of Publications **217**

List of Figures

1.1 The traditional process of drug development, and the cycle time and cost at each step to successfully discover a single new molecular entity (created based on [1, 2]). 2

1.2 Structures and molecular weights of the small molecule drug Aspirin, the linear peptide drug Pluvicto, the cyclic peptide drug Cyclosporin A, and the antibody drug Rituximab. 4

1.3 Examples of natural head-to-tail cyclic peptides (cited from [3]). From the left, Cyclosporin A, and STF-1 and RTD-1 with disulfide bridges are shown. 6

1.4 Various macrocyclization approaches of cyclic peptides (cited from [4]). 7

1.5 FDA-approved macrocyclic drugs ($n = 67$) (cited from [5]). (A) Orally available drugs are shown in blue ($n = 26$; 39%), while those administered parenterally are shown in gold ($n = 41$; 61%). (B) Natural products and their derivatives are indicated in light green ($n = 59$; 88%); *de novo* designed macrocyclic drugs, which are synthetically created from scratch rather than derived from existing natural compounds, are indicated in dark gray ($n = 8$; 12%). 9

1.6 Therapeutic indications (inner circle) and targets (outer circle) of the FDA-approved macrocyclic drugs ($n = 72$) (cited from [5]). Five macrocycles are duplicated because each is used for two therapeutic indications. 10

1.7 Image of the membrane permeation process by passive diffusion. The process of passive diffusion is driven by a compound's net concentration gradient, flowing from a high-concentration region (extracellular) to a low-concentration region (intracellular). 11

1.8 Image of the dynamic conformation changes of Cyclosporin A when transitioning between aqueous and hydrophobic environments (cited from [6]). Cyclosporin A exists in an "open"-conformation inside the aqueous extra- and intracellular environments but adopts a "closed"-conformation containing four intramolecular hydrogen bonds shown as dotted lines upon entering the lipid bilayer. 12
1.9 Examples of modification strategies to improve membrane permeability: backbone N-methylation [7], conformational control [8], amide-to-ester substitution [9], amide-to-thioamide substitution [10], and side-chain modification [11]. 13
1.10 Image of drug binding to plasma protein. 15
1.11 Cocrystal structure of cyclic peptide drug dalbavancin and HSA (gray) (PDB ID: 6M5E). The hydrocarbon side chain of dalbavancin is inserted deeply into the hydrophobic pocket of HSA. 16
1.12 The relationship between chapters. 22

2.1 Image of cyclic peptide across a lipid bilayer during MD simulation. . . 25

3.1 Overall framework of the proposed fusion model. The three-level expression vectors extracted using the three sub-models are concatenated and passed through a shared layer to derive the final permeability prediction value. 33
3.2 Architecture of the peptide model. Peptide descriptors and Morgan FP were used as input for the MLP-based model. 37
3.3 Example of the monomer division: Cyclosporin A and its 11 monomers. 38
3.4 Comparison of the conventional convolution method (1D-CNN) and the proposed convolution method (CyclicConv). In the case of kernel size of three, monomer C is supplemented to the left of A and monomer A is supplemented to the right of C. By this operation, the information of CAB and BCA can be correctly acquired as a result of the CyclicConv. 40
3.5 Architecture of the monomer model. Monomer descriptors are aligned based on the sequence information and used as input for the CNN-based model. 41
3.6 Atom model architecture. Node features and three types of node-pair relative relationship matrices are used as input for the transformer-based model. 43

3.7 Peptide- and monomer-levels data augmentation based on using multiple conformations. We use 60 different conformations per peptide/monomer to incorporate more diverse 3D information. 47

3.8 Monomer-level data augmentation based on sequence arrangement. The aligned monomer descriptors are translated and rotated based on the sequence information. The number of replicas of a cyclic peptide consisting of n monomers is $n \times (max\_len(16) - n + 1)$. 48

3.9 Atom-level data augmentation based on SMILES enumeration. Molecular graphs with different atoms order are constructed from different SMILES representations. 49

4.1 Permeability (LogP_{exp}) distribution of all peptides. The background color of the high permeability ($\text{LogP}_{\text{exp}} \geq -6.0$; 5,113 peptides) range is yellow, and the background color of the low permeability ($\text{LogP}_{\text{exp}} < -6.0$; 2,338 peptides) range is green. 54

4.2 (A) Example of HELM notation (PEPTIDE1{[dL].[dL].L.[dL].P.Y} $PEPTIDE1, PEPTIDE1,1:R1-6:R2$$$) and its constituent parts in CycPeptMPDB. If the simple polymer is a peptide, write the simple polymer as PEPTIDEx (where x is a number, and in the case of RNA is RNAx). The connection section means that R1 of the 1st monomer of PEPTIDE1 and R2 of the 6th monomer of PEPTIDE1 are connected. The hydrogen bonds and attributes sections of all peptides in this study are empty. (B) Example of monomer definition of tyrosine (Y). 55

4.3 LogP distribution of all monomers. The background color for extremely hydrophilic monomers ($\text{LogP} < -0.60$, lower than G:-0.60; 35 monomers) is blue, hydrophilic monomers ($-0.60 \leq \text{LogP} < 0.40$, lower than V:0.43, general hydrophilic amino acids, such as G:-0.60, A:-0.21, and P:0.28; 66 monomers) is light blue, hydrophobic monomers ($0.40 \leq \text{LogP} < 1.40$, general hydrophobic amino acids, such as V:0.43, I:0.82, L:0.82, and F:1.02; 127 monomers) range is orange, and extremely hydrophobic monomers ($1.40 \leq \text{LogP}$, extremely hydrophobic amino acids, such as W:1.50; 84 monomers) is red. 57

4.4 Experimental data distribution. (A) The logarithm of experimentally determined membrane permeability ($LogP_{exp}$). (B) Molecular weight (MolWt descriptor calculated by RDKit). Valid–1 is the dataset used for the first-time evaluation of the validation set; the corresponding training data sets are Train, Valid–2, and Valid–3. 60
4.5 Experimental data distribution in PCA space, with the first principal component (PC1) as the horizontal axis and the second principal component (PC2) as the vertical axis; the contribution rates are shown in the parentheses of axes captions. (A) Distribution of training and validation sets. (B) Distribution of the test set. High and low indicate data with $LogP_{exp} \geq -6.0$ and $LogP_{exp} < -6.0$, respectively. 61
4.6 Top 15 peptide descriptors with the highest RF feature importance from (A) RF model with 2D descriptors and (B) RF model with 3D descriptors, respectively. Selected seven 2D and nine 3D peptide descriptors are shown in black. 64
4.7 Heatmap of absolute correlation coefficient values for 16 selected peptide descriptors. The pair with the highest correlation is MolLogP and logP(o/w) ($|R| = 0.884$). 65
4.8 Top 14 hyperparameters with an Optuna importance > 0.01 on the CycPeptMP hyperparameter search. Optuna importance is calculated based on the fANOVA hyperparameter importance evaluation algorithm [12]; the sum of the importance values is normalized to 1.0. 70
4.9 Prediction results of the test set by CycPeptMP. The predicted values of the test set are the average value of three runs. 76
4.10 Prediction results of the test set by (A) RF, (B) SVM-2D, and (C) SVM-2D3D models. The predicted values of the test set are the average value of three runs. 77
4.11 Prediction results of the test set by (A) MAT and (B) SAT models. The predicted values of the test set are the average value of three runs. . . . 78
4.12 Prediction results of the test set by (A) PharmHGT and (B) FinGAT models. The predicted values of the test set are the average value of three runs. 78
4.13 Ablation results (MAE) for the atom and monomer models using the test set. 84
4.14 Ablation results (MAE) for the fusion model using the test set. (A) Different numbers of input replicas. (B) Different architectures. 86

4.15 (A) Prediction results of the MD-based method. Black dots represent hydrophobic peptides with AlogP ≥ 4; green dots represent the remaining peptides with AlogP < 4. (B) Prediction results of CycPeptMP. . . 89

5.1 Distributions of experimental data. (A) Objective variable $\%PPB_{50-95}$. Data of $\leq 50\%$ is included in the leftmost bar, and data of $\geq 95\%$ is included in the rightmost bar. (B) Molecular weight (MolWt descriptor calculated by RDKit). Blue, orange, and green bars indicate PD (PeptiDream), Tajimi, and DrugBank datasets, respectively. 93

5.2 Experimental data distribution in PCA space, with the PC1 as the horizontal axis and the PC2 as the vertical axis; the contribution rates are shown in the parentheses of axes captions. 96

5.3 Top 15 peptide descriptors with the highest RF feature importance from (A) RF model with 2D descriptors and (B) RF model with 3D descriptors, respectively. Selected three 2D and two 3D peptide descriptors are shown in black. 98

5.4 Heatmap of absolute correlation coefficient values for five selected peptide descriptors. The pair with the highest correlation is vsurf_CW2 and vsurf_CW3 ($|R| = 0.795$). 99

5.5 Top ten hyperparameters with an Optuna importance > 0.01 on the CycPeptMP hyperparameter search. Optuna importance is calculated based on the fANOVA hyperparameter importance evaluation algorithm [12]; the sum of the importance values is normalized to 1.0. 102

5.6 Prediction results of the test set by four baselines and four proposed methods. (A) ADMET Predictor, (B) RF model, (C) SVM-2D model, (D) SVM-2D3D model, (E) atom model, (F) monomer model (1D-CNN), (G) monomer model (CyclicConv), and (H) fusion model. Except for the ADMET Predictor, the predicted values are the average value of three runs. 108

5.7 Prediction results of the DrugBank set by four baselines and four proposed methods. (A) ADMET Predictor, (B) RF model, (C) SVM-2D model, (D) SVM-2D3D model, (E) atom model, (F) monomer model (1D-CNN), (G) monomer model (CyclicConv), and (H) fusion model. Except for the ADMET Predictor, the predicted values are the average value of three runs. 109

5.8 Ablation results (%MAE) for different input replica numbers for the fusion and 1D-CNN monomer models. (A) Results on the internal test set. (B) Results on the external test (DrugBank) set. 111

5.9 Ablation results (%MAE) for different architectures for the fusion and 1D-CNN monomer models. (A) Results on the internal test set. (B) Results on the external test (DrugBank) set. 112

5.10 The monomers obtained by the decomposition procedure and these salience scores of acetyl-daptomycin (Exp.%PPB $\leq$ 50.0%, Pred.%PPB = 50.5%) and daptomycin (Exp.%PPB = 91.5%, Pred.%PPB = 91.9%). The salience score shows the average value of the salience score of all augmentation replicas across three repeated runs (normalized to a maximum value of 1). 114

5.11 Docking results of daptomycin with Site 1 of HSA (cited from [13], added black dotted circle line). The N-terminal fatty acid side chain of daptomycin penetrates deeply into the binding pocket Site 1 of HSA and has extensive and strong hydrophobic interactions with the HSA residues Leu219, Leu260, Ile264, Ile290, and Ala291. 115

5.12 The monomers obtained by the decomposition procedure and these salience scores of colistin (Exp.%PPB = 56.0%, Pred.%PPB = 55.2%) and polymyxin b (Exp.%PPB = 85.5%, Pred.%PPB = 74.0%). The salience score shows the average value of the salience score of all augmentation replicas across three repeated runs (normalized to a maximum value of 1).116

5.13 Results of the superposition of four different HSA structures used. . . . 118

5.14 Coordinates and grid boxes of Site 1 to Site 6 of 1N5U (green: internal box, purple: external box). 120

5.15 Distribution of Site 1, Site 2 docking scores (average of maximum five poses), and objective variables $\%PPB_{50-95}$. 123

6.1 Distribution of logP(o/w) with (A) permeability $LogP_{exp}$ and (B) PPB rate $\%PPB_{50-95}$. 128

6.2 Distribution of logP(o/w) with permeability $LogP_{exp}$ on the data from (A) 2013_CHUGAI [14] and (B) 2016_Furukawa [15]. 130

6.3 Top 16 peptide descriptors with the highest RF feature importance for (A) permeability prediction and (B) PPB rate prediction. The 2D and 3D descriptors are shown in light blue and orange, respectively. 132

6.4 Comparison of six MOE 3D descriptors of 1NMe3 (CycPeptMPDB ID: 2328) calculated from MD simulation conformations (red) and RDKit conformations (light blue). (A) E, (B) dens, (C) FASA-, (D) FCASA+, (E) FASA_P, and (F) vsurf_Wp2. 142

6.5 Comparison of six MOE 3D descriptors of Cyclosporin A (CycPeptMPDB ID: 7353) calculated from MD simulation conformations (red) and RDKit conformations (light blue). (A) E, (B) dens, (C) FASA-, (D) FCASA+, (E) FASA_P, and (F) vsurf_Wp2. 143

6.6 Comparison of RDKit and MD conformations of (A) 1NMe3 and (B) Cyclosporin A in PCA space, with the PC1 as the horizontal axis and the PC2 as the vertical axis; the contribution rates are shown in the parentheses of axes captions. 144

6.7 Comparison of current (top) and ACE-NME (bottom) capping methods when dividing leucine. Where ACE is an N-terminal acetyl group, and NME is a C-terminal N-methyl group. 146

6.8 Structure and comparison of conformations generated with two capping methods in PCA space of (A) Leucine, (B) Pye, and two monomers, Sub25 (C) and Sub27 (D), which contain large side chain portions of the Lariat peptide. 147

6.9 Distribution of the number of heavy atoms of (A) permeability and (B) PPB datasets. 150

6.10 Ablation study results (MAE) for membrane permeability prediction with the fusion model on the test set, showing performance across varying numbers of input replicas at each augmentation level: (A) Peptide-level, (B) Monomer-level, and (C) Atom-level. Note that for each level, the other two levels use duplicated inputs. 153

A.1 Structure of all 67 FDA-approved macrocyclic drugs. Structures and molecular weights are from PubChem and ChEMBL databases. 168

C.1 Basic framework of CycPeptMPDB. CycPeptMPDB data were collected from published papers and patents of pharmaceutical companies and then manually inspected. Information in various formats was deposited into a PostgreSQL-based database for various web-based functions. . . 184

C.2 (A) Classification method selection for browsing peptides and browsing page. The case when Monomer Length is selected is shown as an example. (B) Peptide's list page. The background color of the permeability cell is yellow when the permeability is High ($\text{LogP}_{\text{exp}} \geq -6.00$) and green when it is Low ($\text{LogP}_{\text{exp}} < -6.00$). 186
C.3 (A) Peptide information section of the peptide detail page. (B) The structural information section of the peptide detail page. HELM images and LogP transition diagrams are colored by the LogP value of each monomer. 188
C.4 Monomer (A) browsing and (B) list pages. LogP cell background color is blue when LogP is Extremely Hydrophilic ($\text{LogP} < -0.60$), light blue when Hydrophilic ($-0.60 \leq \text{LogP} < 0.40$), orange when Hydrophobic ($0.40 \leq \text{LogP} < 1.40$), and red when Extremely Hydrophobic ($1.40 \leq$ LogP). 190
C.5 (A) Monomer information section of the monomer detail page. (B) Statistics section of peptides containing current monomer. 191

List of Tables

1.1 Characteristics of major drug modalities: small molecules, linear peptides, cyclic peptides, and antibodies (created based on [16]). 3

2.1 Summary of ML-based permeability prediction methods for cyclic peptides. The data type abbreviations are as follows: CP (cyclic peptides) and MC (macrocycles). The model type abbreviations are as follows: SLR (simple linear regression), MLR (multiple linear regression), PLS (partial least squares), SVM (support vector machine), and RF (random forest). The accuracy abbreviations are as follows: R^2 (coefficient of determination) and RMSE (root mean square error). 26

3.1 Summary of feature selection methods in ML. 36
3.2 Summary of atom-level features. 42

4.1 Source literature list of CycPeptMPDB. The number of peptides, molecular weight range, and assay type of membrane permeability for each source are shown. 53
4.2 Explanation of symbols naming method. 57
4.3 List of source literature for CycPeptMPDB updates. The number of peptides and assay type of membrane permeability for each source are shown. 59
4.4 Description of selected descriptors, arranged in order of RF feature importance. 66
4.5 Search target and range of the hyperparameter search for the proposed fusion model. Hyperparameters marked with * are not subject to search and used fixed values. 68
4.6 Hyperparameter search range and its results for the fusion model-based membrane permeability prediction model (CycPeptMP). 69

4.7 Search range and results of the hyperparameter search (grid search) for the RF and two SVM models. 72
4.8 Search range and results of the hyperparameter search for the MAT and SAT models. Hyperparameters were determined by 150 trials using Optuna software based on the average MSE of three runs. 73
4.9 Search range and results of the hyperparameter search for the PharmHGT and FinGAT models. Hyperparameters were determined by 150 trials using Optuna software based on the average MSE of three runs. 74
4.10 Performance comparison between seven baseline methods and CycPeptMP using the test set. The metrics are the averaged values of three repeated runs; the best result for each metric is indicated in bold. 75
4.11 Prediction performance between seven baseline methods and CycPeptMP (models were trained with PAMPA) for Caco-2, MDCK, and RRCK permeabilities recorded in CycPeptMPDB. The metrics are averaged for three runs. 79
4.12 Performance comparison between seven baseline methods and CycPeptMP by 10-fold cross-validation. The metrics are the averaged values of ten repeated runs; the best result for each metric is indicated in bold. . . . 80
4.13 Performance comparison between six DL-based methods and CycPeptMP. The best result for each metric is indicated in bold. 83
4.14 Prediction performance of CycPeptMP of the test set using 3D conformations regenerated by RDKit (five times with different seeds). The metrics are averages of three runs. 87
4.15 Peptides used in comparison with the MD-based method. The 23 peptides are included in the validation and test sets of this study. The AlogP and MD predicted values ($\log P_{ISMD\_mod}$) are reported by Sugita *et al* [17]. 88

5.1 Amino acid sequence and experimental %PPB of the Tajimi dataset. . 94
5.2 %PPB reported in prior research and our surveyed %PPB (1/2/2021 accessed; experimental values used in this study) of the DrugBank dataset. If multiple %PPB values are listed, the average value is used, and the original range is shown in the parentheses. 95
5.3 Description of selected descriptors, arranged in order of RF feature importance. 97

5.4 Hyperparameter search results of the four proposed models. "-" indicates that the hyperparameter does not apply to the respective model. 101
5.5 Search range and results of the hyperparameter search (grid search) for the RF and two SVM models. 104
5.6 Performance comparison between four baseline methods and four proposed models using the test set. Except for the ADMET Predictor, the metrics are the averaged values of three repeated runs; the best result for each metric is indicated in bold. 105
5.7 Performance comparison between four baseline methods and four proposed models using the DrugBank set. Except for the ADMET Predictor, the metrics are the averaged values of three repeated runs; the best result for each metric is indicated in bold. 106
5.8 Number of binding poses obtained from each site docking of 1N5U for each number of residues (numbers in parentheses indicate the number of peptides). 121
5.9 Total calculation time (CPU core time) for each site docking of 1N5U for each number of residues (numbers in parentheses indicate the number of peptides). 122
5.10 Summary of used 12 glide descriptors. 124
5.11 Performance comparison between three baseline methods and two glide models using the glide test set. The metrics of three baseline methods are calculated from the averaged prediction values of three repeated runs; the best result for each metric is indicated in bold. 124

6.1 Selected descriptors and their RF feature importance for permeability and PPB rate prediction. 129
6.2 Performance comparison between three SVM models for permeability (using the test set) and PPB rate prediction (using the test set and DrugBank set). The metrics are the averaged values of three repeated runs; the best result for each metric is indicated in bold. 133
6.3 Descriptor selection results for permeability prediction by Bolasso when changing the Lasso hyperparameter α. Selected 2D and 3D descriptors and their frequency (up to $k \times m = 60$ times) are shown, respectively. Finally, we used six 2D descriptors (MATS2i, Xch-7d, VSA_EState9, fr_piperdine, AATSC4c, and BCUT_PEOE_0) and seven 3D descriptors (FASA-, FNSA4, vsurf_Wp4 FASA+, FNSA3, vsurf_R, and dens). . . . 135

6.4 Descriptor selection results for PPB rate prediction by Bolasso when changing the Lasso hyperparameter α. Selected 2D and 3D descriptors and their frequency (up to $k \times m = 60$ times) are shown, respectively. Finally, we used four 2D descriptors (logP(o/w), PEOE_VSA-1, EState_VSA3.1, and AATSC4se) and three 3D descriptors (vsurf_D5, vsurf_CW3, and vsurf_CW1). 136
6.5 Performance comparison between four SVM models for permeability (using the test set) and PPB rate prediction (using the test set and DrugBank set). The metrics are the averaged values of three repeated runs; the best result for each metric is indicated in bold. 138

A.1 Full therapeutic indications and target classification for the FDA-approved macrocyclic drugs dataset ($n = 72$) (cited from [5]). Five macrocycles (macrocycle names in bold) are duplicated because each is used in two therapeutic indications. The table is ordered by therapeutic indication (alphabetically) and then by target (alphabetically). Complete target names are reported. NA: Target not available. 164

B.1 Full list of calculated 206 MOE 2D descriptors. 178
B.2 Full list of calculated 117 MOE 3D descriptors. 180
B.3 Full list of calculated 208 RDKit 2D descriptors. 181

Chapter 1

General Introduction of Cyclic Peptide Drug Discovery

1.1 Drug Development Process

Drugs must follow strict regulations to ensure high efficacy, safety, and stability. As shown in Fig. 1.1, developing a new drug is a long process, taking over 10 years and costing over $1.5 billion [1]. The traditional process of drug development can be mainly categorized into three parts:

1. **Discovery and Development**: The aim is to obtain potential drug candidates that can interact with specific targets, and this part is further comprised of the following three steps.

   (a) **Target selection**: A step to identifying a target, such as a nucleic acid sequence or protein involved in a specific disease. This step also confirms that the target is "drug discoverable," i.e., its activity can be modulated by exogenous compounds.

   (b) **Compound screening**: A step to finding lead compounds that can interact with the target from a huge compound library through screening.

   (c) **Lead optimization**: A step to modifying compounds' structure to improve their pharmacological properties, such as activity, bioavailability, and safety.

   Many efforts have been made to accelerate these steps using computational methods instead of wet experiments.

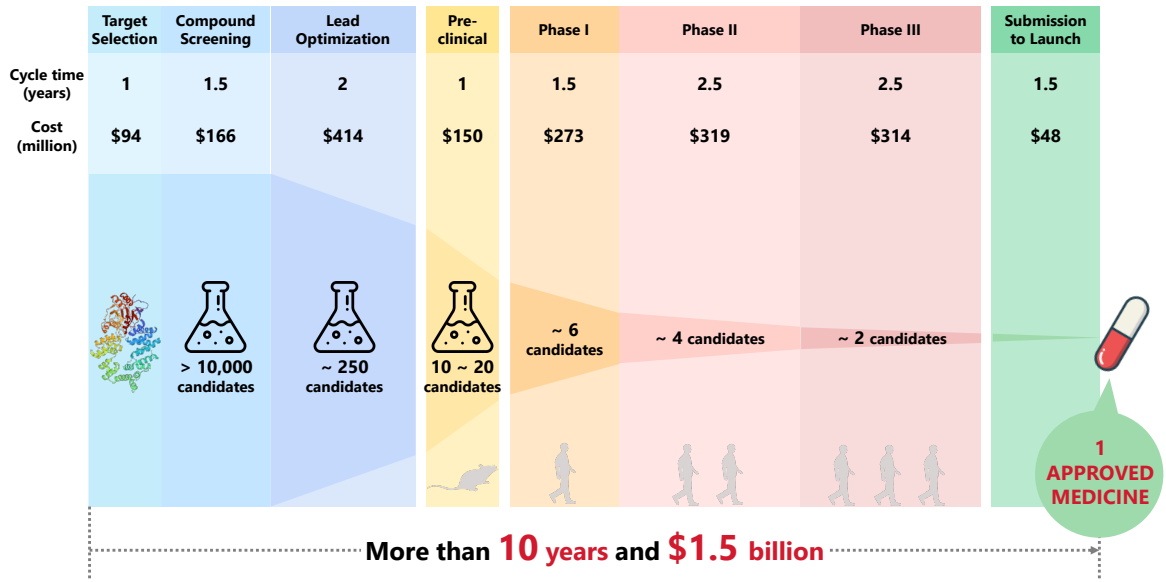


Figure 1.1: The traditional process of drug development, and the cycle time and cost at each step to successfully discover a single new molecular entity (created based on [1, 2]).

2. **Preclinical Research**: Drug candidates are tested in extensive cell and animal experiments to confirm their efficacy and safety.

3. **Clinical Research**: Drug candidates' efficacy, dosage, and safety in humans are evaluated. This part is divided into three phases depending on the number of subjects: Phase I involves 20 to 80 subjects, Phase II involves 100 to 300 subjects, and Phase III involves 300 to 3,000 subjects.

1.2 Major Drug Modalities

Drugs can be broadly classified into three major modalities based on their molecular weight: small molecules, middle molecules, and antibodies. Table 1.1 shows the characteristics, and Fig. 1.2 shown the example structures of each type of drug.

1.2.1 Small molecule drugs

Small molecule drugs have the longest development history for advantages such as oral administration and low manufacturing costs and currently account for about 75% of the global pharmaceutical market [18]. According to the "Rule of five" (Ro5) pro-

Table 1.1: Characteristics of major drug modalities: small molecules, linear peptides, cyclic peptides, and antibodies (created based on [16]).

	Small molecules	Middle molecules		Antibodies
		Linear peptides	Cyclic peptides	
Molecular weight	typically < 600	500 – 6,000	600 – 2,500	$> 150{,}000$
Activity	medium – high	high – very high	very high	very high
Selectivity	low	high	very high	very high
Intracellular targets	possible	generally not possible	possible	not possible
PPI inhibition	difficult, but possible	possible	possible	possible
Plasma stability	low – medium	very low – low	medium – high	very high
Oral administration	possible	generally not possible	possible	not possible
Toxicity / Side effects	high	low	low	low
Manufacturing cost	low	low – medium	low – medium	high

posed by Lipinski *et al.* [19], for the chemical properties of typically small molecule compounds that are likely to become oral drugs, small molecule drugs are generally characterized by moderate lipid solubility and relatively small molecular weight (< 500 Da). Many small molecule drugs are permeable to cell membranes due to their small molecular weight, which allows for easier diffusion through the lipid bilayer of the membrane, making them suitable for targeting intracellular targets. Small molecules typically bind to rigid targets that possess druggable pockets, such as active sites or cavities on the protein surface with a well-defined structure that can accommodate them [20]. However, for neurodegenerative diseases such as Parkinson's disease and Alzheimer's disease, which are still considered incurable, proteins and peptides associated with these disorders are often intrinsically disordered. They lack a well-defined structure in their native state [21], preventing the existence of such druggable pockets. Moreover, small molecules have difficulty inhibiting both intracellular and extracellular protein–protein interactions (PPIs), which are continually being discovered as the major "undruggable" targets [22, 23]. PPI interfaces are typically flat and large or groove-shaped (area of approximately 1,000 to 3,000 $Å^2$ [24, 25]) and lack the deep and narrow binding pockets required by small molecule drugs (volume of approximately 300 to 1,000 $Å^3$ [25, 26]). Furthermore, small molecule drugs suffer from low target selectivity and may bind to proteins other than the target, ultimately causing side effects [27].

On the other hand, many small molecule drugs that exceed the molecular weight of 500 Da, known as beyond Rule of Five (bRo5) compounds, have also shown therapeutic potential. These bRo5 compounds tend to have complex structures, often exhibiting higher lipophilicity and potentially engaging in different modes of interaction with their

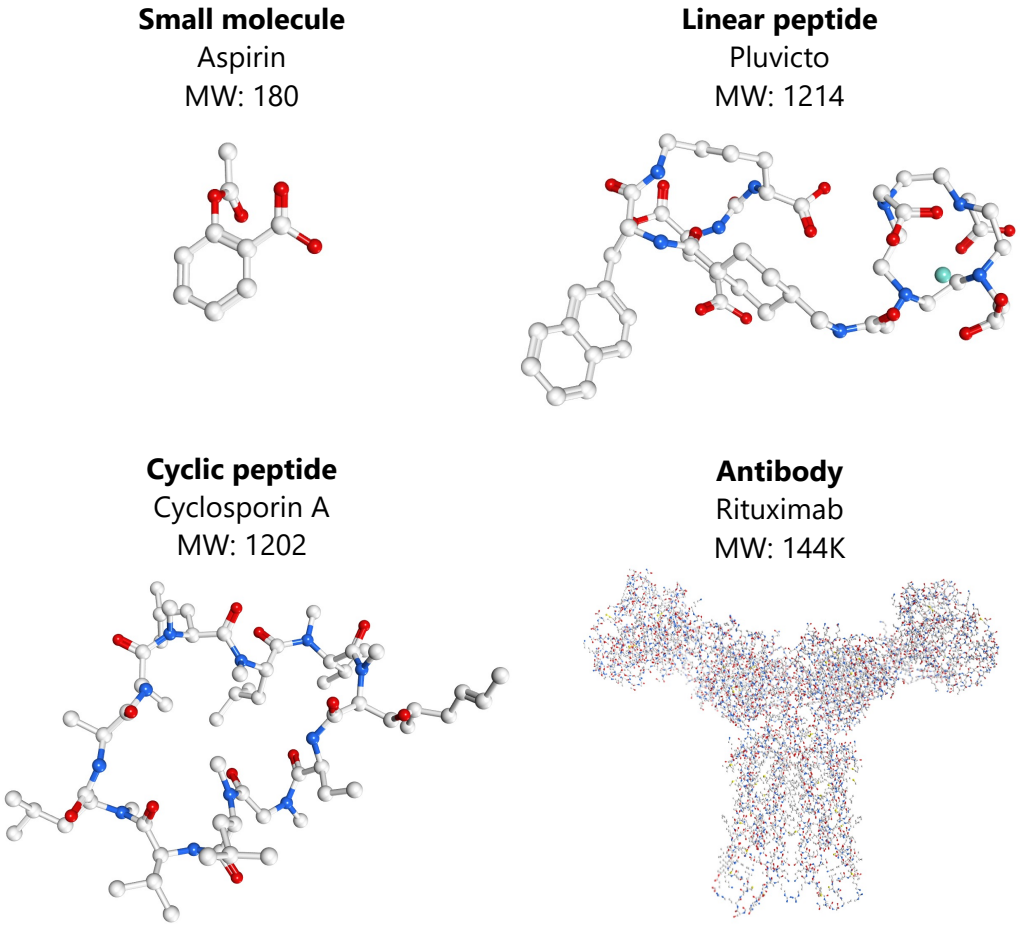


Figure 1.2: Structures and molecular weights of the small molecule drug Aspirin, the linear peptide drug Pluvicto, the cyclic peptide drug Cyclosporin A, and the antibody drug Rituximab.

targets compared to traditional small molecules. Notably, promising compounds such as proteolysis targeting chimeras (PROTACs) also reside within the bRo5 chemical space, offering new opportunities in medicinal chemistry [28].

1.2.2 Antibody drugs

Antibodies are Y-shaped proteins with a much larger molecular weight than small molecules (over 150 kDa), which can recognize and bind to protein targets with high specificity and modulate their toxic behavior [20]. Because of their high specificity to their target, antibody drugs can be expected to have high therapeutic efficacy and rarely cause side effects. Particularly in the field of cancer treatment, starting with

Rituximab (Rituxan®), a chimeric monoclonal antibody that specifically binds to the CD20 antigen on normal and malignant B lymphocytes [29], in the 1990s, the number of antibody drugs for cancer treatment has grown steadily. Combination therapies with other conventional anticancer drugs have been widely studied and are showing great promise [30]. In addition, by binding to large, flat surfaces involved in PPIs, antibodies can effectively disrupt these interactions. The US Food and Drug Administration (FDA) has approved 140 antibody-based drugs until 2024 [31]. However, due to the size of the molecule, antibodies cannot enter the cell unless they are combined with a special molecule using DDS (Drug Delivery System) technology, etc., making it impossible to target intracellular targets. Furthermore, antibodies are produced using complex techniques such as genetic recombination, which results in high production costs [32].

1.2.3 Conventional middle molecule drugs

One of the difficulties in drug development is the decreasing number of targets remaining for conventional small molecule and antibody drugs. It has been reported that approximately 80% of all existing disease-relevant targets, including those involved in intracellular PPIs, cannot be tackled by conventional small molecule and antibody drugs [23, 33]. Under these circumstances, middle molecule drugs have been gaining attention due to their ability to interact with previously "undruggable" proteins by conventional small molecule and antibody drugs [25]. In addition to peptide drugs, including the linear peptide and cyclic peptide drugs described in Table 1.1, the other type of middle molecule drugs is nucleic-acid drugs, which target nucleic acids such as mRNA and miRNA. The sequencing of the human genome and the elucidation of many molecular pathways that are important in disease management have provided unprecedented opportunities for the development of new nucleic-acid therapeutics. Various strategies exist for targeting mRNA, including the use of antisense DNA, antisense RNA, and RNA decoy molecules [34].

Insulin was the first synthesized therapeutic peptide in 1921, the highest-selling peptide drug today [18]. Subsequently, peptides were frequently studied as they combine the advantages of small molecule and antibody drugs. For example, homing peptides (HPs), which can bind to specific sites in the vasculature, can improve drug delivery systems to tumors and are gaining attention in tumor therapy [35]. The growing advancements in genetic engineering, peptide synthesis technologies, and sequence analysis tools have led to the development of new classes of peptide therapeutics for various applications [36, 37, 38, 39]. However, certain limitations of conventional lin-

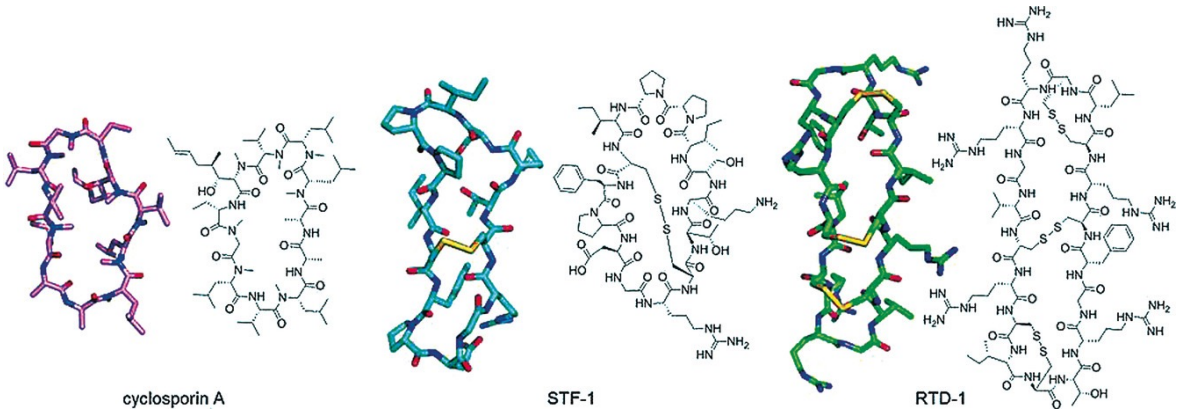


Figure 1.3: Examples of natural head-to-tail cyclic peptides (cited from [3]). From the left, Cyclosporin A, and STF-1 and RTD-1 with disulfide bridges are shown.

ear peptides, such as low stability, selectivity, and cell membrane permeability remain unresolved [3, 4]. For example, linear peptides are easily degraded by stomach acid and intestinal enzymes, making oral administration nearly impossible.

1.3 Cyclic Peptide Drugs

1.3.1 Overview of cyclic peptide drugs

Macrocycles are generally defined as organic molecules that contain a ring of at least 12 heavy atoms and have been treated as a prominent example of new modalities that exist in chemical space beyond the "Rule of five" (bRo5) [5]. Due to the availability of robust synthetic and biological methods that enable the rapid assembly of the amino acid building blocks, cyclic peptides have received the most attention in drug discovery among different classes of macrocycles [40]. A typical head-to-tail cyclic peptide forms a macrocyclic structure via an amide bond between the N-terminal amine and the C-terminal carboxylic acid, and natural cyclic peptides often contain disulfide bridges between cysteine residues (Fig. 1.3) [3]. In addition to the head-to-tail type, various macrocyclization reaction approaches have been developed over the years. Depending on the strategy, backbone cyclizations, side-chain to side-chain cyclizations, lariat peptides, and other more esoteric topologies can now be easily synthesized (Fig. 1.4) [4].

Cyclic peptides exhibit several pharmacological characteristics distinct from other well-established drug classes, resulting in a versatile modality with a unique profile of advantages and limitations [4]. In contrast to linear peptides, the unique structural

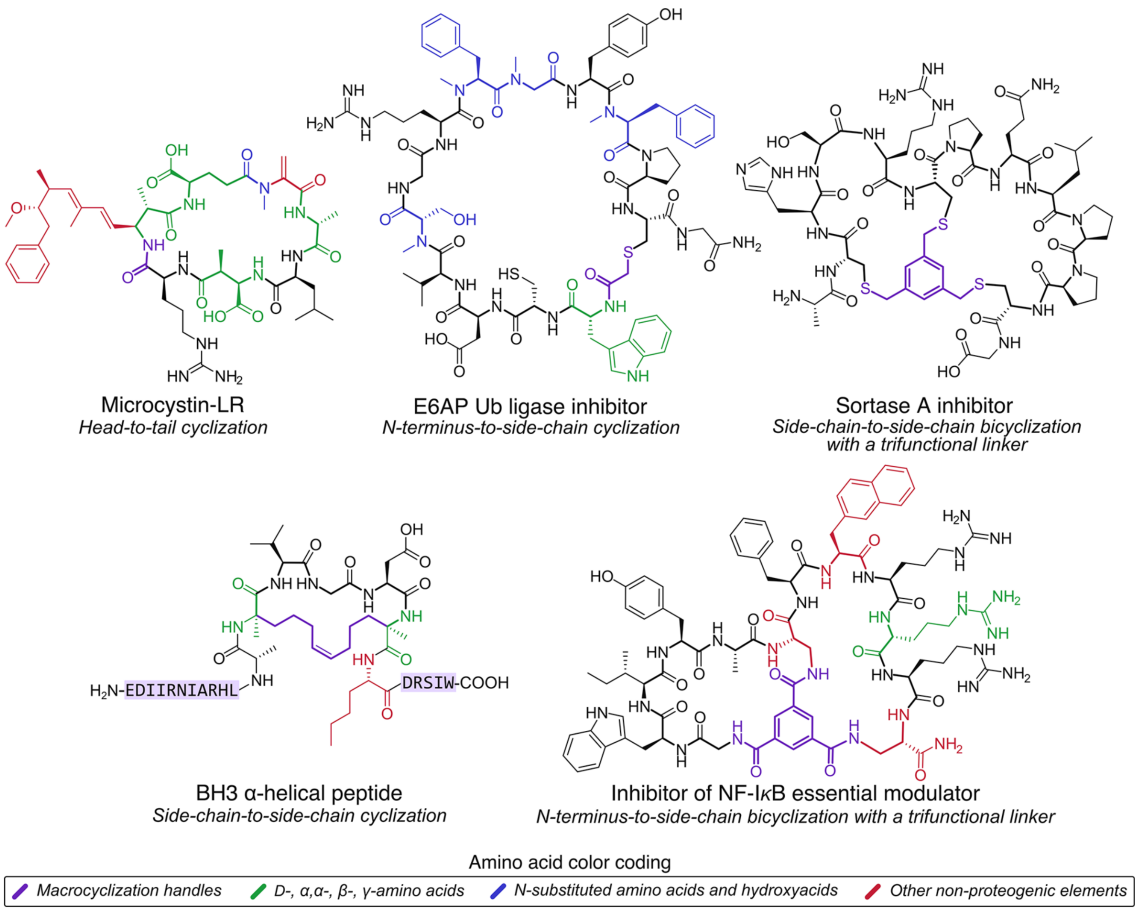


Figure 1.4: Various macrocyclization approaches of cyclic peptides (cited from [4]).

features of macrocyclic peptides, such as their restricted conformational flexibility and local secondary structure motifs, actually help stabilize bioactive conformations, enhancing their affinity, potency, stability, and selectivity for specific targets [5, 10, 20, 41]. In some cases, macrocyclization results in up to a 100-fold increase in potency compared to the corresponding linear compounds [42, 43]. Despite their larger size, some carefully designed cyclic peptides exhibit sufficient cell permeability and bioavailability to reach intracellular targets via oral administration [44, 45]. At the same time, cyclic peptides can bind to "difficult to drug" targets, especially those with large flat, groove-shaped, or tunnel-shaped binding sites [5, 46]. As a result, cyclic peptides have been considered a breakthrough modality because of their ability to inhibit intracellular PPIs [47, 48]. Furthermore, cyclic peptides are easy to synthesize, making them relatively inexpensive to produce compared to antibody drugs, and lead compounds can be optimized through traditional medicinal chemistry efforts to tailor biophysical

properties for specific applications [4].

In the past century, cyclic peptide drugs were predominantly sourced from natural products, including antimicrobial agents and human peptide hormones. Recent advances in novel synthesis and display technologies have led to breakthroughs in cyclic peptide drug discovery and simplified the discovery of functional cyclic peptides against specific targets [10, 18, 47, 49]. For example, the random nonstandard peptides integrated discovery (RaPID) system designs cyclic peptides from a diverse library, including non-natural amino acids, enabling the synthesis and rapid selection of potent binders for a wide range of therapeutic targets [18, 50]. The RaPID system has designed novel cyclic peptides for complex therapeutic targets, including a high-affinity binder to the osteoporosis target PlexinB1 [51], an inhibitor of the ubiquitin-protein ligase E6AP [50], and a selective inhibitor of the oncogenic K-Ras [52]. Although no peptides developed through the RaPID system have advanced to late-stage clinical trials, one peptide (zilucoplan) developed using a similar technology (Ra Pharmaceuticals mRNA display platform) is currently in phase III trials for the treatment of generalized myasthenia gravis [18]. Overall, the FDA has approved more than 60 macrocyclic drugs (Fig. 1.5, Table A.1, Fig. A.1) [5] and has approved approximately one cyclic peptide-based drug per year for the past 20 years [47]. As shown in Fig. 1.6, infectious diseases are the primary therapeutic indication for macrocyclic agents (44.4%), followed by oncology (20.8%) [5].

1.4 Remaining Two Major Challenges for Cyclic Peptide Drug Discovery

Poor pharmacokinetic properties are one of the major causes of new drug development failure during clinical phases [53, 54]. Despite their pharmacological potential, the development of cyclic peptides as therapeutic agents faces challenges in several critical areas: cellular uptake (determining the ability to target intracellular targets), oral bioavailability (essential for oral drug administration), metabolic stability (ensuring stability within the body), and renal clearance (affecting drug elimination rates) [4]. Only less than 40% of the FDA-approved macrocyclic drugs can be administered orally (Fig. 1.5). Notably, these issues primarily stem from the inherent characteristics of cyclic peptides, which often exhibit poor membrane permeability and plasma protein binding (PPB) rates, thereby severely limiting their biological applications and reducing their feasibility as orally available drugs [4, 5, 20, 23, 55]. This section delves

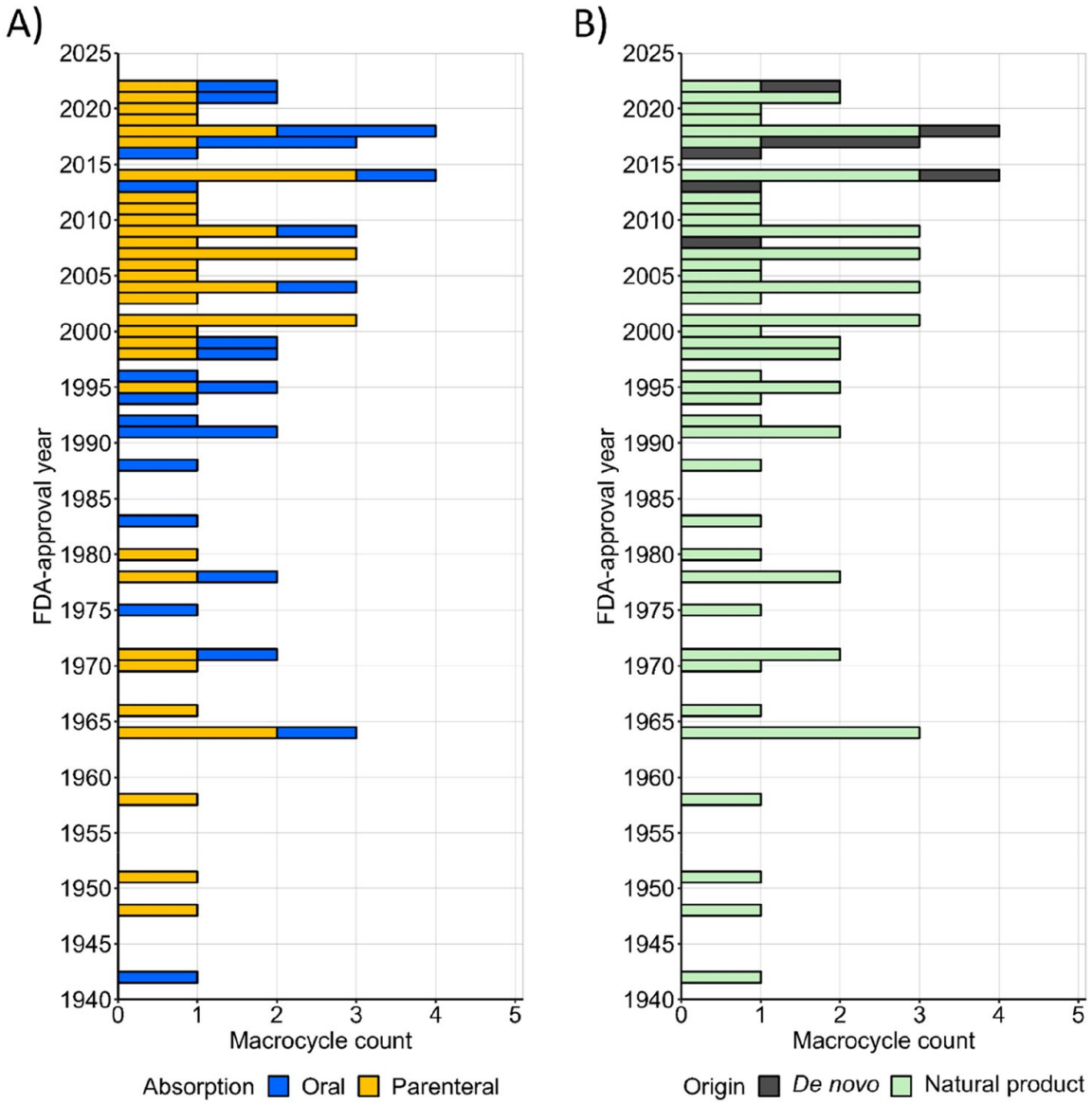


Figure 1.5: FDA-approved macrocyclic drugs ($n = 67$) (cited from [5]). (A) Orally available drugs are shown in blue ($n = 26$; 39%), while those administered parenterally are shown in gold ($n = 41$; 61%). (B) Natural products and their derivatives are indicated in light green ($n = 59$; 88%); *de novo* designed macrocyclic drugs, which are synthetically created from scratch rather than derived from existing natural compounds, are indicated in dark gray ($n = 8$; 12%).

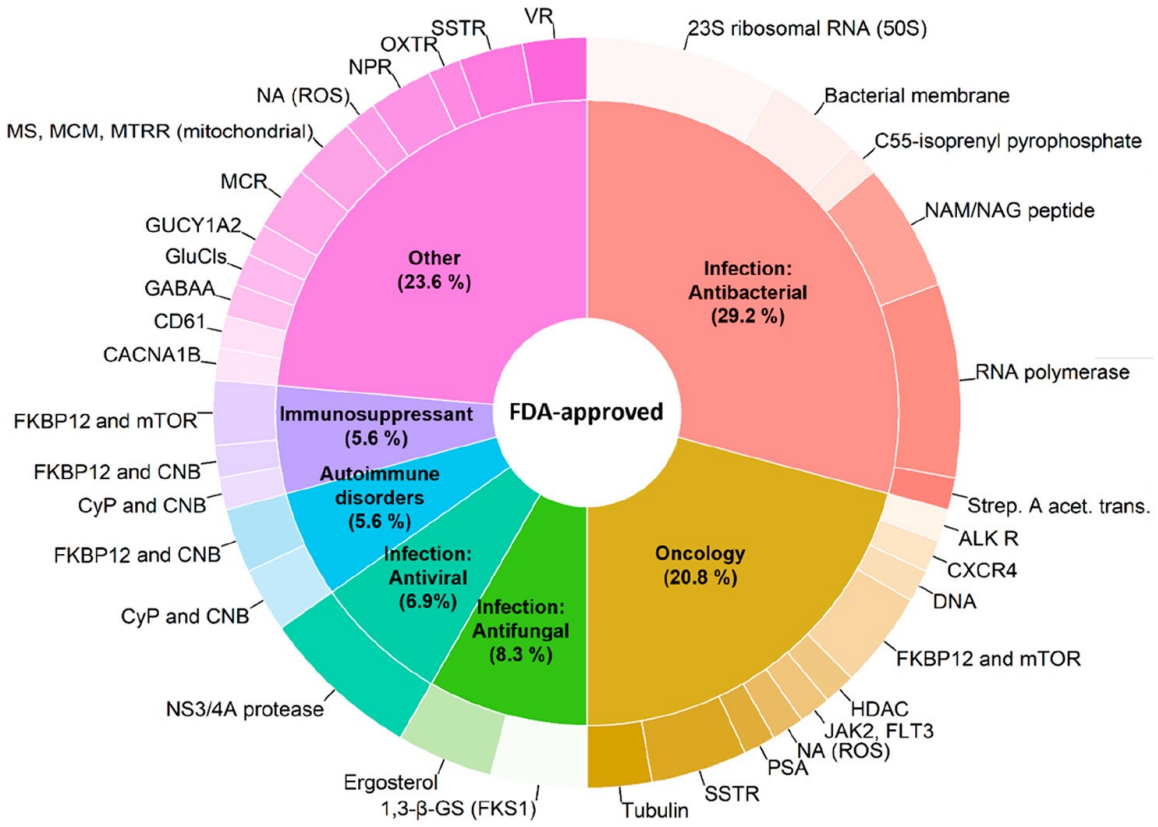


Figure 1.6: Therapeutic indications (inner circle) and targets (outer circle) of the FDA-approved macrocyclic drugs ($n = 72$) (cited from [5]). Five macrocycles are duplicated because each is used for two therapeutic indications.

into the principles of cell membrane permeability and PPB, their experimental determination, and the limitations of current computational approaches for predicting these properties.

1.4.1 Cell membrane permeability

Overview of cyclic peptide membrane permeation

The cell membrane is composed of a phospholipid bilayer with a hydrophilic head group of choline and phosphate and a hydrophobic tail group of hydrocarbon chains, with a thickness of about 50 to 100 Å. Cell membrane permeability is one of the most important indicators for assessing oral bioavailability and the possibility of intracellular targeting. In contrast to some linear peptides, such as cell-penetrating peptides (CPPs), which permeate membranes actively, the most prevalent mechanism by which cyclic peptides permeate membranes is energy-independent passive diffusion, similar to

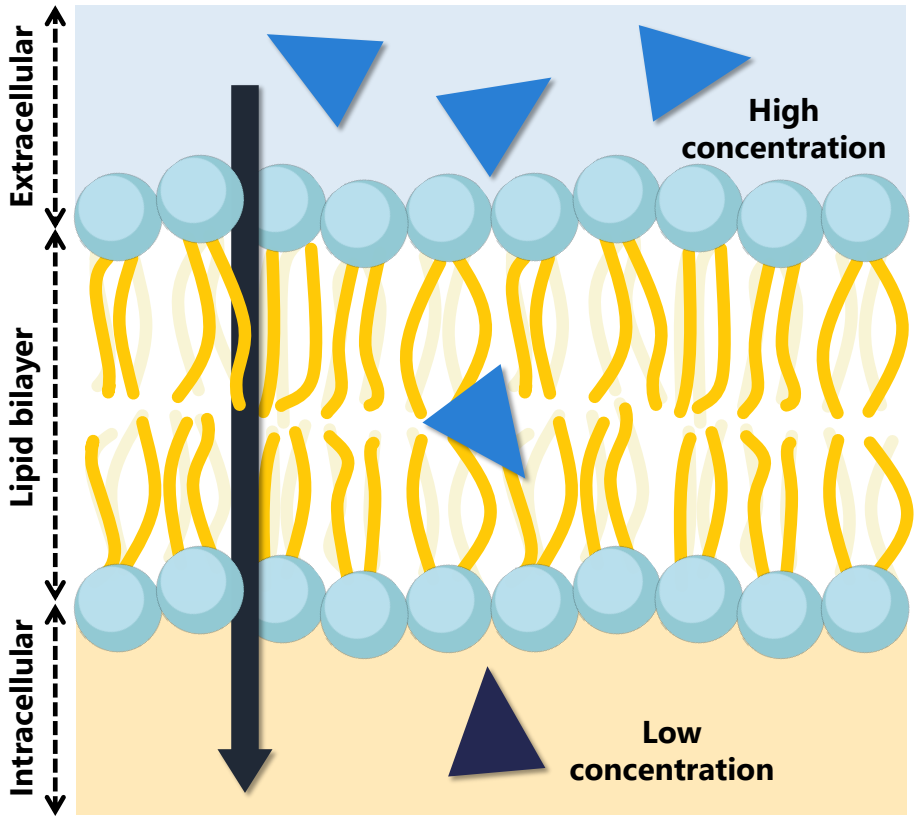


Figure 1.7: Image of the membrane permeation process by passive diffusion. The process of passive diffusion is driven by a compound's net concentration gradient, flowing from a high-concentration region (extracellular) to a low-concentration region (intracellular).

small molecules (Fig. 1.7) [6]. In general, relatively large cyclic peptides exceeding 10 amino acid residues have high target affinity; however, they tend to have low membrane permeability due to their large size. Even for passive diffusion, the process by which cyclic peptides permeate membranes is not well understood. Many cyclic peptides have an environment-dependent property called the "chameleonic" property, namely, the ability to change their molecular conformation and hydrophobicity in response to the surrounding environment. These chameleonic cyclic peptides prefer an "open"-conformation in aqueous environments, where the polar groups are exposed to the outside and interact with water molecules to enhance water solubility, and a "closed"-conformation in hydrophobic environments, where the polar groups are shielded with intramolecular hydrogen bonds or lipophilic side chains, leading to improved membrane permeability (Fig. 1.8) [56, 57, 58]. Therefore, the most common strategy to

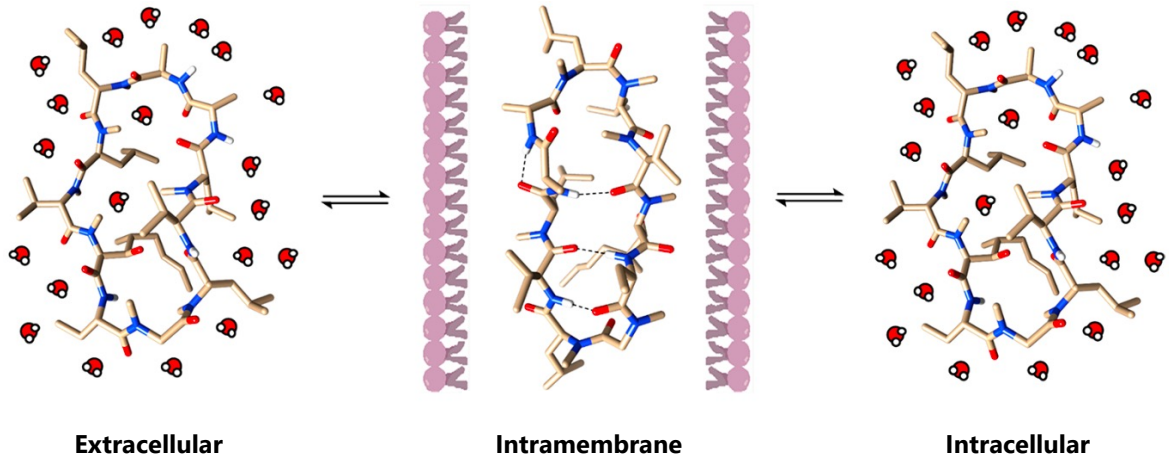


Figure 1.8: Image of the dynamic conformation changes of Cyclosporin A when transitioning between aqueous and hydrophobic environments (cited from [6]). Cyclosporin A exists in an "open"-conformation inside the aqueous extra- and intracellular environments but adopts a "closed"-conformation containing four intramolecular hydrogen bonds shown as dotted lines upon entering the lipid bilayer.

increase membrane permeability has been to allow cyclic peptides to have a "closed"-structure inside the cell membrane by shielding the exposed hydrogen bond donor (-NH group) with backbone N-methylation [7, 59]. In addition, various strategies, such as conformational control [8, 45], amide-to-ester substitution [9], amide-to-thioamide substitution [10], and side-chain modifications [11] have emerged for improving membrane permeability (Fig. 1.9). However, these strategies do not always improve membrane permeability across all cyclic peptides.

Experimental membrane permeability measurement methods

Currently, the most common methods for evaluating the membrane permeability of cyclic peptides are biochemical assays. The following five methods are representative of such measurements:

- **Parallel artificial membrane permeability assay (PAMPA)** [60, 61]
  PAMPA is a primary screen surrogate for gastro-intestinal permeability (passive diffusion); it uses a thin artificial membrane (hexadecane layer) as a model membrane instead of cells of biological origin. This assay is performed, for example, in 96-well plates and measures the ability of compounds to diffuse from a donor into a separated acceptor compartment after an incubation period (four to 24 hours). Since the PAMPA membrane components do not include transporters,

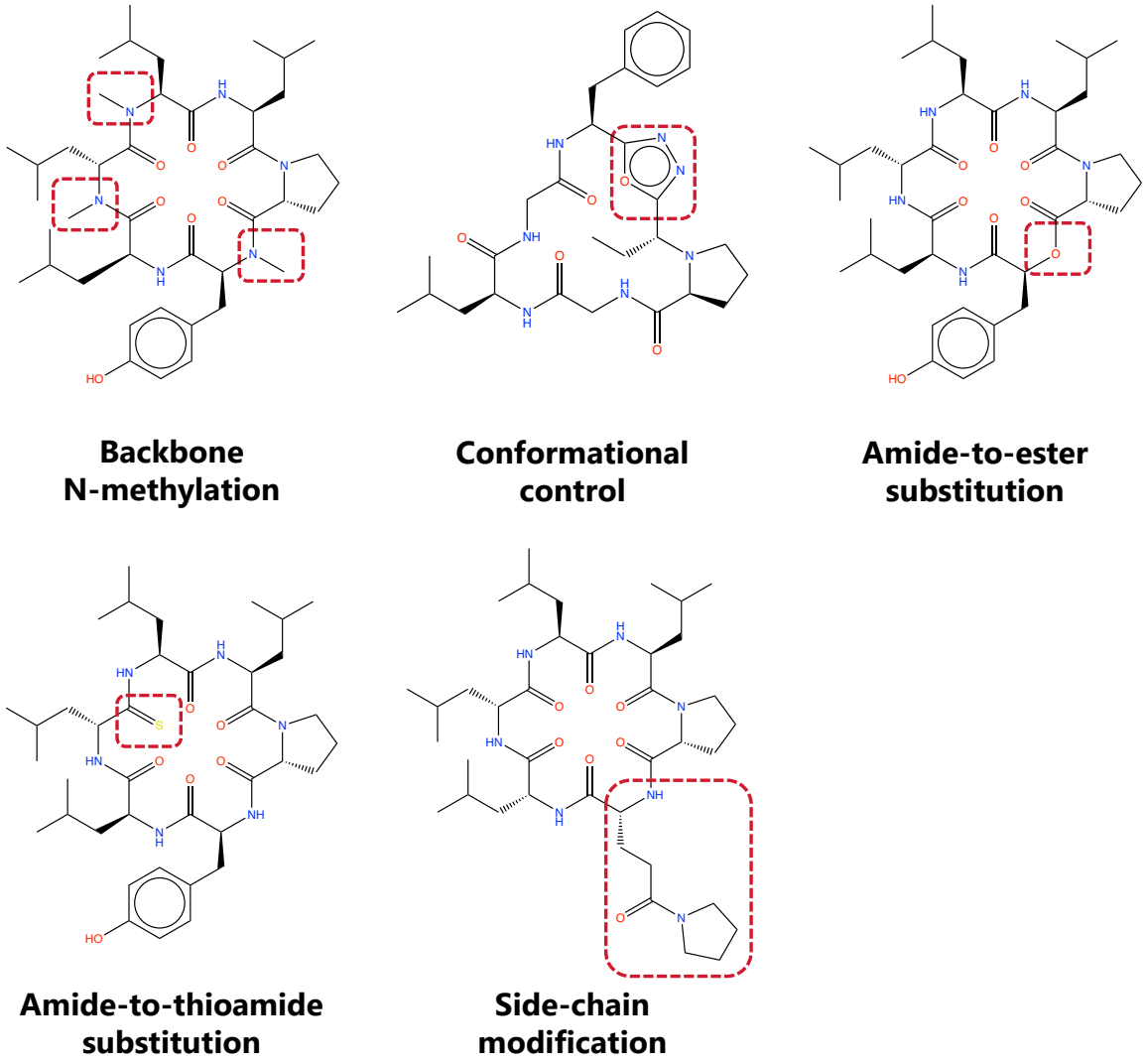


Figure 1.9: Examples of modification strategies to improve membrane permeability: backbone N-methylation [7], conformational control [8], amide-to-ester substitution [9], amide-to-thioamide substitution [10], and side-chain modification [11].

it is possible to measure only the membrane permeability for passive diffusion. Hence, a limitation of PAMPA is that it does not account for active transport mechanisms or metabolism, which can affect the overall permeability and absorption in biological systems. PAMPA also can be performed at a lower cost than cell-based assays.

- **Caco-2** [62]
  Caco-2 assay is a cell-based assay, and Caco-2 cells are derived from human colon

adenocarcinoma. When grown on semipermeable filters, Caco-2 cells spontaneously differentiate in culture to form a monolayer of cells that structurally and functionally resemble small intestinal epithelial cells. It takes about three weeks to culture Caco-2 cells.

- **Madin-Darby canine kidney (MDCK)** [61, 63]
  MDCK assay is a cell-based assay, and MDCK cells are a nonhuman origin epithelial cell line derived from the canine kidney. MDCK cells exhibit low expression of transporter proteins and have minimal metabolic activity. MDCK cells can be cultured faster than Caco-2 cells in about three days.

- **Ralph Russ canine kidney (RRCK)** [64]
  RRCK assay is a cell-based assay, and RRCK cells are derived from the MDCK cell line and are also called MDCKII-LE (low efflux) cells. The primary concern with using MDCK cells is the presence of endogenous transporters of nonhuman origin, such as canine P-glycoprotein (Pgp), which can interfere with permeability and transporter studies, resulting in less reliable data. In contrast to MDCK cells, RRCK cells exhibit over 200-fold lower canine Pgp mRNA levels and less functional efflux activity.

- **Lilly Laboratories cell-porcine kidney 1 (LLC-PK1)** [65]
  LLC-PK1 assay is a cell-based assay, and LLC-PK1 cells are derived from porcine renal epithelial cells.

Membrane permeability obtained from different measurement methods can vary significantly. For example, Wang *et al.* [66] reported both PAMPA and Caco-2 measurements of 62 cyclic peptides, but the correlation coefficient between PAMPA and Caco-2 permeability values was only 0.71.

1.4.2 Plasma protein binding (PPB)

Plasma protein

Plasma is an aqueous solution that comprises 55% of the components of human blood and is composed of 92% water, 7% protein, and 1% other solutes such as inorganic ions. Human serum albumin (HSA) is the most abundant plasma protein, present at a concentration of approximately 7×10^{-4} M and accounting for 55% of the total plasma protein. The primary physiological role of HSA is to transport fatty acids, and HSA is also involved in the maintenance of colloidal osmotic blood pressure, the fluid

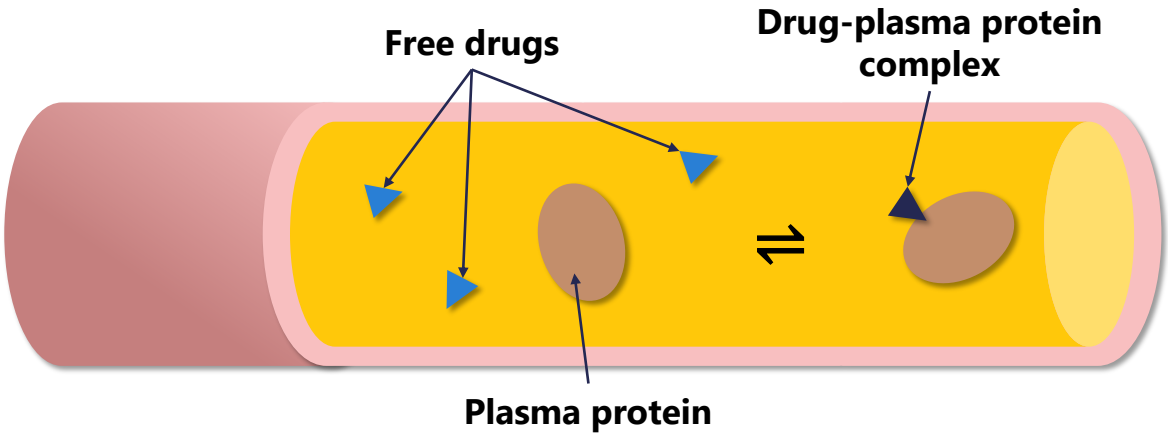


Figure 1.10: Image of drug binding to plasma protein.

distribution between body compartments, the protection of the organism by binding toxic metabolites, and the storage of nitric oxide [67]. HSA exhibits an extraordinary ligand-binding capacity, binding reversibly to acidic and neutral compounds at various binding sites. While it also binds basic drugs, this occurs to a lesser extent. Due to its importance, the term plasma protein binding is often associated with HSA. Following HSA, α1-acid glycoprotein (AGP or AAG) is the second most crucial serum protein for drug binding; it is present at a much lower concentration (approximately 1 to 3×10^{-5} M), representing only 1.8% to 5.5% of HSA [67].

Overview of PPB of cyclic peptide

When drugs enter the body, they exist in two states: as free drugs or as drug-plasma protein complexes bound to plasma proteins Fig. 1.10. The PPB rate is defined using the concentration of the free drugs $[D]$ and the concentration of the drug-plasma protein complexes $[PD]$ in the body as follows:

$$\%PPB = \frac{[PD]}{[D] + [PD]} \times 100 = \frac{(\text{Mass of drugs present as complexes})}{(\text{Mass of drugs administered})} \times 100 \quad (1.1)$$

The binding of drugs to plasma proteins is reversible, and only free drugs can bind to targets or enzymes to exert their pharmacological activity or be converted to the corresponding metabolites [68]. The function of plasma proteins as drug carriers can facilitate the delivery of drugs to their site of action while simultaneously reducing side effects. Therefore, PPB has a substantial effect on the absorption, distribution, metabolism, excretion, and toxicity (ADME-Tox) of drugs and strongly affects drug

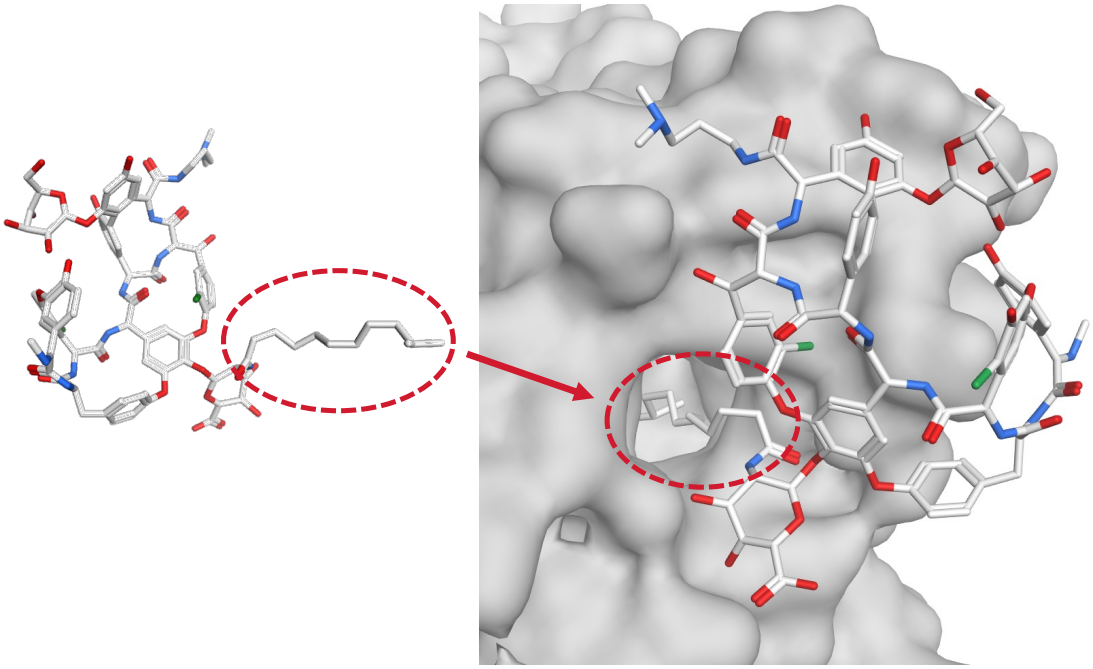


Figure 1.11: Cocrystal structure of cyclic peptide drug dalbavancin and HSA (gray) (PDB ID: 6M5E). The hydrocarbon side chain of dalbavancin is inserted deeply into the hydrophobic pocket of HSA.

distribution and pharmacokinetic behavior with consequences in overall pharmacological action [67]. Overall, the PPB rate for a specific drug is a key indicator for lead optimization.

Although studies on PPB of cyclic peptides are not as advanced as those on membrane permeability, the binding mode of cyclic peptides to plasma proteins has been elucidated. In contrast to small molecules, whose entire molecules fit into the binding pocket of plasma proteins, the hydrophobic local structure of cyclic peptides is known to form hydrophobic bonds with plasma proteins (Fig. 1.11) [13, 55, 69]. Although sufficiently hydrophobic peptides may exhibit satisfactory PPB, it is generally necessary to conjugate a fatty acid or a similar lipophilic moiety to the peptide to enhance the HSA binding rate and extend its plasma half-life [4].

Experimental PPB rate measurement methods

As with membrane permeability, biochemical experiments are the current most common method for measuring the PPB rate of cyclic peptides. Several methods have been used to experimentally measure PPB rates, including equilibrium dialysis (ED), ultrafiltration (UF), and ultracentrifugation (UC) [70, 71, 72].

- **ED**: ED is the most widely used method. A dialysis membrane separates plasma and buffer, and the drug concentration in both solutions is measured after the equilibrium is reached. Recently, rapid equilibrium dialysis (RED), using a 96-well format device, was developed to enhance the method's throughput.

- **UF**: In this method, the plasma solution containing drugs is filtered by an ultrafiltration membrane that allows only the drug to pass through, and the drug concentration in the solution is measured. The analysis speed can be enhanced by applying pressure to force the solution through the membrane. However, non-specific binding of the drug to the filtration membrane and other surfaces can lead to experimental artifacts.

- **UC**: In this method, the plasma solution containing drugs is centrifuged, and the free drug concentration in the supernatant after centrifugation is measured. UC has the advantage of eliminating issues related to membrane effects.

1.4.3 Limitation of existing computational prediction methods of cyclic peptide membrane permeability and PPB rate

Selection of candidate compounds with high membrane permeability and PPB rate is important during the early stages of drug development. However, traditional biochemical assays first require the synthesis of actual compounds. Moreover, both the membrane permeability and PPB rate measurement experiments are time-consuming (measurement of membrane permeability by PAMPA takes approximately 4 to 24 hours, and measurement of PPB rate takes approximately 6 to 24 hours [67] per molecule), and different measurement methods and experimental conditions may obtain different results. Thus, due to the cost associated with randomly measuring the permeability and PPB rate of numerous peptides using biochemical assays, the development of fast computational approaches that enable the prediction of these properties is eagerly anticipated. The approaches developed so far can be broadly categorized into two types: molecular simulation-based and quantitative structure-property relationship (QSPR)-based (most using machine learning (ML) in recent years) methods. We briefly discuss these prediction methods here, and more detailed explanations can be found in later Chapter 2.

Molecular simulation-based methods

Molecular dynamics (MD) simulations are widely used for membrane permeability predictions, and molecular docking methods are typically applied to predict PPB rates. Analyzing simulation results can yield valuable information on molecular behavior for structural optimization and drug development. However, these simulation-based methods are highly computationally intensive. For instance, the latest MD-based method for predicting membrane permeability requires parallel computations on 224 NVIDIA P100 GPUs and takes approximately 90 hours per peptide [73]. Similarly, performing rigid-body docking to predict the PPB rate using small molecule software (e.g., Glide) can take over 30 hours per cyclic peptide.

ML-based methods

ML-based approaches offer significantly faster predictions than simulation-based methods, with computational times ranging from under one second to a few seconds per cyclic peptide for the prediction stage once featurization is complete. However, due to the limited amount of data available for cyclic peptides, ML-based prediction models developed so far often face challenges in generalization performance to extrapolation data, highlighting the need for larger, diverse datasets tailored specifically for cyclic peptides. Moreover, most models rely on traditional ML techniques and fail to account for unique structural features distinguishing cyclic peptides from small molecules.

1.5 Overview of Major Deep Learning Architecture

In recent years, deep learning (DL) has dramatically improved state-of-the-art technologies in various fields, such as object recognition [74], speech recognition [75], and protein structure prediction [76]. Deep learning models are based on neural networks that mimic the behavior of human neurons and have been a subject of research since the 1940s. Unlike traditional machine learning methods, deep learning models extract and learn intricate structures in large data sets by employing the backpropagation algorithm to indicate how a machine should change its parameters used to compute the representation in each layer from the representation in the previous layer [77]. The major deep learning architectures commonly used today include the following four types:

- **Convolutional Neural Network (CNN)**: CNNs are typically composed of distinct convolutional layers followed by pooling layers and fully connected layers

and are primarily used for tasks involving image processing and computer vision [78]. They leverage convolutional layers to automatically detect spatial patterns and hierarchies in data, such as edges, textures, and shapes in images, making them highly effective for tasks like image classification, object detection, and facial recognition.

- **Recurrent Neural Network (RNN)**: RNNs are designed to handle sequential data where the order of inputs matters, such as in time-series analysis, speech recognition, and natural language processing [79]. They maintain a form of memory by feeding the output from previous time steps back into the network, enabling them to learn from and make predictions based on sequences. Common variants of RNNs include Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), which are widely used to improve the handling of long-term dependencies in sequences.

- **Graph Neural Network (GNN)**: GNNs are specialized for data represented in graph structures, where relationships between nodes are important [80]. GNNs can effectively capture the dependencies and interactions between nodes and edges, making them highly suitable for tasks such as social network analysis and molecular property prediction. One of the widely used variants of GNNs is the Graph Convolutional Network (GCN), which extends the idea of convolution from images to graph structures.

- **Transformer**: Transformers have become the most important architecture in recent deep learning developments, revolutionizing many fields, particularly natural language processing [81]. The core innovation of transformers lies in the self-attention mechanism, which allows the model to weigh the importance of different input elements relative to each other. This mechanism enables the model to capture long-range dependencies and relationships in the data more effectively than traditional sequential models like RNNs. Furthermore, attention computation does not require sequential processing of inputs, allowing for parallel computation and making transformers more computationally efficient than CNNs and RNNs.

The summary of DL-based small molecule property prediction methods is described in Section 2.4.

1.6 Purpose of Study

To accelerate cyclic peptide drug discovery, we aim to overcome the two major challenges in establishing fast computational prediction methods for membrane permeability and PPB rate of cyclic peptides: the limited availability of experimental data and the unique structural characteristics of cyclic peptides.

In this study, we describe the development of a DL-based prediction method that can quickly and accurately predict cyclic peptide membrane permeability and PPB rate. Our method demonstrates a prediction speed comparable to traditional ML-based methods, capable of predicting the permeability or PPB rate of a single cyclic peptide in just a few seconds after featurization. To realize model construction, we first collected and organized an unprecedented amount of cyclic peptide data from published papers, patents, and collaborative research with a pharmaceutical company. Next, we designed cyclic peptide features at the peptide-, monomer-, and atom-levels to capture the complex structural characteristic of cyclic peptides. Additionally, we applied various data augmentation techniques to enhance model training efficiency.

The cyclic peptide membrane permeability prediction model is called CycPeptMP. We evaluated its accuracy compared to state-of-the-art DL-based property prediction methods and MD-based methods. In PPB rate prediction, we developed a PPB rate prediction model, CycPeptPPB, and a method to present important substructures that affect the PPB rate to support the design and optimization of candidate compounds.

1.7 Summary of Contributions

The contributions of this thesis are classified into three categories: (1) proposal of the concept of multi-level molecular features design for cyclic peptides, (2) development of data augmentation methods for efficient model training, and (3) experimental data collection and construction of prediction models with excellent performance, CycPeptMP and CycPeptPPB. We now describe these in more detail.

- We proposed the fundamental concept of a novel monomer-level feature design and fused it with peptide- and atom-levels molecular features to comprehensively represent the characteristics of cyclic peptides. In permeability prediction, where it is crucial to capture cyclic peptides' local and global structural features simultaneously, ablation studies demonstrated that all feature levels contributed and were relatively essential for prediction. In PPB rate prediction, where substruc-

tures are essential in principle, monomer-level feature design was able to capture that unique feature of cyclic peptides.

- We developed a data augmentation method that considers their unique structural characteristics, referencing data augmentation techniques commonly used in image processing to learn the structures of cyclic peptides, which are far more complex than small molecules. We unveiled that data augmentation was essential for the successful prediction of both permeability and PPB rate prediction.

- We greedily searched published papers and pharmaceutical company patent documents from scratch, collecting over 7,000 cyclic peptide structures and experimentally measuring membrane permeabilities from 47 publications. We have released this database, together with various additional functions such as sequence notation, as the world's first cyclic peptide membrane permeability database, CycPeptMPDB [82]. This paper has been cited by 19 papers in 18 months since its publication. In the field of permeability prediction, where no database was available, the emergence of CycPeptMPDB made it possible for the first time to apply deep learning techniques, bringing significant advances in the field.

- We designed a deep learning architecture suitable for multi-level molecular features and constructed the CycPeptMP [83] and CycPeptPPB [84] models with hyperparameters determined from exhausted hyperparameter search. These models showed superior performance to existing cyclic peptide prediction methods and state-of-the-art small molecule property prediction methods.

1.8 Thesis Organization

The remaining chapters of this thesis are organized as follows: Chapter 2 reviews the previous research for cyclic peptide membrane permeability and PPB rate prediction. It also reviews the DL-based small molecule properties prediction methods. Chapter 3 describes a novel multi-level molecular features design method to concurrently capture the local sequence variations and global conformational changes in cyclic peptides. It also presents a new data augmentation method tailored to the above multi-levels for efficient model training for complex cyclic peptide structures. Based on these proposed techniques, Chapter 4 and Chapter 5 describe the development of prediction models for membrane permeability (CycPeptMP) and PPB rate (CycPeptPPB) of cyclic peptides, respectively. All experiments of CycPeptMP were from the article

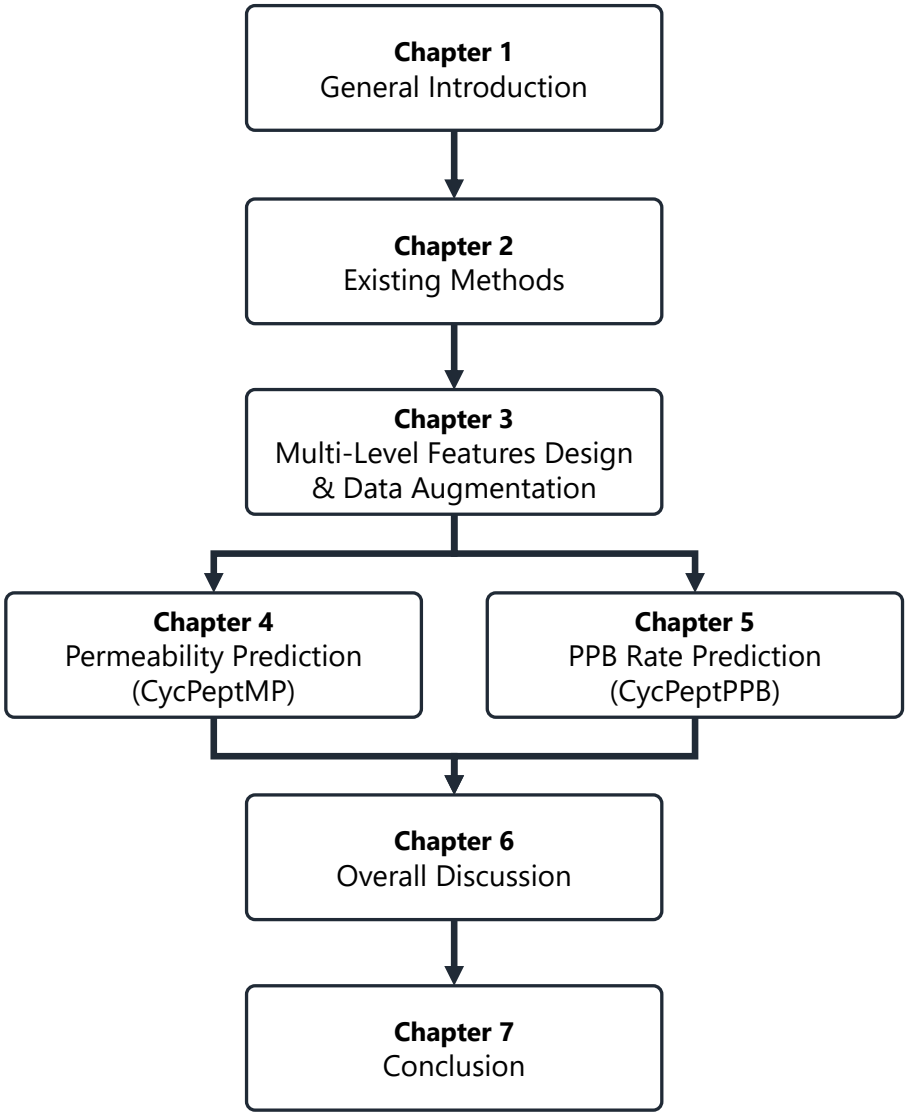


Figure 1.12: The relationship between chapters.

[82, 83], and all experiments of CycPeptPPB were updated from the article [84]. The overall discussions of membrane permeability prediction and PPB rate prediction are given in Chapter 6, and conclusions are presented in Chapter 7. This thesis is based on the following publications by the author: [82, 83, 84].

Chapter 2

Overview of Cyclic Peptide Membrane Permeability and PPB Rate Prediction Methods

2.1 Introduction

The development of cyclic peptides as drugs is often hindered by relatively poor membrane permeability and PPB rate, which are crucial factors influencing their bioavailability. Traditional experimental approaches to evaluate these properties are time-consuming and resource-intensive, prompting the need for fast and reliable computational prediction methods.

In this chapter, we explore the various computational approaches that have been developed to predict the membrane permeability and PPB rate of cyclic peptides. These methods can be broadly categorized into two types: molecular simulation-based and quantitative structure-property relationship (QSPR)-based. While simulation-based methods offer detailed insights into the molecular mechanisms governing these properties, they are often computationally expensive. On the other hand, though QSPR-based methods (machine learning has been widely used in recent years) can provide faster predictions, they may suffer from limited generalizability due to the scarcity of experimental data.

This chapter will first delve into the state-of-the-art methods for predicting cyclic peptide membrane permeability, discussing the strengths and limitations of both MD and ML approaches. We will then shift focus to the prediction of PPB rates, highlighting the challenges unique to cyclic peptides and how computational models initially

developed for small molecules have been adapted for these more complex structures. Finally, we will provide an overview of the recent advancements in DL methods, which are rapidly becoming a dominant force in small molecule properties prediction, and discuss their potential application to cyclic peptides.

2.2 Cyclic Peptide Membrane Permeability Prediction Methods

Computational methods for predicting the membrane permeability of cyclic peptides can be broadly classified into two types: methods based on MD and methods based on ML.

2.2.1 MD-based methods for cyclic peptide membrane permeability prediction

Computational approaches to predict the permeability of cyclic peptides have been primarily based on MD simulation [17, 73, 85, 86, 87]. MD-based predictions do not rely excessively on existing peptide information and can predict novel peptides. Since the membrane permeability of cyclic peptide depends on the conformation during the membrane permeation process, most MD research involves identifying conformations in aqueous solutions or organic solvents [85, 86, 87]. However, the conformations of cyclic peptides in aqueous and hydrophobic environments are always different. Moreover, the conformational transitions are slow relative to simulation time scales due to the high energy barrier between conformations [87]. As a result, many accelerated sampling techniques have been used, such as replica-exchange MD (REMD) [88], metadynamics (MetaD) [89], accelerated MD (aMD) [90], and multicanonical MD (McMD) [91]. Sugita *et al.* performed complex studies simulating the membrane permeation process of cyclic peptides across a lipid bilayer (Fig. 2.1) utilizing replica-exchange umbrella sampling (REUS) method [17, 73]. In addition, analyzing simulation results using methods such as Markov state models (MSMs) can provide information on the behavior of cyclic peptides, which is crucial for elucidating the mechanism of membrane permeation and structural optimization to increase membrane permeability [23, 86]. Markov state models were used to analyze simulation data and elucidate cyclic peptide behavior, which is crucial to understanding membrane permeation mechanisms and optimizing structures to enhance membrane permeability. Conventional MD-based

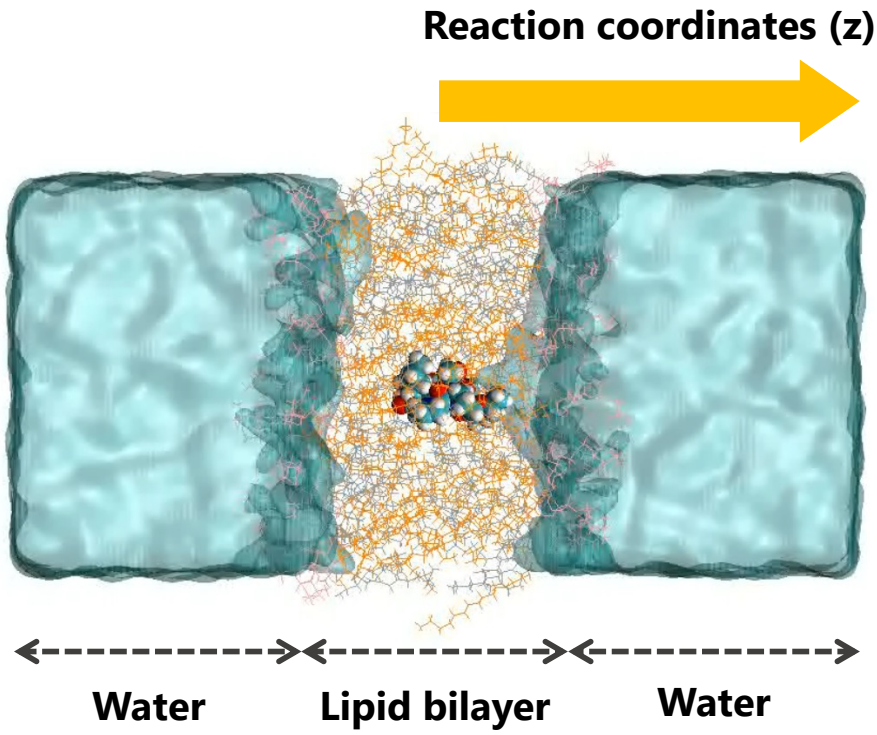


Figure 2.1: Image of cyclic peptide across a lipid bilayer during MD simulation.

methods could only predict relatively small cyclic peptides, such as six residues [85, 87]. Recently, protocols have been developed to predict cyclic peptides over ten residues accurately ($R > 0.8$) [73]. However, the high computational cost of MD-based methods is also a major limitation. For example, the latest method proposed by Sugita *et al.* requires parallel calculations on 224 GPUs, with a computational time of approximately 90 hours per peptide using NVIDIA P100 GPU [73].

2.2.2 ML-based methods for cyclic peptide membrane permeability prediction

In contrast to MD-based methods, several physicochemical or machine learning models have been developed, offering more rapid prediction capabilities [66, 92, 93, 94, 95, 96, 97, 98, 99]. The 2D descriptors for hydrophobicity, such as the octanol-water partition coefficient (LogP), are generally the most important features for prediction [94, 96]. Some studies have reported that the prediction accuracy can be improved by using 3D descriptors such as solvent-accessible surface area (SASA) [93, 95, 96, 97], suggesting that it is necessary to capture the information of 3D structures of cyclic

Table 2.1: Summary of ML-based permeability prediction methods for cyclic peptides. The data type abbreviations are as follows: CP (cyclic peptides) and MC (macrocycles). The model type abbreviations are as follows: SLR (simple linear regression), MLR (multiple linear regression), PLS (partial least squares), SVM (support vector machine), and RF (random forest). The accuracy abbreviations are as follows: R^2 (coefficient of determination) and RMSE (root mean square error).

Group	Data type	N	Model type	R^2	RMSE
Rezai *et al.* [92]	CP	11	SLR	0.96	-
Wang *et al.* [66]	CP	62	MLR	0.63–0.70	-
Leung *et al.* [93]	MC+CP	201	MLR	0.51	0.41
Over *et al.* [94]	MC	> 200	PLS, SVM, RF	0.81–0.84	0.41–0.50
Rossi *et al.* [95]	MC	19	MLR	-	0.71
Digiesi *et al.* [96]	CP	62	PLS	0.46–0.76	0.46–0.60
Poongavanam *et al.* [97]	MC	70	MLR	-	0.33–0.56
Williams *et al.* [98]	MC+CP	1,061	RF	0.47–0.57	-

peptides. However, unlike the property prediction methods for small molecules, which have considerable experimental data, these models were established using limited data sets (Table 2.1). Moreover, biases introduced by different research groups further exacerbate this issue, as molecule type and evaluation criteria often vary significantly. As a result, their reported prediction accuracies, while useful for assessing model performance within a specific dataset, can not be a reliable indicator of the model's generalization performance across the broader landscape of cyclic peptides. For instance, Digiesi *et al.* used 62 cyclic hexapeptides with very similar structures to each other [96]. The topological polar surface area (TPSA) of these peptides ranged from approximately 150 to 250 $Å^2$, which could only cover a very narrow chemical space. The lack of available databases that collect structurally diverse cyclic peptides is a major reason for poor generalization performance and is currently the greatest obstacle to developing comprehensive machine learning predictions. The combinations of amino acids that comprise cyclic peptides are numerous; therefore, the chemical space of possible cyclic peptides is very large. Furthermore, even a single residue change in the amino acid sequence can lead to drastic changes in membrane permeability [100]. Most studies using the available data fixed the majority of structures and measured changes in membrane permeability with very few residue changes. Therefore, building a machine learning prediction model with high generalization performance requires a large amount of training data collected from a larger body of publications than employed in previous studies. This dataset should include various peptides with diverse struc-

tures and peptides in which small structural changes have a large effect on membrane permeability. Additionally, these methods directly apply whole-molecule features, typically used to predict the membrane permeability of small molecules while ignoring the unique structural characteristics of cyclic peptides, such as sequence information and circularity.

2.3 Cyclic Peptide PPB Rate Prediction Methods

Computational methods for predicting the PPB rate of cyclic peptides have made little progress.

2.3.1 Summary of small molecule PPB rate prediction methods

Previous PPB rate prediction methods for small molecules can be classified into the following three types: ML-based methods [68, 101, 102, 103, 104, 105], docking-based methods [106, 107], and composite methods that use both ML- and docking-based methods [108, 109, 110]. Similar to ML-based membrane permeability prediction methods, the 2D descriptors of the hydrophobicity index, such as the octanol-water partition coefficient (LogP), are generally the most important features for ML-based PPB prediction methods [67]. Docking-based methods predict PPB rates using information such as docking scores obtained from docking simulations of plasma proteins (basically HSA) and ligand molecules. In addition to conventional rigid-body docking, some studies used induced-fit docking to consider the flexibility of HSA [107]. Composite methods construct ML models based on descriptors calculated from compounds, as well as docking scores or docking descriptors that provide more information obtained from docking simulations.

2.3.2 Cyclic peptide PPB rate prediction based on small molecule data

Cyclic peptides have a relatively large structure, and therefore the space of possible conformations is vast compared to that of small molecules, making it difficult to search for an energetically stable conformation. Furthermore, most docking software is designed for small molecules, and the crystal structures of cyclic peptides are also limited. This makes applying docking-based and composite methods for small molecules to

cyclic peptides difficult. Thus, finding descriptors that can be used for both cyclic peptides and small molecules has been a major research direction for the computational prediction of the PPB rate of cyclic peptides. Tajimi *et al.* constructed a prediction model using 1,211 experimental data of small molecules and made PPB rate predictions for the 24 public DrugBank cyclic peptides and 16 in-house cyclic peptides [111]. Since the biophysical mechanism of PPB can be expected to be similar for both small molecules and cyclic peptides, their study focused on selecting descriptors with high generalizability from small molecule training data utilizing enumerating lasso solutions and forward beam search techniques. However, their method proposed still has lower prediction accuracy (mean absolute error (MAE) = 21.6, correlation coefficient (R) = 0.46) than that of the traditional prediction methods for small molecule compounds and is not accurate enough for practical use.

2.4 Overview of DL-based Small Molecule Properties Prediction Methods and Potential Application to Cyclic Peptides

DL-based small molecule properties prediction methods based on graph neural networks (GNNs) and transformers have become a major research area. By representing atoms as nodes and bonds as edges, the GNN-based method can effectively capture molecular structural information and integrate the topological structure of molecules with complex atomic features. Nonetheless, most existing approaches, such as GCN (graph convolution network) [112, 113], GAT (graph attention network) [114, 115], and MPNN (message passing neural network) [116] have intrinsic limitations, including a poor ability to process global information and risk of over-smoothing when many atoms are present. In contrast, many transformer-based methods have been proposed that treat SMILES as strings following the successful experience in natural language processing [117, 118, 119]. Since these methods lack 2D structural information, several methods have been developed using molecular graph representations as input that can encode more complex structural information than strings [120, 121, 122, 123]. Furthermore, some studies reported that combining multi-scale molecular features, such as fragment-level and atom-level information, can improve the prediction accuracy [123, 124].

However, successfully applying these DL methods to predict cyclic peptide properties with high accuracy still faces several challenges, primarily due to the scarcity of

experimental data and structural complexity. Therefore, future development should include designing models tailored specifically to the characteristics of cyclic peptides and constructing larger, more diverse datasets to improve the generalization performance and predictive accuracy. Overall, although DL methods are still in the early stages of application to cyclic peptides, their successful application in small molecules suggests they hold significant potential for advancing cyclic peptide drug discovery.

2.5 Summary

In this chapter, we introduced typical computational approaches to predict the membrane permeability and PPB rate of cyclic peptides. We also provided an overview of the recent advancements in DL-based small molecule properties prediction methods and discussed their potential application to cyclic peptides.

Chapter 3

Multi-Level Molecular Features Design and Data Augmentation for Cyclic Peptides

3.1 Introduction

The structural complexity and unique characteristics of cyclic peptides present significant challenges in accurately predicting their membrane permeability and PPB rate. Even a minor alteration in a single residue can lead to substantial changes in their membrane permeability and PPB rate. However, as we mentioned in Chapter 2, existing prediction methods focus on global molecular characteristics while overlooking subtle sequence variations. Furthermore, limited experimental data make it more difficult to build generalizable models.

In this chapter, we proposed multi-level molecular features and model architectures designed to concurrently capture the local sequence variations and global conformational changes in cyclic peptides. Additionally, due to the inherent data scarcity in biological datasets, especially for cyclic peptides, we also proposed data augmentation strategies for cyclic peptides designed to improve model training efficiency. The methods and techniques described here were applied in subsequent chapters to predict permeability (Chapter 4) and PPB rate (Chapter 5).

3.2 Overall of the Proposed Fusion Model

We proposed a fusion model that hierarchically combines multi-level molecular features to simultaneously capture local and global structural information of cyclic peptides. Fig. 3.1 shows the overall architecture of the proposed fusion model. We designed three-level representations of peptides and used each for three different sub-models to extract the peptide-, monomer-, and atom-level molecular representations. Each of the three levels plays an important role in capturing different aspects of cyclic peptide structure. Peptide-level features capture the global structure of the entire molecule, integrating information on the overall properties and substructural characteristics that correlate to some extent with membrane permeabilities and PPB rates. Monomer-level features offer a lower-level perspective by specifically capturing the sequence information of cyclic peptides, which is especially important for these molecules. Since small changes in individual residues can drastically alter the peptide's permeabilities and PPB rates, the monomer-level features allow the model to track these variations and understand their effects on the overall function. Finally, atom-level features are crucial for capturing both local and global molecular information. On the local scale, they provide fine-grained details such as bond types, atom types, and local chemical environments, allowing the model to detect subtle variations like changes in chirality or atomic bonding, which can significantly impact properties such as membrane permeability. At the global scale, atom-level features also contribute to understanding the overall molecular shape by representing the relative positions of atoms within the graph and 3D structure, helping the model capture the structural information of the cyclic peptide.

Initially, the SMILES representation of the peptide was divided into monomers (substructures, such as amino acid residues), and respective 3D conformations were generated from their structural formulas. Subsequently, peptide- and atom-level features were extracted from the peptide conformation and used as input for the peptide and atom models, respectively. Monomer-level features extracted from the monomer conformation were used as inputs for the monomer model. Finally, the three-level latent feature vectors extracted using the three sub-models were concatenated to derive the membrane permeability prediction values.

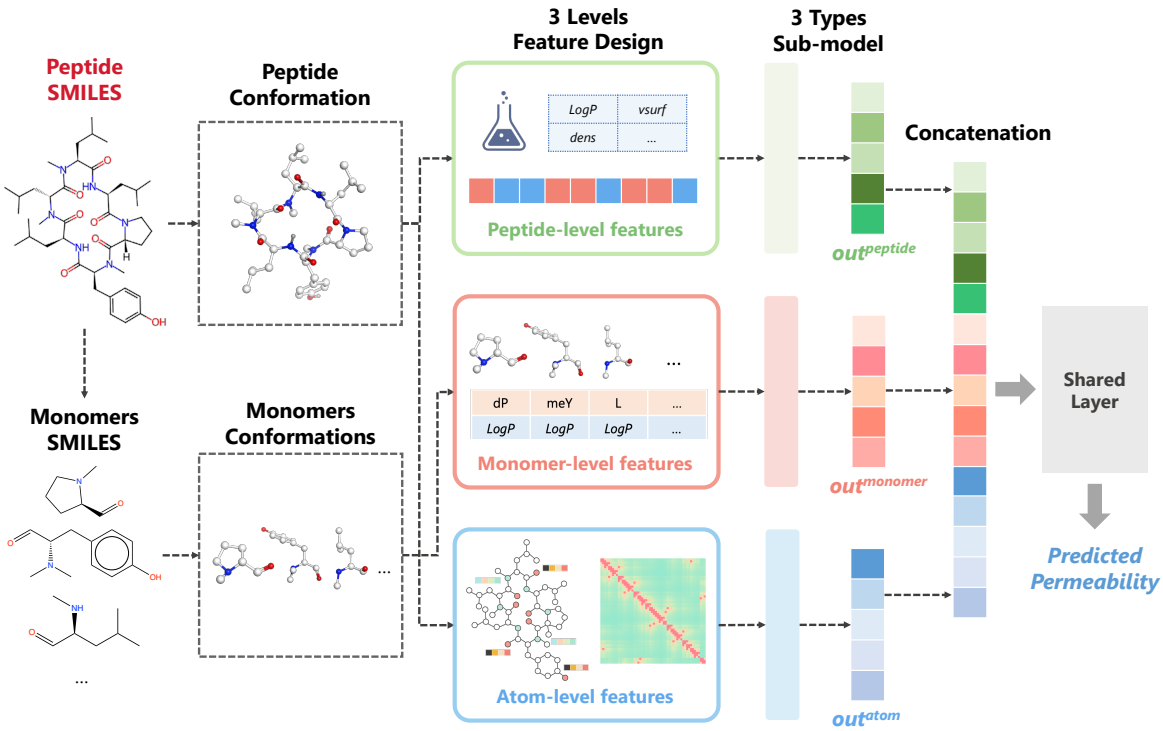


Figure 3.1: Overall framework of the proposed fusion model. The three-level expression vectors extracted using the three sub-models are concatenated and passed through a shared layer to derive the final permeability prediction value.

3.3 Peptide-Level Feature Design

To capture the characteristics of the entire molecule, we used peptide descriptors representing physicochemical properties and Morgan fingerprints (Morgan FP) representing substructural information as peptide-level features.

3.3.1 Conformation generation of peptides

The membrane permeability of cyclic peptides depends on their conformation during the membrane permeation process. The possible conformational space of the cyclic peptide is vast, making it challenging to identify an energetically stable conformation. Generally, MD simulations are considered a method for generating relatively precise conformations. However, simulating cyclic peptides involves numerous complex settings that must be carefully considered, making it far from a straightforward process. For example, determining the appropriate force field is essential, particularly for peptides containing non-canonical amino acids, as these require custom parameterization. Moreover, temperature control and other simulation parameters must be carefully tuned to the specific peptide system, adding further complexity. The biggest challenge with MD simulations of cyclic peptides is the high computational cost. When we generated the conformation of cyclic peptides using the commercial software MOE [125] with its Low mode MD method (a relatively simple MD method), it took 84 hours per cyclic peptide. Furthermore, from the experience of MD specialists in our lab, even with a very simple MD method, it takes approximately 10 to 20 hours to simulate one peptide. Therefore, to generate relatively accurate conformations for thousands of cyclic peptides using the MD method, it is important to use many GPUs to accelerate the computational process (several hours per peptide).

We generated the 3D conformations of peptides using a relatively simple method by the RDKit package (version 2022.09.5) [126] due to computational cost considerations, for 3D descriptors calculation. The computational time for this method was only three hours per peptide. The ETKDG (Experimental-Torsion Knowledge Distance Geometry) method is a stochastic algorithm widely used for generating 3D conformations of molecules [127]. It combines the classical distance geometry method with chemical knowledge, such as flat aromatic rings and linear triple bonds derived from experimental crystal structures [128]. This method ensures that generated 3D conformations are not only geometrically valid but also more likely to resemble physically plausible structures found in nature. With hydrogen atoms added to the SMILES representation of the cyclic peptide, the initial 3D structures are generated using the ETKDG method

(`AllChem.EmbedMolecule`) and subsequently optimized through energy minimization with the UFF force field (`AllChem.UFFOptimizeMolecule`). In this case, the maximum number of attempts to try embedding is set to 1,000,000, and additional torsion profiles for macrocycles are used (`useMacrocycleTorsions=True`).

3.3.2 Descriptor calculation of peptides

Subsequently, 2D and 3D descriptors of cyclic peptides are calculated from SMILES representations and single-conformation 3D structures, respectively. A total of 1,857 descriptors (1,689 2D and 168 3D descriptors) are calculated using MOE software (version 2019.01, 206 2D and 117 3D descriptors) [125], the Mordred package (version 1.2, 1,275 2D and 51 3D descriptors) [129], and RDKit package (208 2D descriptors). The full list of 1,857 descriptors is shown in Appendix B. The 2D descriptors are conformation-independent and are calculated from the SMILES representation of the molecule. These include descriptors derived from physical properties, such as water solubility indicators, and descriptors representing neighbor information when the molecule is viewed as a graph. Before calculating the 2D descriptors, hydrogen atoms are added, and the ionization state at pH 7.0 is estimated using the `Wash` function of MOE (`Protonation:Dominant`) when calculating MOE 2D descriptors. The 3D descriptors are calculated from the 3D structure of a molecule and depend only on relative coordinates within the molecule. Before calculating the MOE 3D descriptors, the charge of the RDKit-generated conformations is calculated using the `Partial Charges` function of MOE for correct calculation. MOPAC descriptors of the 3D descriptors (e.g., AM1 descriptors using the AM1 Hamiltonian) are not computed due to computational cost.

3.3.3 Preprocessing and selection of peptide descriptors

To remove the meaningless descriptors, we perform descriptors preprocessing as follows:

1. Descriptors with constant values among all cyclic peptides within the dataset are removed.

2. For descriptor pairs with the absolute value of correlation coefficient ($|R|$) > 0.9, the one with the lower correlation with the objective variable is excluded.

Table 3.1: Summary of feature selection methods in ML.

Type	Description
Filter method	Filter methods use univariate statistics to evaluate the relationship between explanatory and objective variables. Features are then ranked based on these scores to determine their suitability for use in prediction. These methods are faster than wrapper methods. A drawback is that the combined effect of multiple features is not considered. Methods for determining scores include analysis of variance (ANOVA), Chi-square test, and Fisher's score.
Wrapper method	Wrapper methods follow a greedy search approach by evaluating all the possible combinations of features. Wrapper methods usually result in better performance than filter methods. Its drawbacks include the tendency to overfit and the large amount of computation required. Typically, forward feature selection, which adds features, and backward feature elimination, which reduces features, are used.
Embedding method	Embedding methods simultaneously perform training and feature selection by ML algorithms. These methods combine the advantages of both wrapper and filter methods by accounting for interactions among multiple features while still keeping computational costs at a reasonable level. Commonly used algorithms include lasso regression, ridge regression, and random forests.

3. The remaining descriptors are standardized using Z-score as shown in Equation (3.1) (μ is the mean of descriptor x, and σ is the standard deviation of x).

$$z_i = \frac{x_i - \mu}{\sigma} \tag{3.1}$$

Performing further feature selection and using only important features can prevent overfitting and improve model interpretability. There are three main methods for feature selection in ML: filter, wrapper, and embedding methods (Table 3.1) [130]. Random forest (RF) is commonly used as an embedding method for robust feature selections, even with many variables. It can provide quantitative measures of the importance of each variable in prediction. Therefore, we constructed two RF models (`n_estimators=500`) with the 2D or 3D peptide descriptors, respectively. Subsequently, 2D and 3D peptide descriptors were selected based on the assigned feature importance.

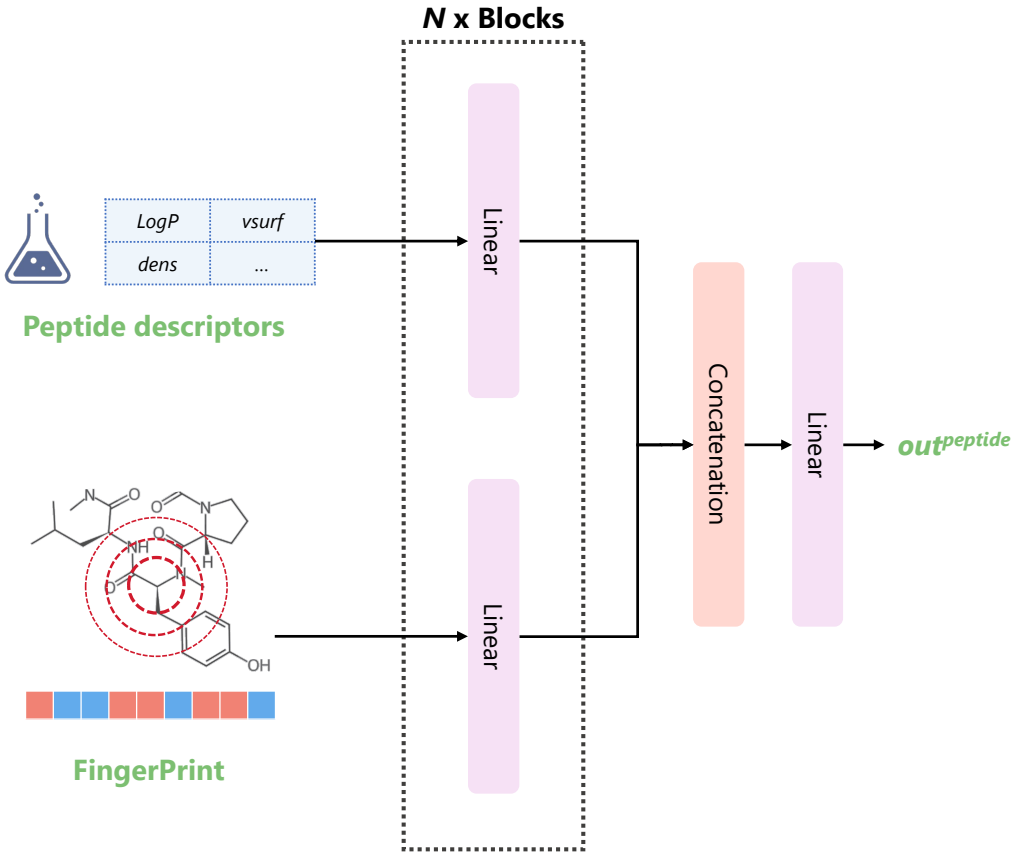


Figure 3.2: Architecture of the peptide model. Peptide descriptors and Morgan FP were used as input for the MLP-based model.

3.3.4 Architecture of the peptide model

In addition to selected peptide descriptors, we calculate 2048-bit Morgan FP (1024-bit, radius: 2; 1024-bit, radius: 3) to represent substructural information as peptide-level features. The selected peptide descriptors and Morgan FP are each trained with different MLPs (Fig. 3.2), and the latent feature vectors $x^{desc^{out}}$ and $x^{fp^{out}}$ are concatenated and used to derive the final output $out^{peptide}$ of the peptide model as follows:

$$out^{peptide} = \texttt{Linear}(\texttt{Concat}(x^{desc^{out}}, x^{fp^{out}})) \tag{3.2}$$

3.4 Monomer-Level Feature Design

Unlike small molecules, most cyclic peptides comprise a combination of monomers that are the standard building blocks in their chemical synthesis. Since the specific

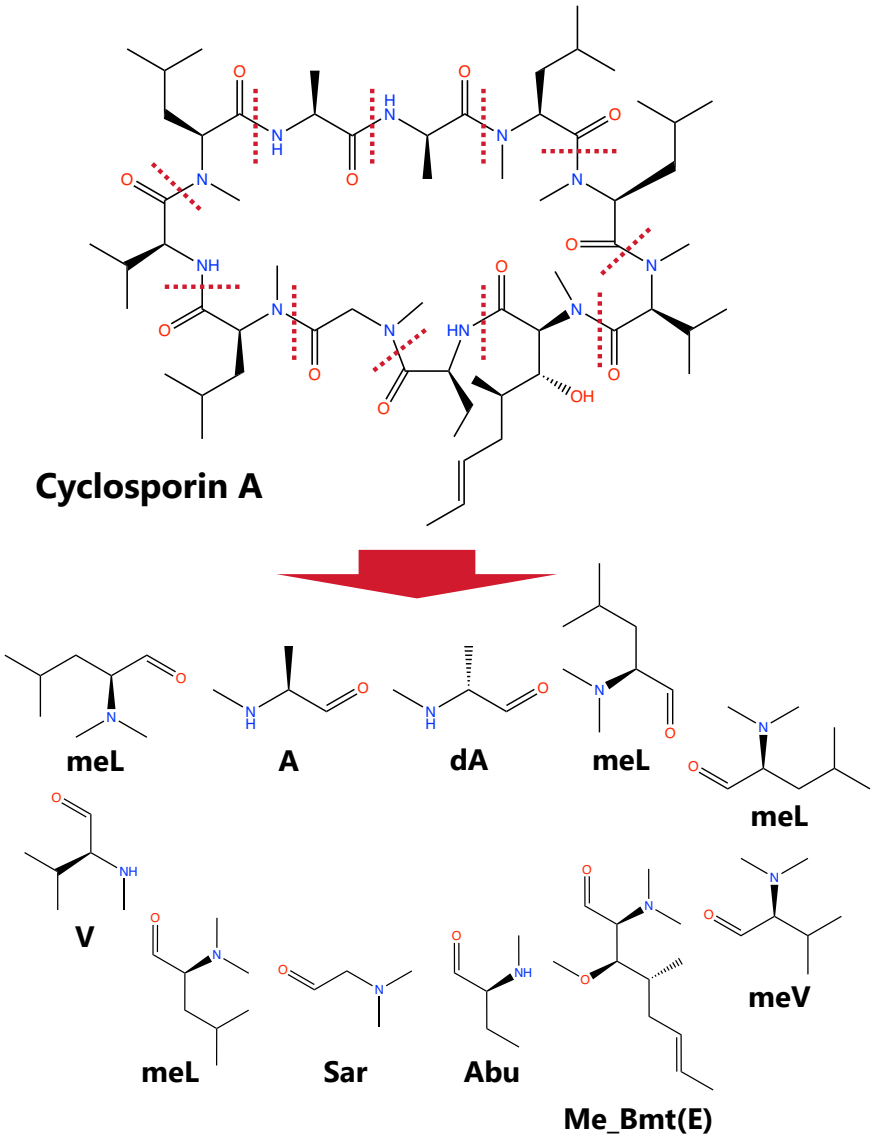


Figure 3.3: Example of the monomer division: Cyclosporin A and its 11 monomers.

local structure of a cyclic peptide has a great influence on the membrane permeability and PPB rate, most studies of cyclic peptides performed modifications at the monomer level. Therefore, we designed monomer-level features to capture the subtle sequence variations of cyclic peptides accurately.

3.4.1 Division of the main chain of cyclic peptide into monomers

We decompose the macrocycle ring of cyclic peptides into units of monomers and calculate descriptors from each monomer (Fig. 3.3). The division is applied for peptide, ester, and disulfide bonds, the most common bonds connecting monomers in the experimental data. Bonds existing anywhere other than the macrocycle (main chain) are not subjected to division to fully express the properties of the local structure. Merely hydrolyzing the peptide bond could generate a new hydrogen bond donor, potentially misrepresenting the substructure's original physicochemical properties. Hence, an appropriate capping is required when decomposing peptides into monomers. When generating the conformation and calculating the monomer descriptor, the cleaved amide group or O atom of the amide-to-ester substitution is methylated (addition of CH3), and the carboxyl group is converted to an aldehyde (addition of H). When dividing the disulfide bond, hydrogen atoms are added to both sulfur atoms.

3.4.2 Conformation generation and descriptor calculation of monomers

Similar to cyclic peptides, the 3D conformations of monomers are generated using the RDKit package for 3D descriptors calculation. On the other hand, compared to peptides, energetically stable conformations of monomers are easier to calculate. Therefore, multiple conformations (up to 200) are generated for each monomer using the ETKDG method (`AllChem.EmbedMultipleConfs`), and the conformation with the lowest energy is used after being optimized through energy minimization with the UFF force field. Then, 2D and 3D monomer descriptors are calculated using procedures similar to peptide descriptors and standardized using Z-score. Finally, the same monomer descriptors as the selected peptide descriptors mentioned in Section 3.3.3 are used.

3.4.3 Architecture of the monomer model

Monomer descriptors are aligned based on sequence information to generate the input feature map of the monomer model. If the peptide has fewer than 16 monomers, the blanks of the feature map are padded with zero. This input format can explicitly express the sequence information and is sensitive to changes in the local structure.

Since the CNN model can extract partial features, we thought it would suit our sequence-based input format and constructed a 1D-CNN-based monomer model. However, the conventional 1D-CNN cannot express circularity. To overcome this architecture-

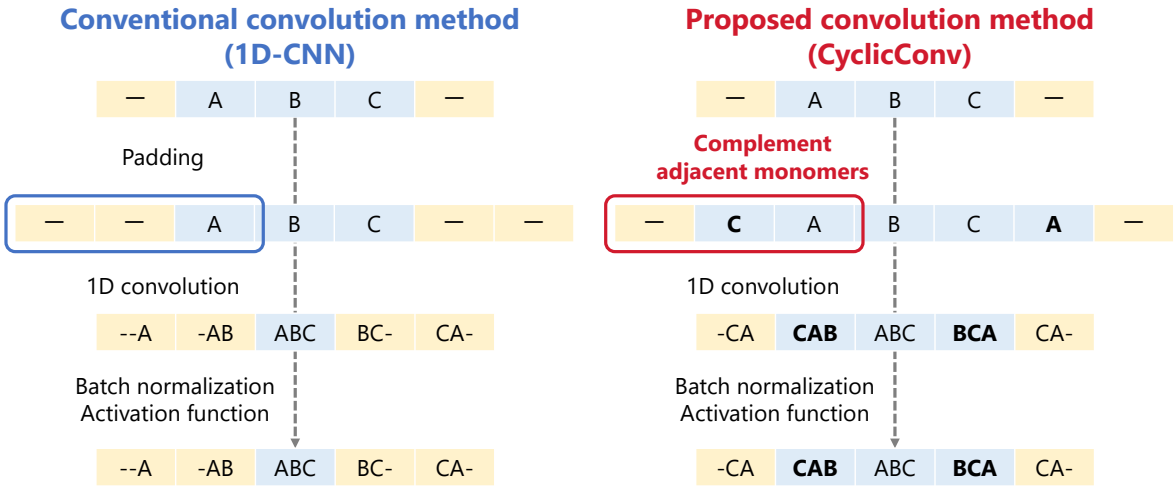


Figure 3.4: Comparison of the conventional convolution method (1D-CNN) and the proposed convolution method (CyclicConv). In the case of kernel size of three, monomer C is supplemented to the left of A and monomer A is supplemented to the right of C. By this operation, the information of CAB and BCA can be correctly acquired as a result of the CyclicConv.

level limitation, we proposed a new convolution method called CyclicConv that supplemented adjacent monomers at both ends of an input peptide sequence (Fig. 3.4). The architecture of the monomer model is shown in Fig. 3.5. The use of the CyclicConv or 1D-CNN layer is determined by hyperparameter tuning. Finally, the monomer model outputs the latent feature $out^{monomer}$.

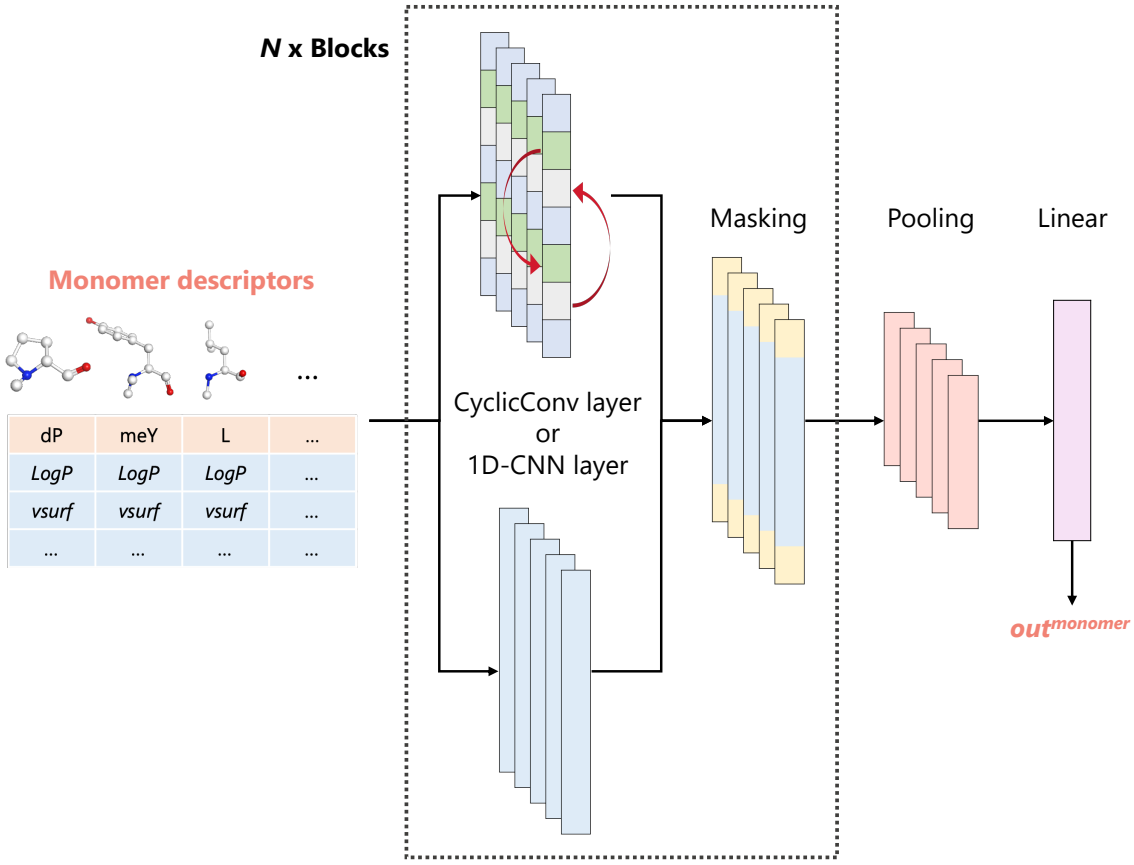


Figure 3.5: Architecture of the monomer model. Monomer descriptors are aligned based on the sequence information and used as input for the CNN-based model.

Table 3.2: Summary of atom-level features.

Type	Feature	Size	Description
Atom	Atom type	8	Type of atom by atomic number (one-hot)
	Degree of atom	5	Number of bonds (one-hot)
	Number of hydrogen	5	Number of bonded hydrogen atoms (one-hot)
	Formal charge	4	Electrical charge (one-hot)
	Hybridization	3	Type of atom hybridization (one-hot)
	Chirality	3	Type of atom chirality (one-hot)
	Is in a ring	1	Whether included in a ring structure
	Is aromatic	1	Whether aromatic
Bond	Bond type (*Bond*)	1	Single(1.0), Double(2.0), Triple(3.0), Aromatic(1.5), Conjugated(1.4) and No-bond(0)
Distance	Graph distance (*Graph*)	1	Distance calculated from graph representation
	3D distance (*Conf*)	1	Euclidean distance (Å) calculated from 3D conformation

3.5 Atom-Level Feature Design

To capture the detailed changes at the atom level of cyclic peptides, we have designed atom-level features that incorporate information from local atomic interactions to the shape of the entire molecule.

3.5.1 Atom-level features calculated from molecular graph representation

The SMILES representations of the cyclic peptides are first converted to a molecular graph using the RDKit package [126]. We designed atom-level features so that the atom model could capture minor changes, such as enantiomers, by node features (*Node*, 30 dimensions) and bond type matrix (*Bond*), and global changes in the entire molecule by two types of node-pair relative relationship matrices (*Graph* and *Conf*). As shown in Table 3.2, heavy atoms are considered nodes, and node features are represented as $Node \in \mathbb{Z}^{N_{atoms} \times 30}$, bonded interaction weights are represented as $Bond \in \mathbb{R}^{N_{atoms} \times N_{atoms}}$, graphic pairwise distances are represented as $Graph \in \mathbb{Z}^{N_{atoms} \times N_{atoms}}$, and 3D pairwise distances are represented as $Conf \in \mathbb{R}^{N_{atoms} \times N_{atoms}}$.

3.5.2 Architecture of the atom model

Since the transformer can effectively learn relationships between distant atoms even when the number of atoms is large (for example, the maximum number of heavy atoms

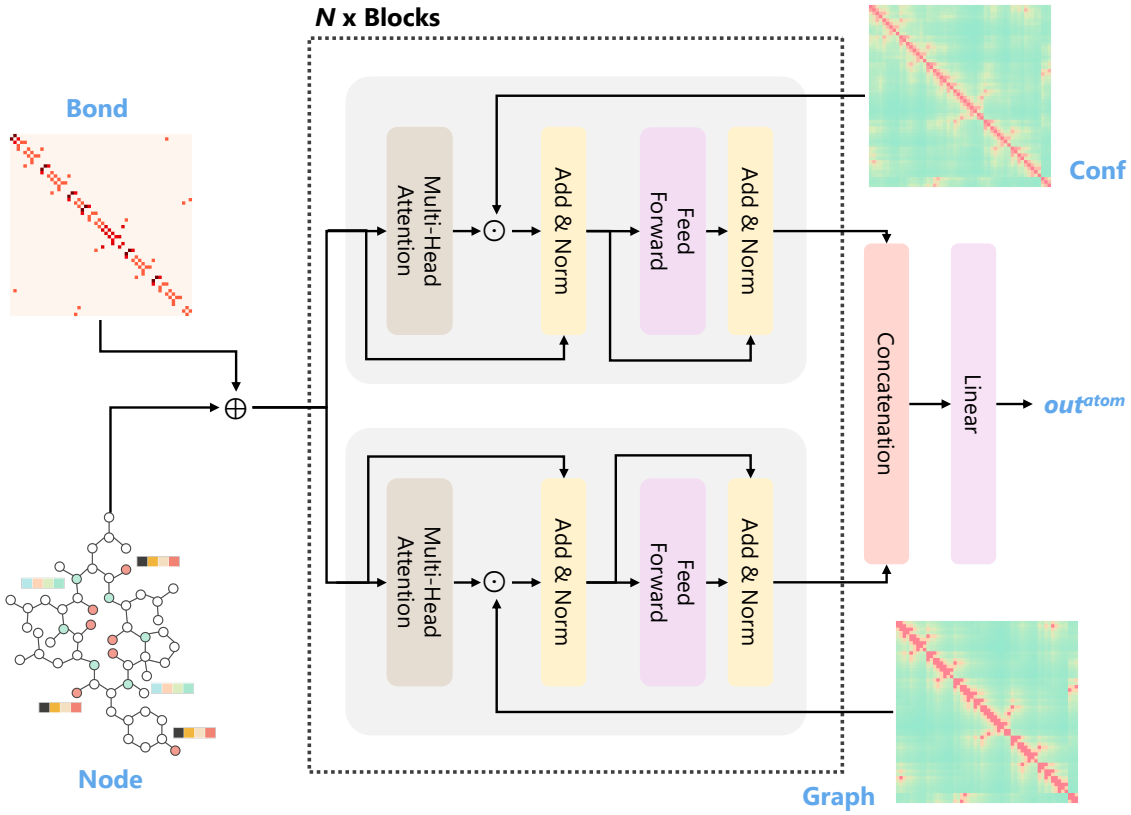


Figure 3.6: Atom model architecture. Node features and three types of node-pair relative relationship matrices are used as input for the transformer-based model.

in the experimental data of permeability was 128, and that in the experimental data of PPB rate was 162) [131], we constructed a transformer-based atom model to capture the overall graph structure and 3D conformation of the peptide (Fig. 3.6). *Bond* records the molecule bond information and controls message propagation between neighboring nodes by assigning weights to each bond type. According to the chemical bonding principle, bonds with more electron participation (such as unsaturated bonds) are assigned higher weights to enhance the exchange of information between atoms [122]. Meanwhile, many transformer-based models record the positional relationships between nodes or tokens using traditional absolute positional encoding [119, 132]. However, some studies have reported that relative positional encoding can improve prediction accuracy [121, 131]. To capture the local relationship between each node, embedded *Bond* is used for positional encoding and added to embedded *Node* to serve as the

input for the encoder block as follows:

$$x = \frac{W_{node}Node + W_{bond}Bond}{\sqrt{d_{model}}} \tag{3.3}$$

where d_{model} is the attention dimension of the atom model, $W_{node} \in \mathbb{R}^{N_{node\text{-}features} \times d_{model}}$ and $W_{bond} \in \mathbb{R}^{N_{atoms} \times d_{model}}$ are trainable parameters, and $x \in \mathbb{R}^{N_{atoms} \times d_{model}}$ is the updated input for the encoder block.

Two types of distance matrices, $Graph$ and $Conf$, i.e., the shortest pairwise graph distance and 3D Euclidean distance of each atom, are used to capture the peptide's overall structure and 3D conformation. The distance maps are processed through an attenuation function to weaken distant interactions as follows:

$$Strength_{i,j}^{graph} = \begin{cases} 1 & (i = j) \\ \frac{1}{Graph_{i,j}} & (i \neq j) \end{cases} \tag{3.4a}$$

$$Strength_{i,j}^{conf} = \begin{cases} 1 & (i = j) \\ \frac{1}{Conf_{i,j}} & (i \neq j) \end{cases} \tag{3.4b}$$

where $Graph_{i,j}$ and $Conf_{i,j}$ are the distance between atom pairs i and j from the graph representation and 3D conformation. Furthermore, we designed a structure-enhanced transformer encoder to learn the structural and 3D conformational information of peptides using focused attention. The encoder includes two blocks, one using $Strength^{graph}$ and another using $Strength^{conf}$. This approach attenuates attention between less relevant pairs based on the distance, providing a simplified approach to modeling complex molecular structures as follows (the case of *graph* block):

$$Q_i = x^{graph^{l-1}} W_i^Q,\ K_i = x^{graph^{l-1}} W_i^K,\ V_i = x^{graph^{l-1}} W_i^V \tag{3.5a}$$

$$head_i = \texttt{Softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_{model}}}\right) V_i \tag{3.5b}$$

$$multihead = \texttt{Concat}(head_1, ..., head_h) W^O \tag{3.5c}$$

$$focus = multihead \odot Strength^{graph} \tag{3.5d}$$

$$residual = \texttt{LayerNorm}(x^{graph^{l-1}} + focus) \tag{3.5e}$$

$$x^{graph^l} = \texttt{LayerNorm}(residual + \texttt{FFN}(residual)) \tag{3.5f}$$

where $x^{graph^{l-1}}$ and x^{graph^l} are the updated latent features of the *graph* block in $(l-1)$-th and l-th layers, respectively, h is the head number of multi-head attention, $W_i^Q \in \mathbb{R}^{d_{model} \times d_{model}/h}$, $W_i^K \in \mathbb{R}^{d_{model} \times d_{model}/h}$, $W_i^V \in \mathbb{R}^{d_{model} \times d_{model}/h}$, and $W^O \in \mathbb{R}^{d_{model} \times d_{model}}$

are trainable parameters. In the case of the $conf$ block, x^{conf^l} can be calculated from $x^{conf^{l-1}}$ and $Strength^{conf}$ in the same process as the $graph$ block.

Finally, the outputs $x^{graph^{out}}$ and $x^{conf^{out}}$ of the two blocks were weighted using the hyperparameter λ_g and the concatenated feature vector was used to derive the final output out^{atom} of the atom model as follows:

$$out^{atom} = \texttt{Linear}(\texttt{Concat}(\lambda_g * x^{graph^{out}}, (1-\lambda_g) * x^{conf^{out}})) \tag{3.6}$$

3.6 Architecture of the Fusion Model

As shown in Fig. 3.1, the output latent feature vectors out^{atom}, $out^{monomer}$, and $out^{peptide}$ of the three sub-models are concatenated to generate the final molecular feature vector, which is passed through a shared layer to derive the final permeability prediction value out^{fusion}. As the model becomes more complex, problems such as the vanishing gradient problem may occur, causing input information to not be transmitted. Auxiliary loss is a technique in which additional losses are incurred to optimize the neural network (NN) training process. Directly propagating errors to the middle network layer can prevent gradient vanishing and improve embedding and training efficiency [133]. Hence, we designed the three sub-model losses L_{atom}, $L_{monomer}$, and $L_{peptide}$ derived from the output of each sub-model (the definition of $L_{monomer}$ is shown in Equation (3.7a)), and layer losses L_{layer_a}, L_{layer_m}, and L_{layer_p} derived from the averaged outputs of the layers in each block (transformer, CNN, and MLP) of the three sub-models (the definition of L_{layer_m} is shown in Equation (3.7b)) in addition to the main loss L_{fusion} calculated from the output of the fusion model:

$$L_{monomer} = \texttt{Lossfunc}(\texttt{Linear}(out^{monomer})) \tag{3.7a}$$

$$L_{layer_m} = \texttt{Lossfunc}(\texttt{Linear}(\texttt{Mean}(x^{mono^1}, ..., x^{mono^{num\text{-}cnn}}))) \tag{3.7b}$$

The loss function during training is presented in Equation (3.8):

$$\begin{aligned} Loss = L_{fusion} &+ \gamma_{sub} * (L_{atom} + L_{monomer} + L_{peptide}) \\ &+ \gamma_{layer} * (L_{layer_a} + L_{layer_m} + L_{layer_p}) \end{aligned} \tag{3.8}$$

where the weight parameter γ_{sub} is set to 0.10 and γ_{layer} is set to 0.05. Only the output value out^{fusion} of the fusion model is used during inference.

3.7 Data Augmentation for Cyclic Peptides

Thus far, we have proposed multi-level molecular features and integrated them into advanced model architectures, enabling us to incorporate domain knowledge to better capture the complex characteristics of cyclic peptides. Although the amount of available biological data has increased, experimental data remains limited compared to data for natural language processing and computer vision. For example, the Tox21 dataset deals with the toxicity classification of small molecules and has only approximately 8,000 data points. This limitation in biological data, particularly the scarcity of data with measurement values, has motivated the increased use of self-supervised learning approaches, such as contrastive learning [134] and pre-training [117, 119]. These methods are commonly employed in scenarios where labeled data is scarce while large amounts of unlabeled structural data are available. However, these techniques remain challenging for cyclic peptides given a more limited availability compared to small molecules. Apart from these techniques, data augmentation (such as oversampling and data warping) has been commonly used in the image processing field to increase training efficiency when the data are insufficient. The augmented data represent a more comprehensive set of possible data points that minimizes the distance between the training and any future testing sets and reduces the risk of overfitting [135].

We proposed the data augmentation method dedicated to cyclic peptides to generate 60 replicas based on the complexity of cyclic peptide conformational changes, the nature of cyclic peptide sequences, and the properties of SMILES to improve the training efficiency of the model. First, considering the complex conformational changes during membrane permeation of cyclic peptides, 60 different conformations per peptide/monomer are generated using RDKit to incorporate more diverse 3D information into the model (Fig. 3.7). Cyclic peptide conformations are used to calculate the $Conf$ matrix for the atom model and 3D peptide descriptors for the peptide model. Monomer conformations are used to calculate the 3D monomer descriptors for the monomer model. Next, the input of the monomer model is rearranged using sequence arrangement considering the circularity of the cyclic peptide—the aligned monomer descriptors are augmented by the combination of sequence translation and rotation (change the start point of sequence) as shown in Fig. 3.8. Here, the peptide with the most monomers in the experimental data of permeabilities and PPB rates in Chapter 4 and Chapter 5 comprises 15 monomers. Nevertheless, to reduce the duplication of the replica generated by augmentation, padding for one monomer is added, and the input sequence length (max_len) of the monomer model is set to 16. Finally, the SMILES

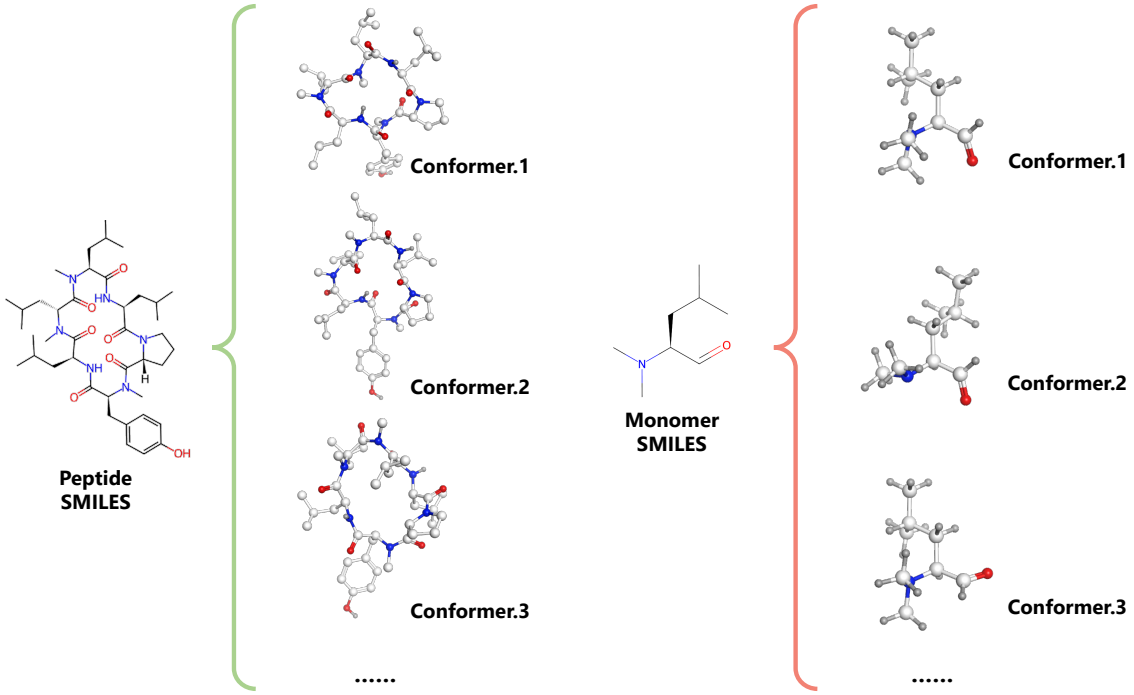


Figure 3.7: Peptide- and monomer-levels data augmentation based on using multiple conformations. We use 60 different conformations per peptide/monomer to incorporate more diverse 3D information.

enumeration technique [136] is used to permute the atom order and generate input for the atom model with a different ordering (Fig. 3.9). Introducing variations of the input data enables our model to become more robust to cyclic peptide conformational flexibility and allows it to partially consider circularity. Data augmentation effectively increases the size of the training set, leading to more efficient and stable training. During training, each replica is given the same label and treated as independent data. During inference, relying on a single conformation could introduce bias, considering the conformational flexibility of cyclic peptides. Therefore, during inference, 60 replicas are also generated for each peptide, and predictions are made individually. The average of 60 replicas is used as the final predicted value to represent the conformational ensemble of peptides.

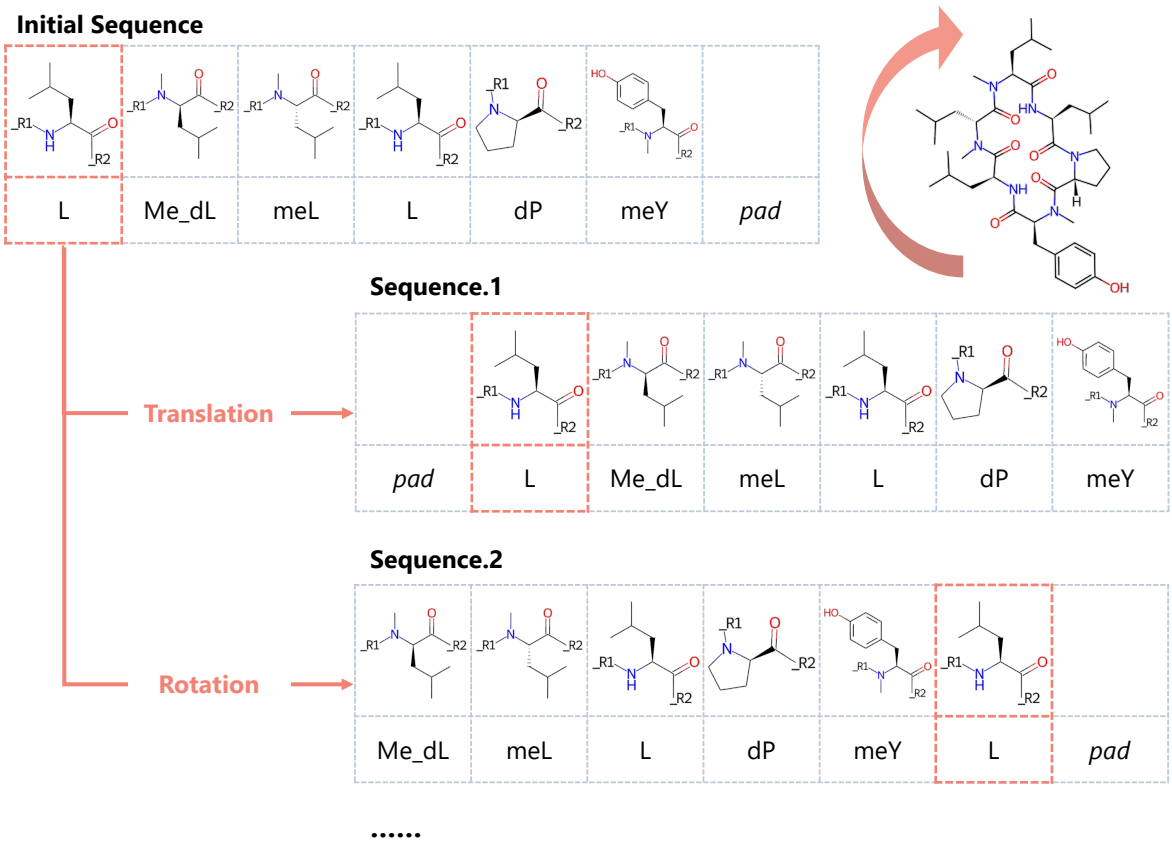


Figure 3.8: Monomer-level data augmentation based on sequence arrangement. The aligned monomer descriptors are translated and rotated based on the sequence information. The number of replicas of a cyclic peptide consisting of n monomers is $n \times (max\_len(16) - n + 1)$.

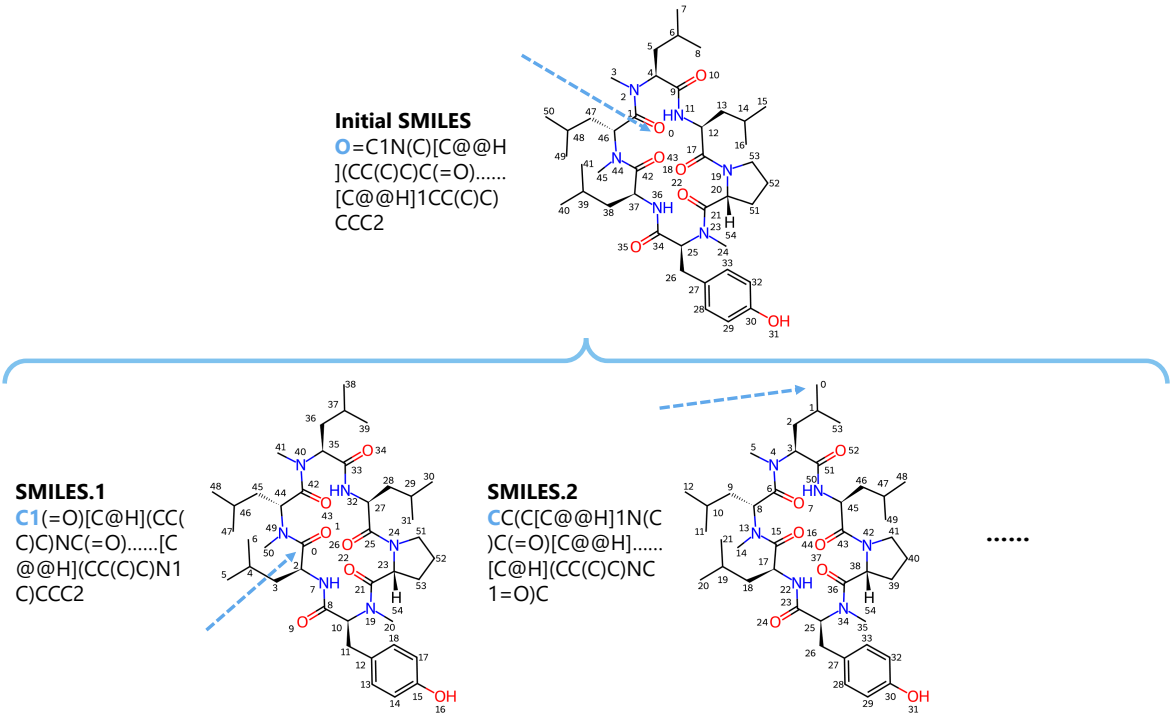


Figure 3.9: Atom-level data augmentation based on SMILES enumeration. Molecular graphs with different atoms order are constructed from different SMILES representations.

3.8 Summary

In this chapter, we introduced a novel model for cyclic peptides that effectively integrates multi-level features with state-of-the-art DL techniques. Considering the findings that cyclic peptide permeability is influenced by both their global conformation and local structural features, while PPB is heavily affected by local interactions, we engineered features at the peptide-, monomer-, and atom-levels to hierarchically capture both the global and local structures of cyclic peptides. We also proposed various data augmentation techniques to address the challenges posed by the limited availability of experimental data for cyclic peptides. The methods were applied in subsequent chapters to predict permeability (Chapter 4) and PPB rate (Chapter 5).

Chapter 4

Development of a Membrane Permeability Prediction Model of Cyclic Peptides (CycPeptMP)

4.1 Introduction

As we mentioned in Chapter 1 and Chapter 2, a rapid cyclic peptide membrane permeability prediction method with high generalization performance still does not exist and is one of the bottlenecks in cyclic peptide drug discovery. Existing ML-based prediction methods suffer from a lack of experimental data and further fail to account for the unique membrane permeation mechanism of cyclic peptides.

To overcome these challenges, we presented CycPeptMP: an accurate and efficient cyclic peptide membrane permeability prediction method. First, we collected information on a total of 7,334 cyclic peptides, including the structure and experimentally measured membrane permeability, from 45 published papers and two patents from pharmaceutical companies, with the intention of developing a DL-based permeability prediction model with high prediction and generalization performance. We then applied the novel multi-level molecular features design and data augmentation methods described in Chapter 3. CycPeptMP demonstrated better prediction performance than existing methods; its code has been published on GitHub (`https://github.com/akiyamalab/cycpeptmp`).

4.2 Development of a Comprehensive Database of Membrane Permeability of Cyclic Peptides (CycPeptMPDB)

4.2.1 Data collection

Although many experimental data on the membrane permeability of cyclic peptides have been reported, a comprehensive database is not yet available. We collected a total of 7,334 structurally diverse cyclic peptide data points (the number of peptides including duplicated structures from all publication sources was 7,451) and their measured membrane permeability from 45 published papers and two patents from pharmaceutical companies. We believed that these data could contribute to the entire field of cyclic peptide drug discovery, and therefore made it available as the first web-accessible database of cyclic peptide membrane permeability, CycPeptMPDB (`http://cycpeptmpdb.com`). The source list of CycPeptMPDB is shown in the Table 4.1. Membrane permeability in CycPeptMPDB is expressed as $\mathrm{LogP}_{\mathrm{exp}}$, the logarithm of experimentally determined permeability. For peptides whose membrane permeability could not be measured due to the detection limit, etc., $\mathrm{LogP}_{\mathrm{exp}}$ was set to the minimum value, -10.0 (1×10^{-10}cm/s, detailed records such as the detection limit can be viewed on the peptide detail page). The structures of peptides were recorded in CycPeptMPDB using the SMILES notation. Structural errors in the original publication were corrected (for example, when the SMILES structure attached to the publication differs from the sequence described in the publication, the structure was corrected based on the sequence information). When there was a new source directly citing the membrane permeability values of the old source, the number of these data was counted only in the old source. As shown in Table 4.1, most previously reported studies included a small number of peptides, with only six publications reporting over 100 peptides. In addition, as mentioned above, most of the publications deal with structurally similar cyclic peptides; therefore, the molecular weight range of peptides reported in a single paper is narrow. By collecting more than 40 publications, CycPeptMPDB covers a wide range of cyclic peptides with molecular weights ranging from 342.4 to 1777.7 and TPSA from 73.0 Å^2 to 702.0 Å^2. Molecular weight and TPSA were calculated by the MolWt and TPSA descriptors calculated using the RDKit package, respectively. Interestingly, over 99.6% of cyclic peptides include non-natural amino acids, suggesting that they were created to enhance permeability through chemical modifications, such as N-methylation, or by deliberately

Table 4.1: Source literature list of CycPeptMPDB. The number of peptides, molecular weight range, and assay type of membrane permeability for each source are shown.

Source	N	MW	PAMPA	Caco-2	RRCK	MDCK
2006_Rezai_1 [137]	10	712.9–1202.6	√			
2006_Rezai_2 [92]	11	710.9–840.1	√			
2011_White [7]	10	712.9–1202.6			√	
2012_Rand [138]	16	712.9–828.1		√	√	
2013_CHUGAI [14]	878	813.0–1777.7	√			
2013_Zaretsky [139]	2	471.6–627.8		√		
2014_Nielsen [140]	4	709.9–778.0			√	
2015_Ahlbach [141]	34	414.5–1620.7	√			
2015_Bockus_1 [142]	16	755.0–1202.6	√		√	
2015_Bockus_2 [143]	17	707.9–778.0	√	√		
2015_Hewitt [100]	18	712.9–755.0		√		
2015_Lewis [61]	2	712.9–755.0	√			√
2015_Marelli [144]	10	454.5–849.0	√	√		
2015_Nielsen [145]	3	724.9–785.1	√			
2015_Schwochert [146]	13	712.9–793.0			√	
2015_Wang [66]	62	454.5–882.1	√	√		
2016_Fouché [147]	15	790.9–1199.6				√
2016_Frost [8]	12	542.6–699.9	√			
2016_Furukawa [15]	688	606.8–944.2	√	√		
2016_Hickey [148]	18	662.8–1202.6	√			
2016_Schwochert [149]	8	696.9–731.0				√
2017_Boehm [150]	14	848.1–929.1			√	
2017_Price [41]	2	1019.4–1202.6			√	
2017_Pye [151]	21	785.0–1151.5	√		√	
2018_Buckton [152]	19	596.7–744.8	√	√		
2018_CHUGAI [153]	374	1062.3–1664.2		√		
2018_García-Pindado [154]	4	725.9–883.7	√			
2018_Kaneda [155]	7	842.1–870.1	√			
2018_Lee [156]	6	639.7–653.7	√			
2018_Naylor [157]	81	578.8–1218.6	√		√	
2018_Ramalho [158]	10	767.0–851.0	√	√		
2019_Ono [85]	8	712.9–712.9	√			
2020_Barlow [159]	26	537.7–753.0	√			
2020_Furukawa [160]	36	987.3–1197.5	√			√
2020_Hosono [161]	11	712.9–727.0	√			
2020_Le Roux [162]	47	342.4–486.7	√	√		
2020_Townsend [163]	3,086	654.9–958.2	√			
2021_Comeau [164]	42	430.5–458.6	√	√		
2021_Golosov [165]	27	758.0–1076.5	√		√	
2021_Kelly [166]	1,519	974.3–1220.4	√			√
2021_Lee [58]	5	1160.6–1231.7	√	√		
2021_Wang [23]	24	959.2–1169.5	√			
2022_Bhardwaj [45]	136	622.8–1299.7	√	√		
2022_Lee [167]	24	1160.6–1251.7	√			
2022_Saunders [168]	11	542.6–623.7	√			
2022_Taechalertpaisarn [11]	52	661.8–856.1	√			
2022_Tamura [169]	12	792.0–950.2	√			

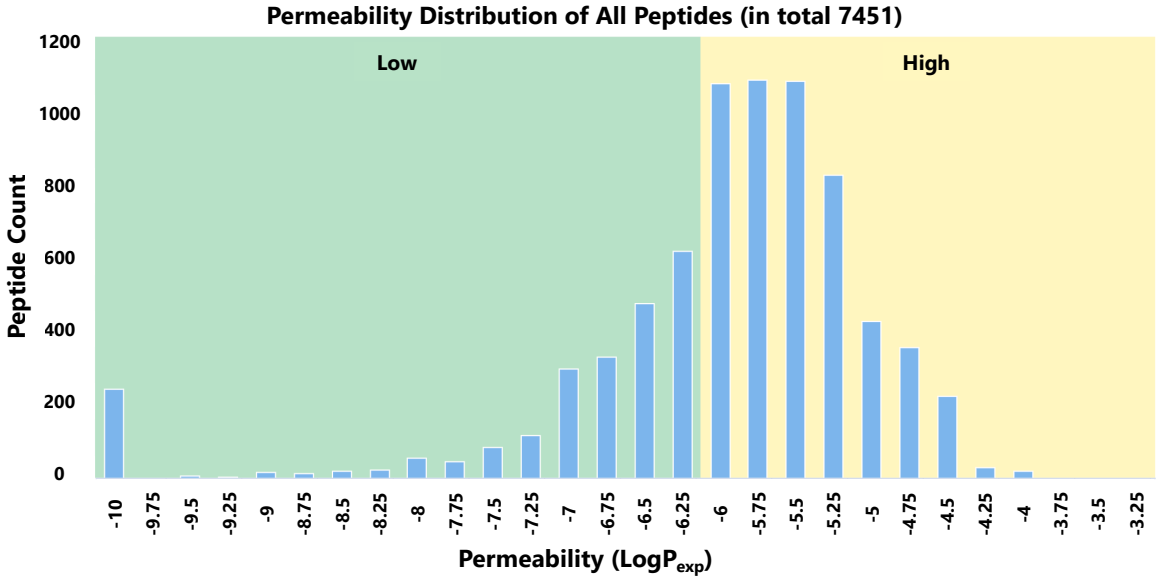
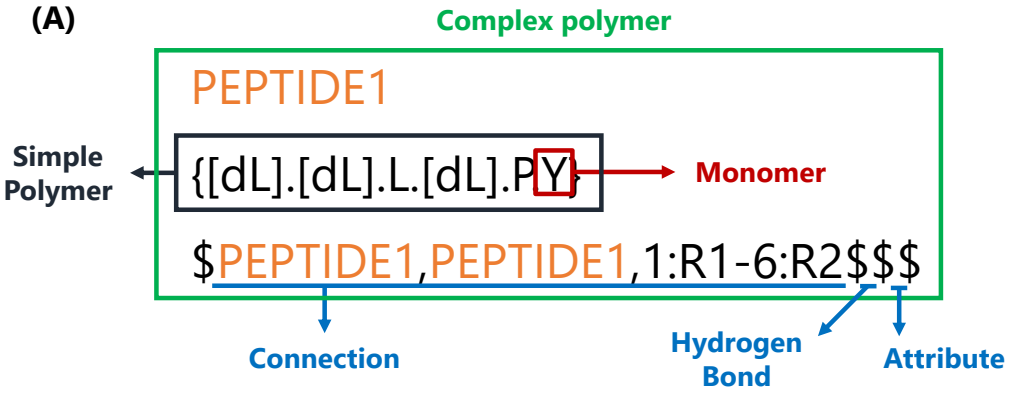


Figure 4.1: Permeability ($\mathrm{LogP_{exp}}$) distribution of all peptides. The background color of the high permeability ($\mathrm{LogP_{exp}} \geq -6.0$; 5,113 peptides) range is yellow, and the background color of the low permeability ($\mathrm{LogP_{exp}} < -6.0$; 2,338 peptides) range is green.

incorporating non-natural building blocks in their design. In the selected publications, there were 6,941 measurements by PAMPA, 649 measurements by Caco-2 assay, 40 measurements by MDCK assay, and 186 measurements by RRCK assay. All measured values were recorded when measurements obtained by multiple assays were reported in a single publication. Of all identified peptides, the membrane permeability measurements of 365 peptides were determined using two different assays. When a peptide was measured by two assays in a single publication, the permeability of the assay with more data was used as the representative membrane permeability measurement. If the permeability from both assays was similar, the representative value was determined according to the following assay rank: (1) PAMPA, (2) Caco-2, (3) RRCK, and (4) MDCK. The distribution of the representative permeabilities of all peptides (including duplicates) is shown in Fig. 4.1 Furthermore, a detailed description of the assay protocol used in each study was also recorded as experimental conditions such as reaction time and initial concentration affect membrane permeability measurements.



(B)

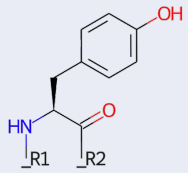
Structure		CXSMILES	Oc1ccc(C[C@H](N[*])C([*])=O)cc 1 |$;;;;;;;;_R1;;_R2;;;$|
		Symbol	Y
		Natural Analog	Y
		Polymer Type	PEPTIDE
		Monomer Type	Backbone
		Attachment Points	R1-H, R2-OH

Figure 4.2: (A) Example of HELM notation (PEPTIDE1{[dL].[dL].L.[dL].P.Y} $PEPTIDE1, PEPTIDE1,1:R1-6:R2$$$) and its constituent parts in CycPeptMPDB. If the simple polymer is a peptide, write the simple polymer as PEPTIDEx (where x is a number, and in the case of RNA is RNAx). The connection section means that R1 of the 1st monomer of PEPTIDE1 and R2 of the 6th monomer of PEPTIDE1 are connected. The hydrogen bonds and attributes sections of all peptides in this study are empty. (B) Example of monomer definition of tyrosine (Y).

4.2.2 Sequence representation of cyclic peptides

Cyclic peptides are relatively large compared to small molecules, and appropriate sequence representation is essential for good readability. Therefore, we used the hierarchical editing language for macromolecules (HELM) notation [170] to generate a unified sequence representation of collected cyclic peptides. HELM can hierarchically represent complex structures with relatively high molecular weights, such as antisense oligonucleotides, short interference RNAs, peptides, proteins, and antibody drug conjugates. HELM consists of four level hierarchies: complex polymer, simple polymer, monomer, and atom (Fig. 4.2 (A)). First, a complex polymer expresses information about the chemical structure of the entire macromolecule. Its components are simple

polymers and their connections (including hydrogen bonds and attributes). Second, a simple polymer is composed of monomers of the same polymer type. A simple polymer is defined as a single linear chain; branching and cycling structures are not covered in this hierarchy. Certain polymer types have explicit rules for connections between monomers, and the position and rules of connections can express the direction of monomer sequences (e.g., PEPTIDE notation represents amino acid sequences from N-terminus to C-terminus). Moreover, monomers are composed of atoms and bonds and can be represented in formats such as Molfile and CXSMILES (Chemaxon Extended SMILES) (Fig. 4.2 (B)). Each monomer was given a unique symbol similar to the amino acid code represented in the peptide sequences (e.g., A, G). Here, the definition of monomer also includes the positions of its connections (i.e., attachment points).

When describing linear peptides, the original HELM definition dictates that monomers are connected by peptide bonds, and the attachment point R1 is defined as the N atom of the amino group, R2 is defined as the C atom of the carboxyl group (the attachment points after R3 is the branch of the side chain). R1 and R2 in the terminal can only be used to form the main chain of the linear peptide (R1 and R3, R2 and R3 can form a ring). Contrary to the original definition, the N-terminus and C-terminus of linear peptides are often connected in the case of cyclic peptides. Therefore, in this study, N-terminal R1 and C-terminal R2 were able to be used to form a ring (like 1:R1-6:R2 in Fig. 4.2 (A), HELM in PubChem and ChEMBL databases are also like our definition). In addition, the O atom changed from the N atom was also set as R1 because there were many cyclic peptides with amide-to-ester substitutions.

4.2.3 Monomer definition in CycPeptMPDB

As mentioned in the previous section, the method used to define monomers that comprise peptides is important and should be standardized for all entries in the database. However, many of the selected publications did not record sequence representations, and even when records were available, the representation of special amino acids tended to differ notably between these publications. Therefore, we defined the partial structure obtained after cleaving the peptide bonds and ester bonds of the cyclic peptide as a monomer (Fig. 4.2 (B), CycPeptMPDB has no peptide containing disulfide bonds). As a result, a total of 312 types of monomers were obtained. The LogP (MolLogP descriptor calculated by RDKit software) distribution of all monomers is shown in Fig. 4.3. There were 305 monomers with the backbone monomer type (having two or more at-

Table 4.2: Explanation of symbols naming method.

Explanation of naming method	Example of symbol
1. Natural amino acids	A, L, dV
2. Monomers with a general compound name	Abu, Sar, dCha
3. Monomers with side chain modifications	Ala(tBu), dGlu(OMe), dPhe(4-F)
4. Monomers with N- terminal modifications	Me_Cha, Bn_Gly, 3-pyridylethyl_Gly
5. Monomers with C- terminal modifications	Glu_NH2
6. Monomers with amide-to-ester substitution	(N->O)Val, d(N->O)Val
7. Combination of above 1. to 7.	Me_Phe(3-Cl), Cys(EtO2H)_NH2
8. Terminal modification of cyclic peptides	ac-, -pip
9. Monomers could not named by 1. to 8.	Mono1 – Mono112

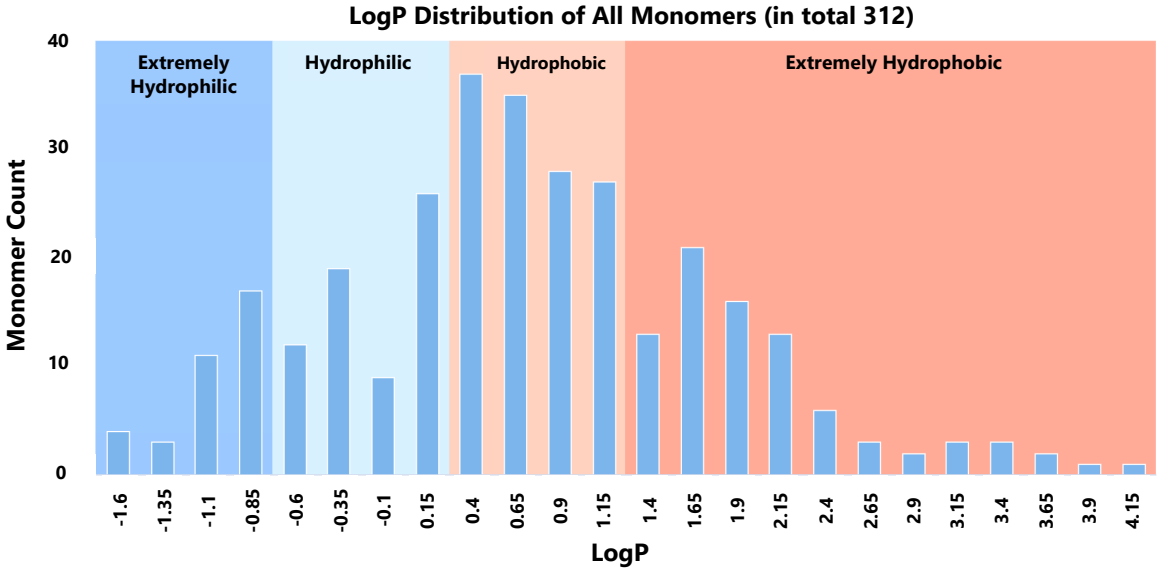


Figure 4.3: LogP distribution of all monomers. The background color for extremely hydrophilic monomers (LogP < -0.60, lower than G:-0.60; 35 monomers) is blue, hydrophilic monomers ($-0.60 \leq$ LogP < 0.40, lower than V:0.43, general hydrophilic amino acids, such as G:-0.60, A:-0.21, and P:0.28; 66 monomers) is light blue, hydrophobic monomers ($0.40 \leq$ LogP < 1.40, general hydrophobic amino acids, such as V:0.43, I:0.82, L:0.82, and F:1.02; 127 monomers) range is orange, and extremely hydrophobic monomers ($1.40 \leq$ LogP, extremely hydrophobic amino acids, such as W:1.50; 84 monomers) is red.

tachment points) and 7 monomers with the terminal monomer type (used for terminal modification of peptide sequences with only one attachment point). Monomers were described in CXSMILES to represent the positions of attachment points. For monomers

included in the PubChem database, their general compound and International Union of Pure and Applied Chemistry (IUPAC) names were recorded. When not included in the PubChem database, their IUPAC names were generated from SMILES using STOUT software (version 2.0, `https://github.com/Kohulan/Smiles-TO-iUpac-Translator`) [171]. Furthermore, when setting the symbol (the monomers short display name in HELM) and natural analog of monomers, we referred to the PubChem database and the monomer library of ChEMBL database (version 29, contained 2,851 types of monomers). At this stage, there were 112 types of monomers that did not have suitable symbols, and their symbols were set as Mono1 to Mono112. The explanation of the naming method of the symbol is shown in the Table 4.2. Additionally, we defined two types of peptide molecule shapes: Circle and Lariat. This classification was based on HELM sequence information. Peptides with cyclization positions at both the N- and C-terminal ends of the sequence were considered Circle peptides, and peptides with cyclization positions not at the end of the sequence, i.e., cyclized between a terminal and a side chain, were considered Lariat peptides.

More detailed descriptions of CycPeptMPDB, such as descriptions of the web pages, are shown in Appendix C.

4.2.4 Future update schedule for CycPeptMPDB

Since its release, CycPeptMPDB has been widely used by many research groups around the world, significantly contributing to the advancement of cyclic peptide drug discovery. The paper of CycPeptMPDB [82] has been cited by 19 papers within 18 months since its publication. We plan to continue to update the CycPeptMPDB in the future. Our near-term plan is to add the 409 peptides included in the new 8 papers (Table 4.3) that reported membrane permeability measurements of cyclic peptides to the database after the CycPeptMPDB was published, by the end of 2024. At the same time, we are also exploring implementing the function that allows users to upload their data, enabling a more collaborative use of CycPeptMPDB.

Table 4.3: List of source literature for CycPeptMPDB updates. The number of peptides and assay type of membrane permeability for each source are shown.

Source	N	PAMPA	Caco-2	RRCK	MDCK
2023_Ghosh [172]	36	√	√		
2023_Ohta [173]	22		√		
2023_Tanada [174]	45		√		
2024_Bergeron [175]	3	√	√		
2024_Faris [176]	234	√			√
2024_Kage [177]	39		√		
2024_Ly [178]	20	√	√		
2024_Otani [179]	10	√			

4.3 Materials and Methods

4.3.1 Experimental dataset

Data selection and preprocessing

We used the structure and $\mathrm{LogP}_{\mathrm{exp}}$ value of peptides in CycPeptMPDB as the experimental data for prediction model construction. CycPeptMPDB contains permeability data based on the PAMPA, Caco-2, MDCK, and RRCK assays. Because measurements by different assays are often different, we selected PAMPA entries with the largest number of data points. The value recorded in the latest publication was used if the same peptide was measured in multiple publications. Consequently, 6,889 peptides were selected, covering a wide range of molecular weights, from 342.4 to 1777.7. Considering that the lower limit of $\mathrm{LogP}_{\mathrm{exp}}$ in CycPeptMPDB was -10 (1×10^{-10}cm/s, 240 peptides), but the detection limit in most publications was -8 (1×10^{-8}cm/s), we rounded the permeabilities of 314 peptides with values lower than -8 to -8. Similarly, the permeability of one peptide with a value higher than -4 was rounded to -4.

Division of training, validation, and test sets

The validation and test sets were extracted from the overall data for model evaluation. First, we employed the Kennard-Stone (KS) algorithm to extract 5% of all data (344 peptides) as the test set, which should uniformly cover the multidimensional space [180]. Subsequently, we generated 2048-bit Morgan FP (radius: 2) and selected the test set so that the Euclidean distance (calculated from Morgan FP) between each data point was maximized by the KS algorithm (Algorithm 4.1). From the remaining data, we randomly extracted 5% validation sets (344 peptides) three times for param-

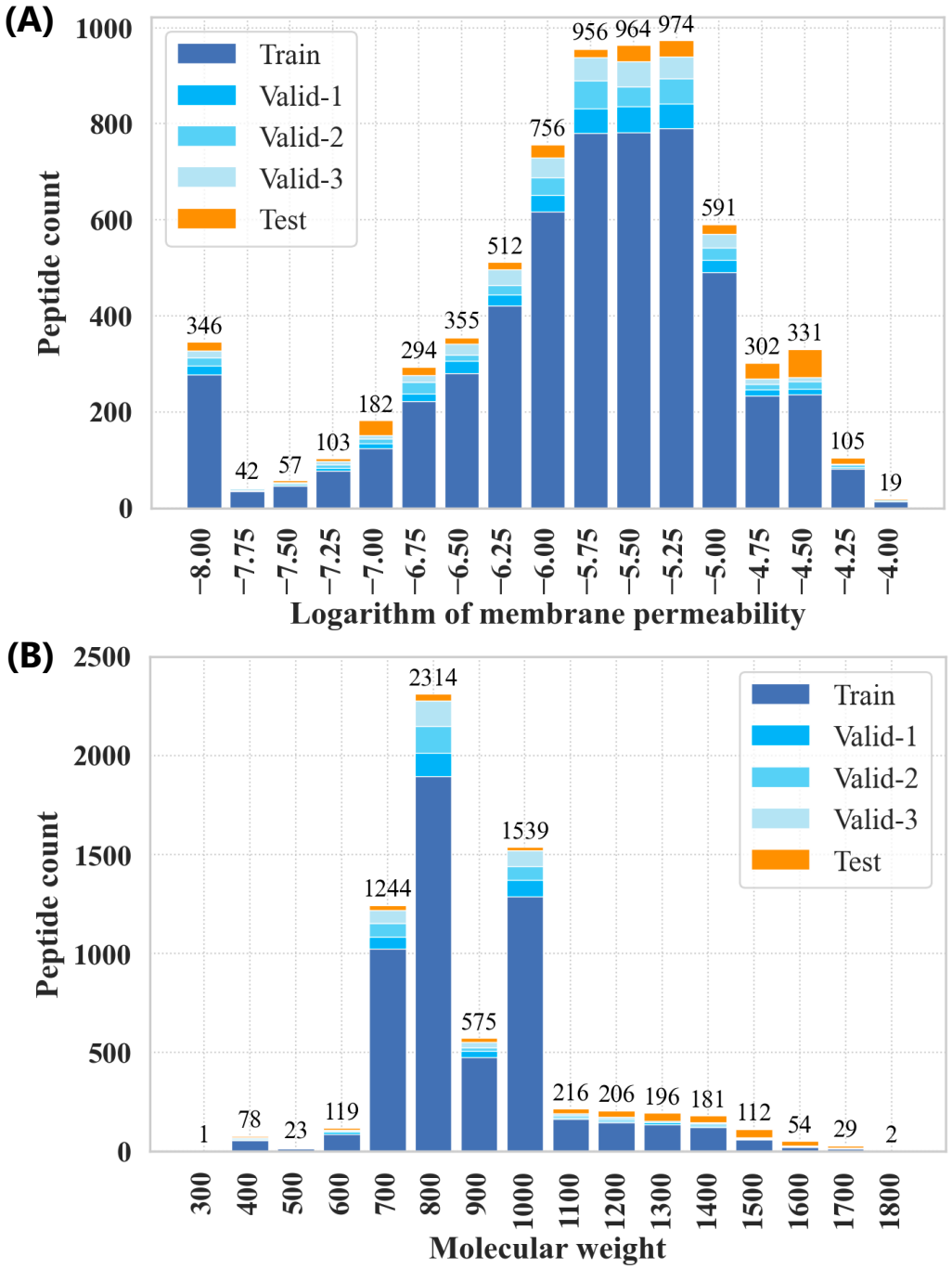


Figure 4.4: Experimental data distribution. (A) The logarithm of experimentally determined membrane permeability ($LogP_{exp}$). (B) Molecular weight (MolWt descriptor calculated by RDKit). Valid–1 is the dataset used for the first-time evaluation of the validation set; the corresponding training data sets are Train, Valid–2, and Valid–3.

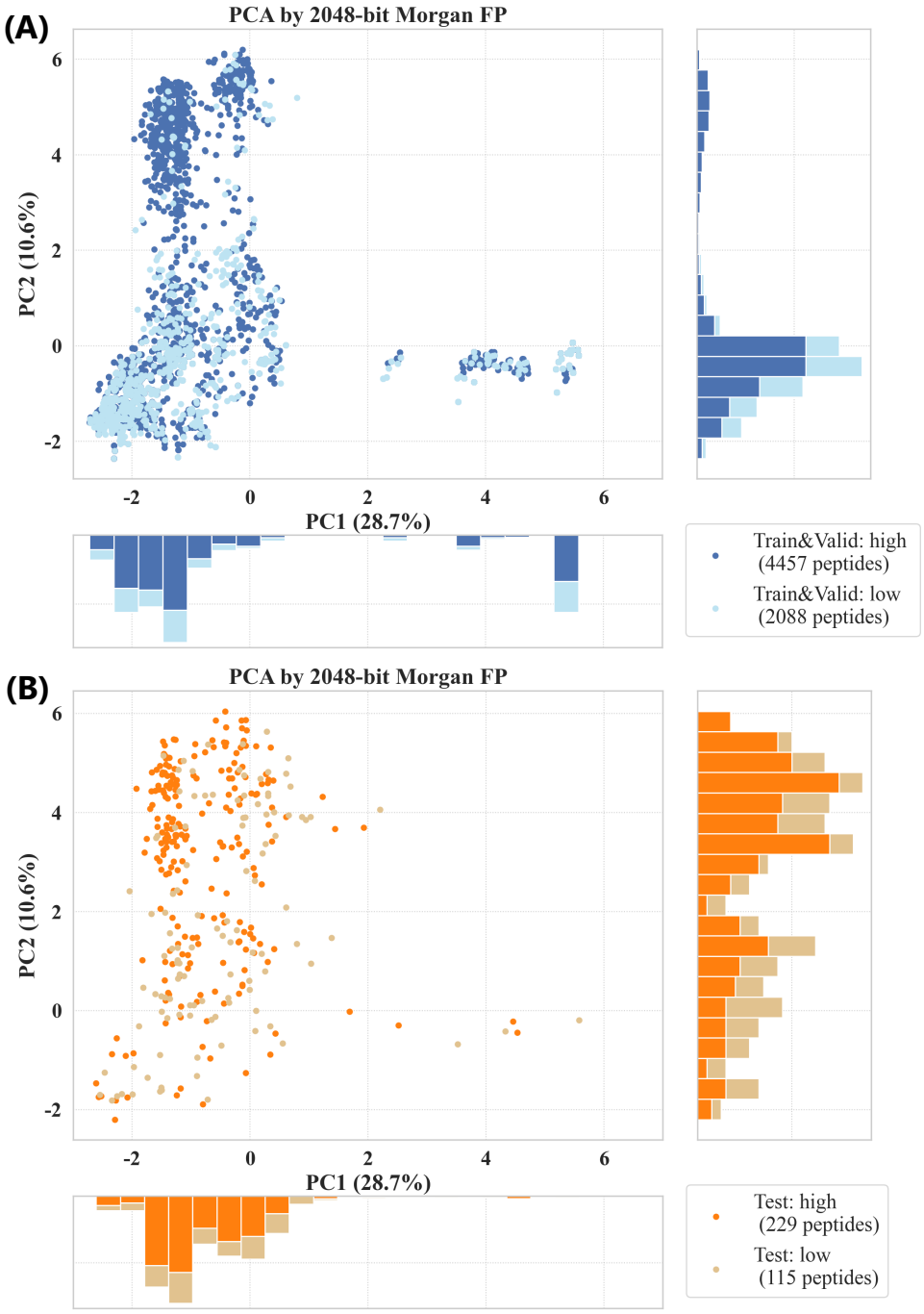


Figure 4.5: Experimental data distribution in PCA space, with the first principal component (PC1) as the horizontal axis and the second principal component (PC2) as the vertical axis; the contribution rates are shown in the parentheses of axes captions. (A) Distribution of training and validation sets. (B) Distribution of the test set. High and low indicate data with $\mathrm{LogP}_{\mathrm{exp}} \geq -6.0$ and $\mathrm{LogP}_{\mathrm{exp}} < -6.0$, respectively.

Algorithm 4.1 Details of the KS algorithm used in this study

Input:

2048-bit Morgan FP: $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_{2048})$
Total number of data: N
Number of data in the test set: n
Initial test set: $\boldsymbol{T} = \{\}$

Calculate the average $\bar{\boldsymbol{X}} = (\bar{x}_1, \bar{x}_2, ..., \bar{x}_{2048})$ of Morgan FP
for $i = 1$ to N **do**
    Calculate the Euclidean distance $D[i]$ between the sample i and the average $\bar{\boldsymbol{X}}$
end for
Select the sample i with the largest Euclidean distance from the average $\bar{\boldsymbol{X}}$ and put it into $\boldsymbol{T}$
while number of elements t of $\boldsymbol{T}$ is less than n **do**
    for $i = 1$ to $N - t$ **do**
        for $j = 1$ to t **do**
            Calculate the Euclidean distance $D[i][j]$ between the sample i that has not been selected yet and the sample j of $\boldsymbol{T}$
        end for
        Calculate the minimum value $\hat{D}[i]$ of $D[i][1]$–$D[i][t]$
    end for
    Calculate the maximum value of $\hat{D}[1]$–$\hat{D}[N - t]$ and put the sample i into $\boldsymbol{T}$
end while

eter tuning, with no overlap between the three datasets. The membrane permeability and molecular weight distributions for each set are shown in Fig. 4.4, and distribution in the principal component analysis (PCA) space based on 2048-bit Morgan FP are shown in Fig. 4.5. From the results shown in Fig. 4.4 and Fig. 4.5, structural diverse cyclic peptides spreading in the PCA space were selected as the test data by the KS algorithm. The test set had a similar ratio of peptides with high ($\mathrm{LogP_{exp}} \geq -6.0$) and low ($\mathrm{LogP_{exp}} < -6.0$) permeability ($1.99 : 1$) as the training and validation sets ($2.13 : 1$). At the same time, it contained more structurally diverse peptides distributed in the range of two to six of PC2.

4.3.2 Evaluation metrics

The average mean absolute error (MAE), mean squared error (MSE), correlation coefficient (R), and coefficient of determination ($\mathrm{R^2}$) from three repeated runs were

used as evaluation metrics:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{4.1}$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4.2}$$

$$\text{R} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{\hat{y}})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \overline{\hat{y}})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{4.3}$$

$$\text{R}^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{4.4}$$

where y_i is the experimental value of the i-th data, $\overline{y}$ is the average value of the experimental values, $\hat{y}_i$ is the predicted value of the i-th data, and $\overline{\hat{y}}$ is the average value of the predicted values.

4.3.3 Descriptors selection

We used the fusion model proposed in Chapter 3 to construct a model for predicting the membrane permeability of cyclic peptides, CycPeptMP. As we mentioned in Section 3.3.3, after the preprocessing, 407 (335 2D and 72 3D descriptors) peptide descriptors were selected. Subsequently, seven 2D and nine 3D peptide descriptors were selected based on the assigned feature importance from two RF models (one used 2D, and the other one used 3D descriptors), including LogP and polar surface area, and the same 16 monomer descriptors were selected. Fig. 4.6 shows the RF feature importance, Fig. 4.7 shown the heatmap of correlation matrix, and Table 4.4 shows the description of selected peptide descriptors.

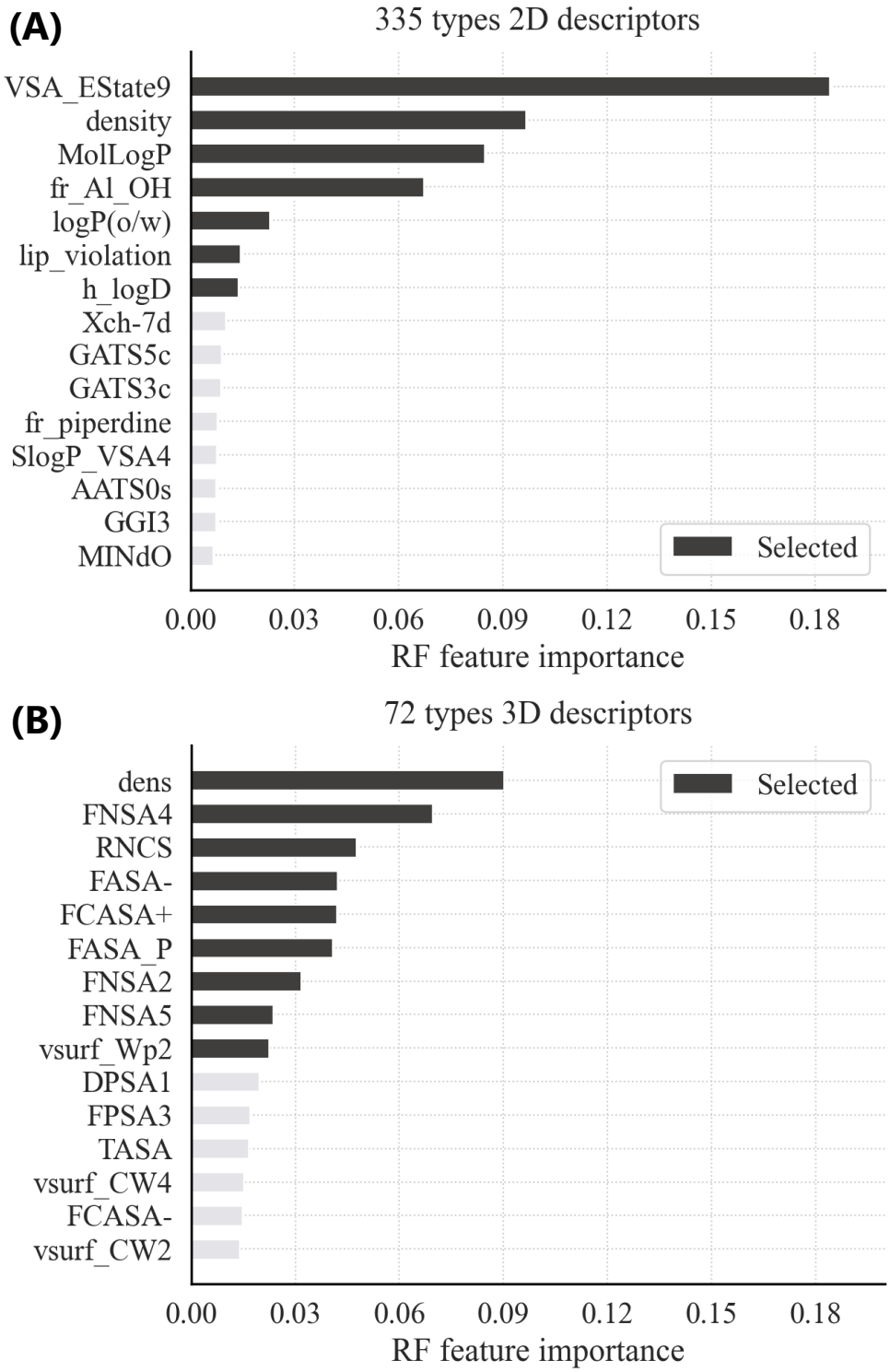


Figure 4.6: Top 15 peptide descriptors with the highest RF feature importance from (A) RF model with 2D descriptors and (B) RF model with 3D descriptors, respectively. Selected seven 2D and nine 3D peptide descriptors are shown in black.

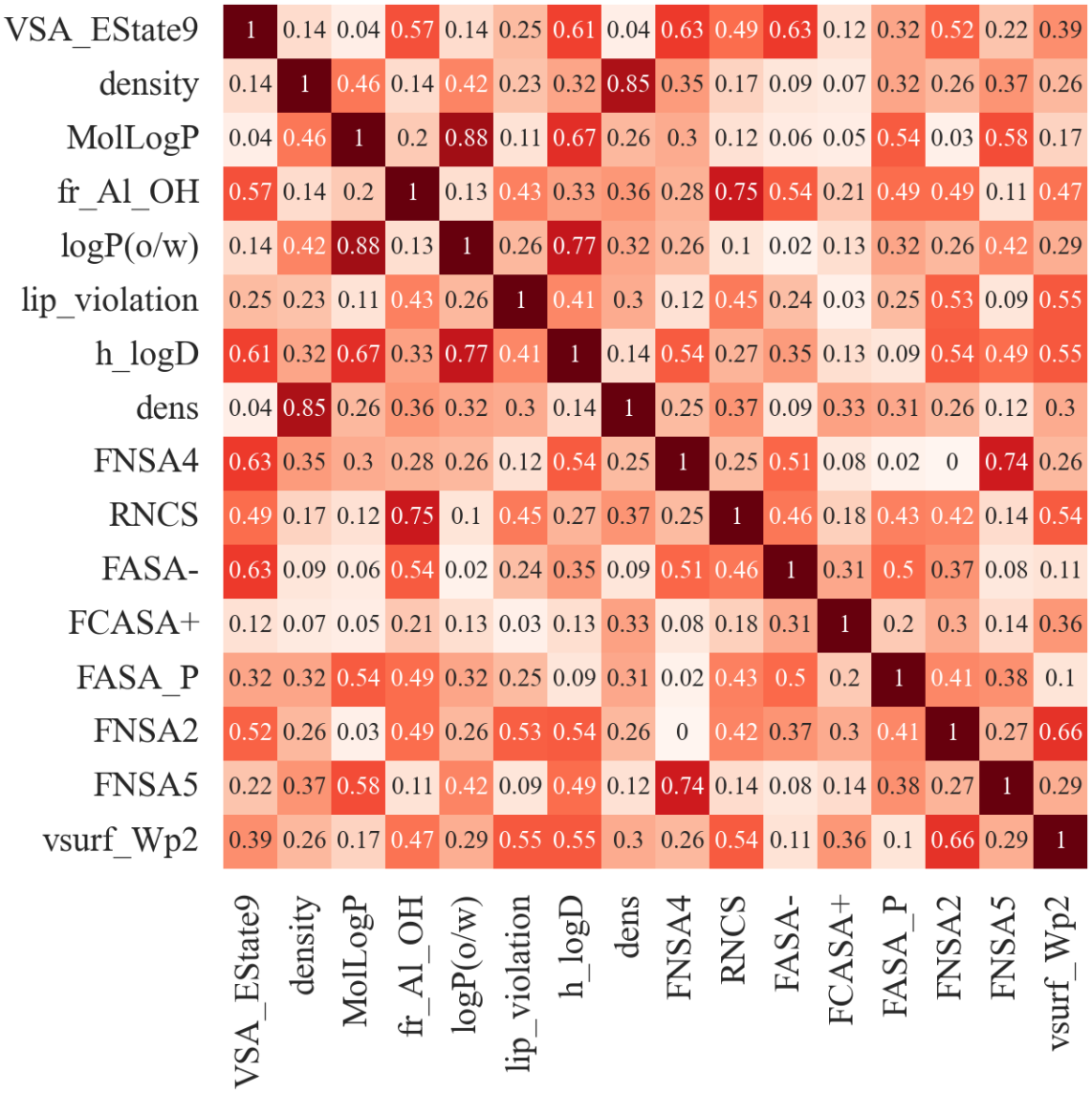


Figure 4.7: Heatmap of absolute correlation coefficient values for 16 selected peptide descriptors. The pair with the highest correlation is MolLogP and logP(o/w) ($|R| = 0.884$).

Table 4.4: Description of selected descriptors, arranged in order of RF feature importance.

Type	Name	Software	Description
2D	VSA_EState9	Mordred	Van der Waals surface area using EState indices and surface area contribution
	density	MOE	Molecular mass density (molecular weight divided by approximated van der Waals volume)
	MolLogP	RDKit	Wildman–Crippen LogP value
	fr_Al_OH	RDKit	Number of aliphatic hydroxyl groups
	logP(o/w)	MOE	Log of the octanol/water partition coefficient
	lip_violation	MOE	Number of violations of Lipinski's Rule of Five
	h_logD	MOE	Log of the octanol/water distribution coefficient at pH 7
3D	dens	MOE	Molecular mass density (molecular weight divided by 3D van der Waals volume)
	FNSA4	Mordred	Fractional charged partial negative surface area (version 4)
	RNCS	Mordred	Relative negative charge surface area
	FASA-	MOE	Fractional water accessible surface area of all atoms with negative partial charge
	FCASA+	MOE	Fractional positive charge weighted surface area
	FASA_P	MOE	Fractional water accessible surface area of all polar atoms
	FNSA2	Mordred	Fractional charged partial negative surface area (version 2)
	FNSA5	Mordred	Fractional charged partial negative surface area (version 5)
	vsurf_Wp2	MOE	VolSurf polar volume

4.3.4 Hyperparameter search

The fusion model has many hyperparameters, and hyperparameter search is a major challenge in constructing deep neural network models. Hyperparameter search methods based on Bayesian optimization have recently attracted attention because of their efficient search capabilities [181]. In this study, we performed a hyperparameter search based on the Tree-structured Parzen Estimator (TPE) algorithm [182], a type of Bayesian optimization. TPE offers the advantage of easily handling not only continuous variables, but also discrete, categorical, and conditional variables—types that are challenging to process with standard Bayesian optimization algorithms based on Gaussian processes [183]. Additionally, TPE has lower computational complexity compared to Gaussian processes. We used Optuna software (version 2.2.0) [184] as the implementation of TPE. The hyperparameters of the CycPeptMP model were determined by 150 Optuna trials; the search target and range are shown in Table 4.5; the search results are summarized in Table 4.6; hyperparameters with high Optuna importance are shown in Fig. 4.8. In this search, none of the monomer model hyperparameters were ranked highly important.

Table 4.5: Search target and range of the hyperparameter search for the proposed fusion model. Hyperparameters marked with * are not subject to search and used fixed values.

Objective	Hyperparameter	Description	Search range
Training	n_epochs*	Epoch number	50
	criterion*	Loss function of training	Mean Squared Error (MSE)
	n_earlystop*	Number of patience for early stopping	5
	scheduler*	Adjustment scheduler of learning rate	NoamLR
	warmup_epochs*	Warming up epochs of NoamLR	10 (20% of n_epochs)
	init_lr*	Initial learning rate of NoamLR	1e-4
	max_lr*	Maximal learning rate of NoamLR	1e-3
	final_lr*	Final learning rate of NoamLR	1e-5
	batch_size	Batch size	64, 128, 256
	optimizer	Type of optimizer	AdamW, NAdam, RAdam
	weight_deacy	Rate of L2 regularization	5e-6, 1e-5, 5e-5, ..., 5e-2, 1e-1
All models	d_linear	Dimension of linear layers	64, 128, 256, 512
	d_subout	Dimension of sub-model output	16, 32, 64
	ac	Activation function	ReLU, LeakyReLU, SiLU, GELU
Atom model	n_encoders	Number of encoders	1, 2, 3, 4, 5, 6
	dropout	Dropout rate	0.0, 0.05, 0.10, ..., 0.3
	n_head	Head number of multi-head attention	4, 8, 16, 32
	d_model	Dimension of encoder input	32, 64, 128, 256
	d_feedforward	Dimension of feedforward network	64, 128, 256, 512
	n_linears	Number of linear layers follows encoders	1, 2
	λ_g	Weight of concatenating *Graph* and *Conf* blocks	0.1, 0.2, ..., 0.9
Monomer model	n_conv	Number of convolutional layers	1, 2, 3, 4, 5, 6
	conv_type	Type of convolutional layers	1D-CNN, CyclicConv
	padding*	Padding size of convolutional layers	1 (1D-CNN), 0 (CyclicConv)
	d_conv	Dimension of each convolutional layer	32, 64, 128, 256
	pooling	Type of pooling layer	Max, Ave
	n_linears	Number of linear layers follows convolutional layers	1, 2
Peptide model	n_mlps	Number of linear layers	1, 2, 3, 4, 5, 6
	dropout	Dropout rate	0.0, 0.05, 0.10, ..., 0.3
	d_mlp	Dimension of linear layers	64, 128, 256, 512
Shared layer	n_shared	Number of shared linear layers	1, 2, 3

Table 4.6: Hyperparameter search range and its results for the fusion model-based membrane permeability prediction model (CycPeptMP).

Objective	Hyperparameter	Search result
Training	batch_size	256
	optimizer	AdamW
	weight_deacy	1e-1
All models	d_linear	512
	d_subout	64
	ac	LeakyReLU
Atom model	n_encoders	2
	dropout	0.2
	n_head	16
	d_model	32
	d_feedforward	512
	n_linears	1
	λ_g	0.9
Monomer model	n_conv	6
	conv_type	1D-CNN
	padding	1
	d_conv	[128, 128, 32, 256, 256, 64]
	pooling	Ave
	n_linears	1
Peptide model	n_mlps	1
	dropout	0.25
	d_mlp	256
Shared layer	n_shared	2

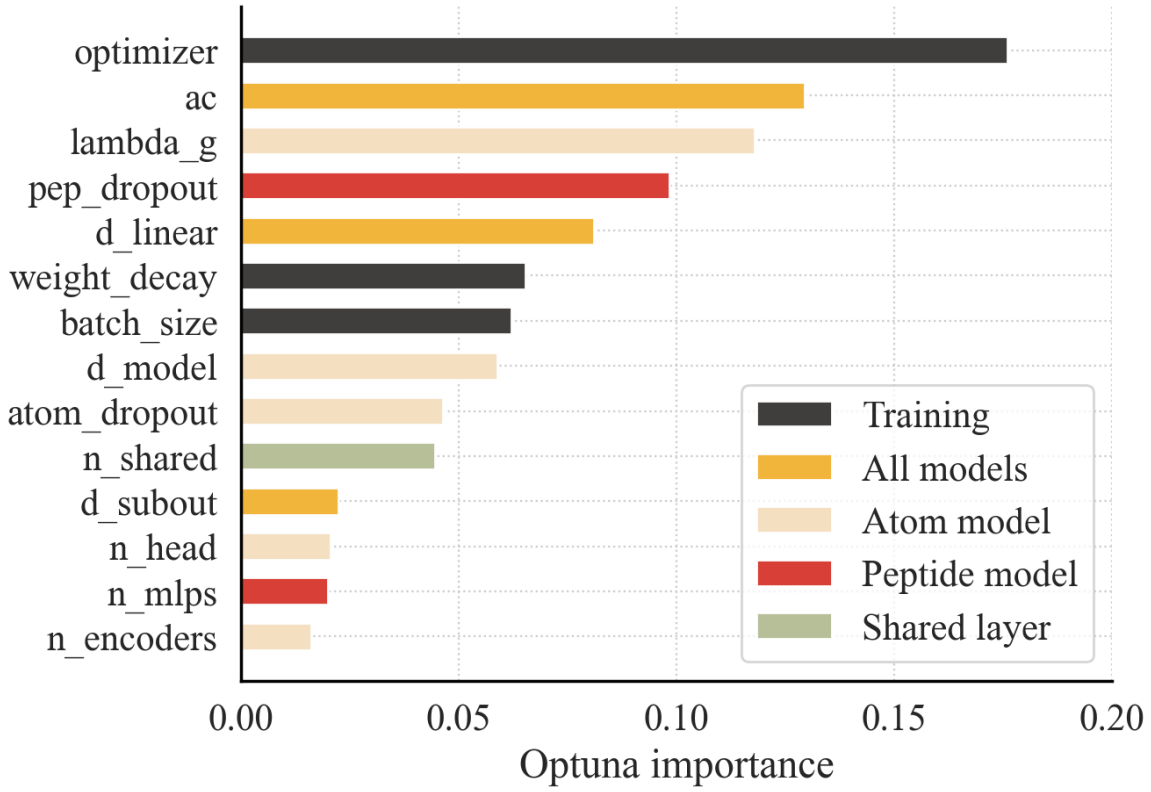


Figure 4.8: Top 14 hyperparameters with an Optuna importance > 0.01 on the CycPeptMP hyperparameter search. Optuna importance is calculated based on the fANOVA hyperparameter importance evaluation algorithm [12]; the sum of the importance values is normalized to 1.0.

4.3.5 Baseline methods

We validated the performance of the CycPeptMP model based on comparisons with seven baseline methods.

- Three traditional baselines: We constructed an RF model with 2048-bit Morgan FP, a support vector machine (SVM) model with seven 2D peptide descriptors, and an SVM model with 16 2D and 3D peptide descriptors to represent traditional cyclic peptide membrane permeability prediction methods. The hyperparameters of the RF and SVM models were determined by a grid search (Table 4.7).

- Two transformer-based methods: MAT [120] and SAT [121] were compared as state-of-the-art transformer-based methods for predicting small-molecule properties. MAT augments the transformer's self-attention mechanism with domain-specific knowledge, incorporating inter-atomic distances and molecular graph structure into the attention calculation to capture structural information. SAT focuses on the problem of traditional transformers in that positional encoding does not necessarily capture the structural similarity between nodes. It proposes a structure-aware transformer incorporating structural information into self-attention by extracting a subgraph representation rooted at each node before computing the attention.

- Two multi-level feature methods: PharmHGT [123] designs features on the atom and fragment levels and constructs a heterogeneous graph considering the correspondence between atoms and fragments for a transformer-based model. FinGAT [115] uses a GAT model to extract atom-level information and combines it with Morgan FP to capture the molecular structure from multiple perspectives.

The hyperparameters of four DL-based models were determined by 150 trials using Optuna software based on the average RMSE of three runs (Table 4.8 and Table 4.9).

Table 4.7: Search range and results of the hyperparameter search (grid search) for the RF and two SVM models.

Objective	Hyperparameter	Description	Search range	Search result
RF model	n_estimators	Number of decision trees	50, 100, 200, 300, 500, 750, 1000	750
	max_depth	Maximum depth of each decision tree	None, 2, 5, 10, 20, 30	20
SVM model	kernel	Kernel function	-	Gaussian kernel (rbf)
	C	Penalty parameter	2^{-3}, 2^{-2}, 2^{-1}, 2^{0}, 2^{1}, 2^{2}, 2^{3}, 2^{4}, 2^{5}	2^{3} (SVM-2D), 2^{2} (SVM-2D3D)
	γ	Kernel coefficient of Gaussian kernel	2^{-6}, 2^{-5}, 2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}, 2^{0}	2^{-1} (SVM-2D), 2^{-4} (SVM-2D3D)

Table 4.8: Search range and results of the hyperparameter search for the MAT and SAT models. Hyperparameters were determined by 150 trials using Optuna software based on the average MSE of three runs.

Objective	Hyper-parameter	Description	Search range (* is the value used in the original paper)	Search result
MAT	batch_size	Batch size	16, 32, 64, 128, 256*	16
	optimizer	Type of optimizer	Adam*, AdamW, NAdam, RAdam	NAdam
	weight_deacy	Rate of L2 regularization	5e-6, 1e-5, 5e-5, . . . , 5e-2, 1e-1	5e-6
	d_model	Dimension of model	32, 64, 128, 256, 512, 1024*	64
	N	Number of encoder module repeats	1, 2, 3, 4, 5, 6	5
	h	Number of molecule self-attention heads	2, 4, 8, 16*, 32	16
	N_dense	Number of dense layers in the FFN	1*, 2, 3, 4	4
	λ_att	Self-attention weight	0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1	1
	λ_dist	Distance weight	0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1	0.8
	distance_matrix_kernel	Function used to transform distance matrix	exp*, softmax	softmax
	dropout	Dropout rate	0*, 0.1, 0.2, 0.3	0
	aggregation_type	Type of global pooling	mean*, add	mean
SAT	batch_size	Batch size	16, 32, 64	16
	optimizer	Type of optimizer	Adam*, AdamW, NAdam, RAdam	NAdam
	weight_deacy	Rate of L2 regularization	5e-6, 1e-5*, 5e-5, . . . , 5e-2, 1e-1	1e-4
	abs_pe_dim	Dimension of absolute positional encoding	3, 7, 10, 20*, 30	7
	k_hop	Size of subtree	1, 2, 3*, 4, 5	5
	d_model	Dimension of model	32, 64*, 128, 256	64
	dim_feedforward	Dimension of feedforward network	64, 128*, 256, 512	64
	dropout	Dropout rate	0, 0.1, 0.2, 0.3*, 0.4	0
	num_head	Head number of multi-head attention	4, 8*, 16, 32	16
	num_layers	Number of encoders	1, 2, 3, 4, 5, 6*	5
	norm	Type of normalization	batch norm*, layer norm	layer norm
	gnn_type	Type of GNN-based subtree extractor	Graph, SAGE, GCN, GIN, GINE, PNA, PNA2*, PNA3, MPNN	Graph
	global_pool	Type of global pooling	mean*, add	mean

Table 4.9: Search range and results of the hyperparameter search for the PharmHGT and FinGAT models. Hyperparameters were determined by 150 trials using Optuna software based on the average MSE of three runs.

Objective	Hyper-parameter	Description	Search range (* is the value used in the original paper)	Search result
PharmHGT	batch_size	Batch size	32, 64*, 128, 256	32
	optimizer	Type of optimizer	Adam*, AdamW, NAdam, RAdam	AdamW
	weight_deacy	Rate of L2 regularization	5e-6, 1e-5, 5e-5, ..., 5e-2, 1e-1	1e-3
	act	Activation function	ReLU*, LeakyReLU, SiLU, GELU	GELU
	hid_dim	Dimension of model	60, 120, 180, 300*, 420, 540	420
	depth	Depth of message passing	1, 2, 3*, 4, 5, 6	2
FinGAT	batch_size	Batch size	32*, 64, 128, 256	32
	optimizer	Type of optimizer	Adam*, AdamW, NAdam, RAdam	Adam
	weight_deacy	Rate of L2 regularization	5e-6, 1e-5, 5e-5, ..., 1e-3*, ..., 5e-2, 1e-1	5e-5
	ac	Activation function	ReLU*, LeakyReLU, SiLU, GELU	GELU
	hidden_gat	Dimension of graph attention network	50*, 100, 150, 200, 300, 500	500
	in_head	Head number of multi-head attention	3, 4, 5*, 6, 7, 8	8
	global_pool	Type of global pooling	mean*, max, add	max
	hidden_linear_1	Dimension of the 1st linear layer	10, 25, 50, 100*, 150, 200	100
	hidden_linear_2	Dimension of the 2nd linear layer	10, 25*, 50, 100, 150, 200	200
	hidden_linear_3	Dimension of the 3rd linear layer	10*, 25, 50, 100, 150, 200	50

Table 4.10: Performance comparison between seven baseline methods and CycPeptMP using the test set. The metrics are the averaged values of three repeated runs; the best result for each metric is indicated in bold.

Metrics	RF	SVM-2D	SVM-2D3D	MAT
MAE	0.485 ± 0.003	0.488 ± 0.005	0.418 ± 0.001	0.503 ± 0.025
MSE	0.380 ± 0.004	0.449 ± 0.014	0.345 ± 0.002	0.461 ± 0.033
R	0.815 ± 0.003	0.781 ± 0.007	0.834 ± 0.001	0.768 ± 0.019
R^2	0.657 ± 0.003	0.595 ± 0.012	0.689 ± 0.002	0.584 ± 0.030
Metrics	SAT	PharmHGT	FinGAT	CycPeptMP
MAE	0.591 ± 0.053	0.443 ± 0.028	0.505 ± 0.017	$\mathbf{0.355 \pm 0.007}$
MSE	0.641 ± 0.081	0.375 ± 0.046	0.447 ± 0.035	$\mathbf{0.253 \pm 0.013}$
R	0.660 ± 0.053	0.825 ± 0.019	0.785 ± 0.010	$\mathbf{0.883 \pm 0.003}$
R^2	0.422 ± 0.073	0.662 ± 0.041	0.597 ± 0.031	$\mathbf{0.772 \pm 0.011}$

4.4 Results and Discussion

4.4.1 Performance comparison for the test set

The prediction accuracy results for the test set are shown in Table 4.10. CycPeptMP ranked first in all evaluation metrics, reflecting a significant improvement in prediction performance over all existing methods (MAE = 0.355, R = 0.883, Fig. 4.9). Considering the structural diversity of the test set, CycPeptMP showed good generalization performance and could learn the complex structures of cyclic peptides which is difficult to apply pre-training through augmentation. The RF model constructed based on Morgan FP showed good prediction performance and ranked fourth among all methods (MAE = 0.485, R = 0.815, Fig. 4.10 (A)). SVM with 2D peptide descriptors (MAE = 0.488, R = 0.781, Fig. 4.10 (B)) had lower prediction accuracy than the RF model, whereas SVM with 3D descriptors improved prediction accuracy, making it superior to the RF model for the test set and ranked second among all methods (MAE = 0.418, R = 0.834, Fig. 4.10 (C)). Cyclic peptide membrane permeation by passive diffusion negatively correlated with molecule size. SVM could partially predict permeability by using lipophilicity descriptors, such as LogP, which are largely dependent on molecular weight. CycPeptMP effectively combined Morgan FP and 16 2D and 3D peptide descriptors as peptide-level information to comprehensively characterize peptide structures from a topological and physicochemical perspective, leading to an improvement in prediction capabilities. The graph represen-

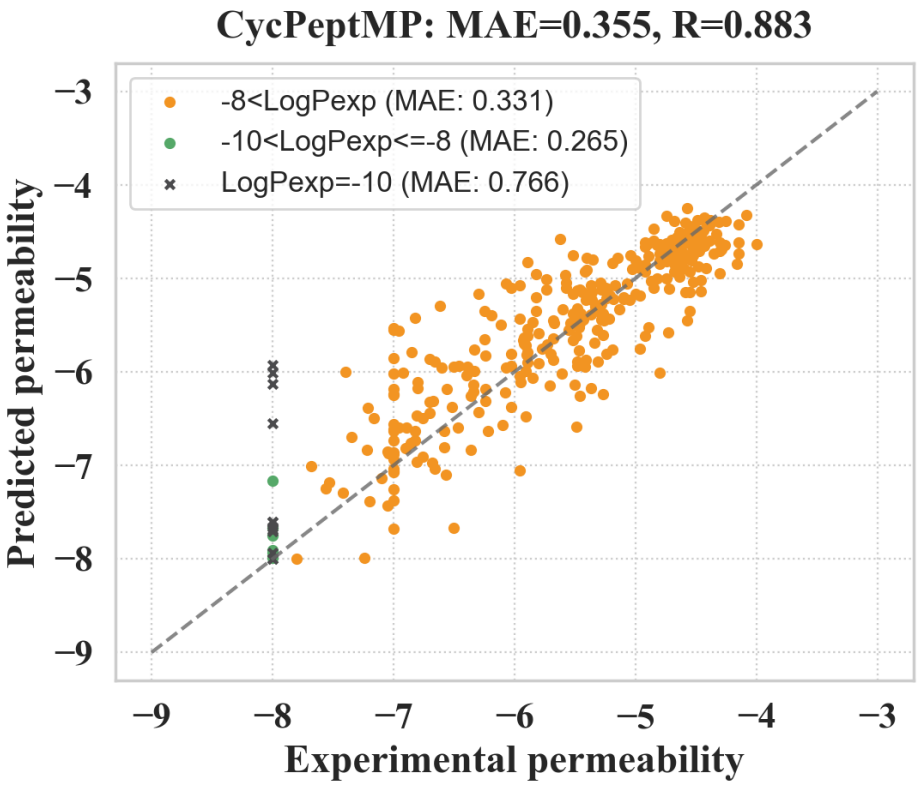


Figure 4.9: Prediction results of the test set by CycPeptMP. The predicted values of the test set are the average value of three runs.

tation transformer-based MAT performed close to 2D SVM (MAE = 0.503, R = 0.768, Fig. 4.11 (A)), but SAT could not predict membrane permeability (MAE = 0.591, R = 0.660, Fig. 4.11 (B)). Although MAT and SAT are state-of-the-art methods for predicting small-molecule properties, they could not effectively learn the structures of more complex cyclic peptides since they only utilize the atom-level information without augmentation technique. In addition to atom-level information, PharmHGT with fragment MACCS Keys (MAE = 0.443, R = 0.825, Fig. 4.12 (A)) and FinGAT with molecular FP (MAE = 0.505, R = 0.785, Fig. 4.12 (B)) had significantly improved prediction accuracies compared to MAT and SAT. PharmHGT ranked third among all methods, and FinGAT showed the same level of accuracy as the RF and 2D SVM models. Hence, designing features from various perspectives may be key to successfully predicting the membrane permeability of cyclic peptides. These findings indicated that CycPeptMP effectively employed three levels of features to capture a wide range of structural information from the smallest atomic detail to the broader peptide-level conformation.

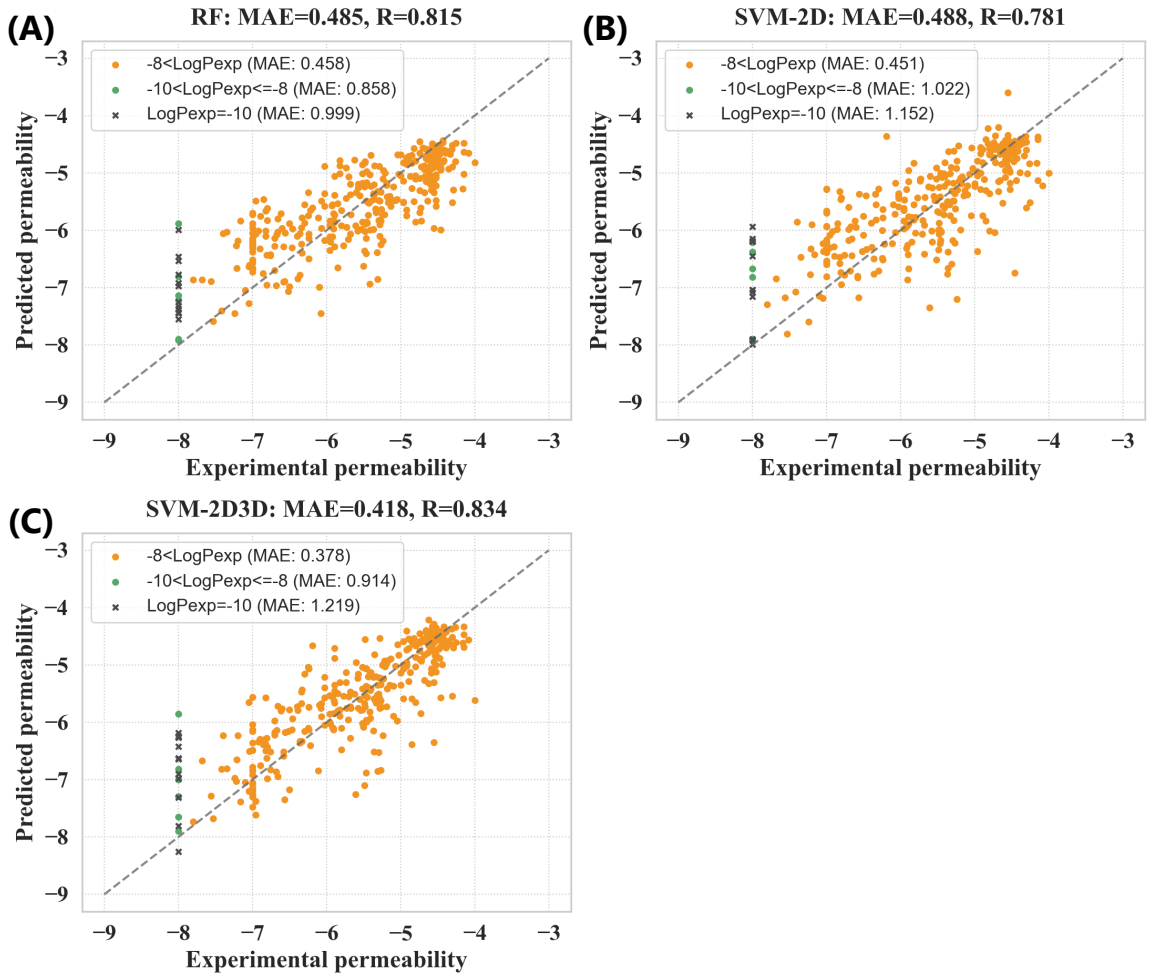


Figure 4.10: Prediction results of the test set by (A) RF, (B) SVM-2D, and (C) SVM-2D3D models. The predicted values of the test set are the average value of three runs.

Lower limit processing of permeability

We rounded the permeability with $-10 \leq \mathrm{LogP}_{\mathrm{exp}} < -8$ to -8 because the detection limit for most literature is -8. Among them, most peptides were recorded in CycPeptMPDB with $\mathrm{LogP}_{\mathrm{exp}} = -10$. Most were not measured as -10 and were set to -10 by CycPeptMPDB as there was no clear value. Therefore, their membrane permeability was unreliable. To discuss the effect of these data, we calculated the accuracy of 12 peptides with $\mathrm{LogP}_{\mathrm{exp}} = -10$, 6 peptides with $-10 < \mathrm{LogP}_{\mathrm{exp}} \leq -8$, and 326 other peptides with $-8 < \mathrm{LogP}_{\mathrm{exp}}$ of the test set, respectively. Peptides with $\mathrm{LogP}_{\mathrm{exp}} = -10$ could not be predicted by any method (MAE = 0.766 to 1.219, Fig. 4.9 to Fig. 4.12). We have included these unreliable experimental values in our data to incorporate as

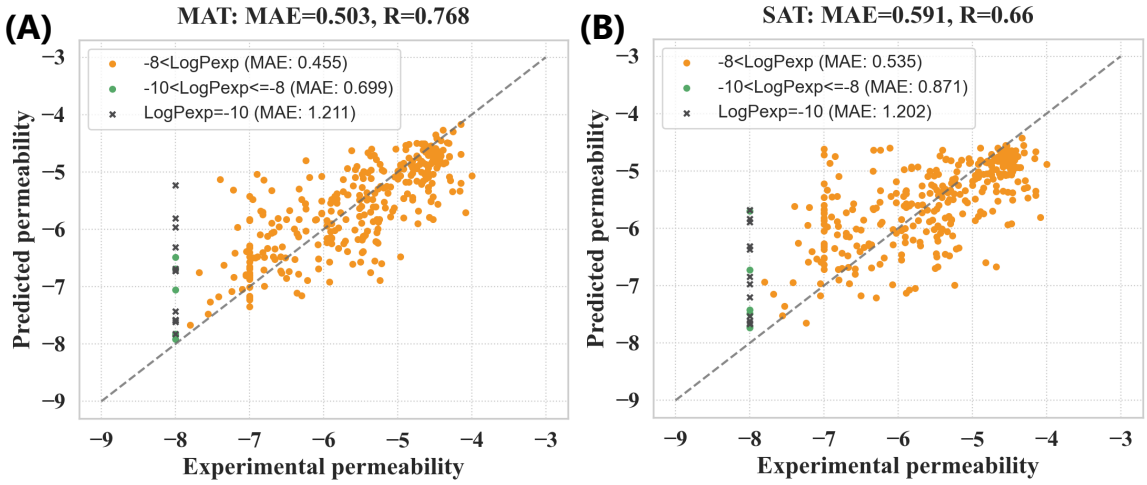


Figure 4.11: Prediction results of the test set by (A) MAT and (B) SAT models. The predicted values of the test set are the average value of three runs.

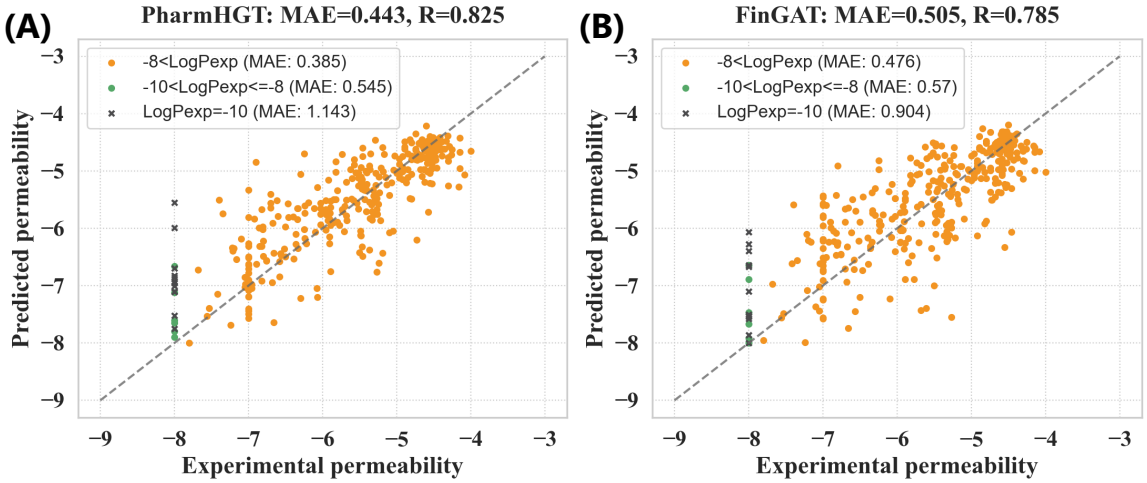


Figure 4.12: Prediction results of the test set by (A) PharmHGT and (B) FinGAT models. The predicted values of the test set are the average value of three runs.

much data as possible; however, it may be more appropriate to eliminate them. For the peptides that have clear measurement values with $-10 < \mathrm{LogP}_{\mathrm{exp}} \leq -8$, no baseline methods could predict them (MAE = 0.545 to 1.022). Nevertheless, CycPeptMP accurately predicted them (MAE = 0.265), further demonstrating the superiority of the proposed method over baselines.

Table 4.11: Prediction performance between seven baseline methods and CycPeptMP (models were trained with PAMPA) for Caco-2, MDCK, and RRCK permeabilities recorded in CycPeptMPDB. The metrics are averaged for three runs.

Assay	Metrics	RF	SVM-2D	SVM-2D3D	MAT
Caco-2	MAE	1.124 ± 0.006	0.784 ± 0.007	0.766 ± 0.007	1.073 ± 0.128
	R	0.181 ± 0.002	0.279 ± 0.016	0.290 ± 0.004	0.277 ± 0.012
MDCK	MAE	0.913 ± 0.016	0.706 ± 0.022	0.778 ± 0.005	0.918 ± 0.028
	R	0.283 ± 0.021	0.377 ± 0.072	0.618 ± 0.011	0.529 ± 0.020
RRCK	MAE	0.683 ± 0.026	0.662 ± 0.007	0.598 ± 0.005	0.558 ± 0.020
	R	-0.044 ± 0.045	0.245 ± 0.006	0.291 ± 0.016	0.404 ± 0.028
Assay	Metrics	SAT	PharmHGT	FinGAT	CycPeptMP
Caco-2	MAE	0.813 ± 0.027	0.808 ± 0.122	1.191 ± 0.033	1.148 ± 0.113
	R	0.180 ± 0.066	0.199 ± 0.010	0.132 ± 0.022	0.209 ± 0.064
MDCK	MAE	0.859 ± 0.092	0.977 ± 0.047	0.931 ± 0.026	0.821 ± 0.009
	R	0.543 ± 0.128	0.631 ± 0.089	0.254 ± 0.009	0.570 ± 0.044
RRCK	MAE	0.557 ± 0.119	0.480 ± 0.020	0.557 ± 0.036	0.678 ± 0.041
	R	0.322 ± 0.229	0.269 ± 0.019	0.107 ± 0.066	-0.181 ± 0.027

Measurement experiment error

Meanwhile, different experimental conditions can significantly alter the measurements. CycPeptMPDB records all reported values from different literature assays for the same peptide (this study used values from the most recent literature). For example, cyclosporin A is a peptide with PAMPA measurements reported from five literature sources with permeabilities of -5.01, -6.20, -6.15, -5.71, and -5.72 (max: -5.01, min: -6.20, std: 0.427) in chronological order of publication; 1NMe3 is a peptide with PAMPA measurements reported from six literature sources with permeabilities of -4.50, -4.40, -6.00, -6.24, -6.40, and -5.52 (max: -4.40, min: -6.40, std: 0.798) in chronological order of publication. Since these errors are already present in the measurement experiment, the prediction accuracy MAE = 0.355 of CycPeptMP may be close to the limit of prediction.

Table 4.12: Performance comparison between seven baseline methods and CycPeptMP by 10-fold cross-validation. The metrics are the averaged values of ten repeated runs; the best result for each metric is indicated in bold.

Metrics	RF	SVM-2D	SVM-2D3D	MAT
MAE	0.400 ± 0.010	0.396 ± 0.010	0.386 ± 0.008	0.397 ± 0.022
MSE	0.318 ± 0.021	0.349 ± 0.022	0.334 ± 0.020	0.322 ± 0.018
R	0.748 ± 0.018	0.725 ± 0.017	0.739 ± 0.015	0.750 ± 0.022
R^2	0.557 ± 0.028	0.516 ± 0.027	0.537 ± 0.025	0.553 ± 0.032
Metrics	SAT	PharmHGT	FinGAT	CycPeptMP
MAE	0.403 ± 0.020	0.398 ± 0.023	0.400 ± 0.019	$\mathbf{0.352 \pm 0.015}$
MSE	0.325 ± 0.022	0.316 ± 0.031	0.321 ± 0.023	$\mathbf{0.271 \pm 0.023}$
R	0.748 ± 0.019	0.758 ± 0.023	0.755 ± 0.022	$\mathbf{0.786 \pm 0.019}$
R^2	0.549 ± 0.031	0.562 ± 0.043	0.554 ± 0.029	$\mathbf{0.613 \pm 0.033}$

4.4.2 Application of trained PAMPA model to other assay data

The predicting results of Caco-2 (378 peptides), MDCK (17 peptides), and RRCK (53 peptides) permeabilities recorded in CycPeptMPDB (the duplicate peptides between each assay and PAMPA were deleted) using models trained by PAMPA data are shown in Table 4.11. No model could predict these assays (Caco-2: MAE = 0.766–1.191, MDCK: MAE = 0.706–0.977, RRCK: MAE = 0.480–0.683). Such direct predictive applications proved difficult because the membrane permeability values obtained from different assays often differed. For example, Wang *et al.* [66] reported both PAMPA and Caco-2 measurements of 62 cyclic peptides, but the correlation coefficient between PAMPA and Caco-2 permeability values was only 0.71.

4.4.3 Performance comparison for 10-fold cross-validation

Considering generalization performance, the Kennard-Stone algorithm was employed to maximize the distance between data points in the chemical space of the test set to extract the most diverse test set possible from the CycPeptMPDB data. For multiple random sampling evaluation, we performed a new 10-fold cross-validation with different random seeds for each run without altering the determined hyperparameters. As shown in Table 4.12, CycPeptMP consistently demonstrated the highest prediction performance for the difficult-to-predict test set (MAE = 0.355) and 10-fold

cross-validation (MAE = 0.352). All baseline methods had higher accuracy for the 10-fold cross-validation than the test set and showed similar performance (MAE = 0.386–0.403).

4.4.4 Comparison to DL-based methods based on CycPeptMPDB

Since we published CycPeptMPDB in 2023 [82], it has been widely used by research groups worldwide. It has been the subject of active research on membrane permeability prediction of cyclic peptides. At present (November 2024), six DL-based methods for predicting membrane permeability of cyclic peptides have been reported using the CycPeptMPDB data. A brief description of these methods in chronological order is summarized below:

- PepLand [185] performed pre-training using a training set composed of three sub-datasets: a protein dataset containing 7,924,509 sequences with a length of less than 30 from UniProt, all 7,334 cyclic peptides from CycPeptMPDB, and 1,643 peptides from PDB. The model architecture is based on PharmHGT modified for peptides (heterogeneous graph transformer). It is important to note that PepLand used CycPeptMPDB data in both the pre-training phase and the performance evaluation phase.

- Multi_CycGT [186] incorporated both atom- and peptide-level features into their prediction model, but the model architecture used the simple GCN and transformer and was not modified for cyclic peptides. It is important to note that their training and test datasets contain leaks.

- CyclePermea [187] performed contrastive learning based on the canonical SMILES and non-canonical SMILES representations of the same cyclic peptide. During prediction, in addition to atom-level features, the model also incorporates peptide-level fingerprints to enhance the feature representation.

- Tan *et al.* [188] collected 823 permeability data of linear peptides in addition to the data from CycPeptMPDB. They removed data with low reliability of $\mathrm{LogP}_{\mathrm{exp}}$ less than -8 and greater than -4. In addition to the atom-level features, they used a total of 2,066 descriptors calculated from four software and three types of fingerprints as peptide-level features.

- MuCoCP [189] performed contrastive learning using the HELM sequence representation and SMILES representation from CycPeptMPDB.
- PeptideCLM [190] performed pre-training based on four different datasets with diverse molecular structures: 10 million small molecules from the PubChem database, 2.2 million small molecules from the SureChEMBL database, 825,632 peptides from the SmProt database (less than 100 amino acids), and 10 million modified peptides generated using an updated version of CycloPs software, which included 100 unnatural amino acids from SwissSidechain. They removed data with low reliability of $\mathrm{LogP}_{\mathrm{exp}} = -10$ in CycPeptMPDB.

To compare these DL-based methods developed for cyclic peptide membrane permeability prediction, the accuracy metrics reported by each method and the results for CycPeptMP (test set) are summarized in Table 4.13. It is important to note that these values were not derived from the same training and test data, so direct comparisons may not be entirely fair. As shown in Table 4.13, although it is not a completely fair comparison, CycPeptMP showed prediction accuracy equal to or better than these state-of-the-art DL-based methods. While the MAE and MSE for CycPeptMP were slightly higher than the model by Tan *et al.* [188], CycPeptMP achieved the highest R and R^2 value. There are two primary reasons for the relatively strong prediction performance of the model by Tan *et al.*: First, their model uses a wide variety of descriptors and fingerprints, which allows it to capture a more comprehensive range of features. Second, they removed data with $\mathrm{LogP}_{\mathrm{exp}}$ less than -8, especially data with $\mathrm{LogP}_{\mathrm{exp}} = -10$, which could have improved the overall model accuracy.

Table 4.13: Performance comparison between six DL-based methods and CycPeptMP. The best result for each metric is indicated in bold.

Method	Architecture	CycPeptMPDB Data	MAE	MSE	R	R^2
PepLand [185]	(Pre-training) Transformer	All	-	-	0.520	-
Multi_CycGT [186]	MLP, GNN, Transformer	PAMPA	0.394	0.269	-	0.338
CyclePermea [187]	(Contrastive learning) Transformer	All	0.486	0.411	0.518	-
Tan *et al.* [188]	MLP, GNN	PAMPA ($-8 \leq \text{LogP}_{\text{exp}} \leq -4$)	**0.317**	**0.185**	-	0.672
MuCoCP [189]	(Contrastive learning) GNN, Transformer	All	-	0.714	-	0.503
PeptideCLM [190]	(Pre-training) Transformer	PAMPA ($\text{LogP}_{\text{exp}} > -10$)	-	-	-	0.658
CycPeptMP	MLP, CNN, Transformer	PAMPA	0.355	0.253	**0.883**	**0.772**

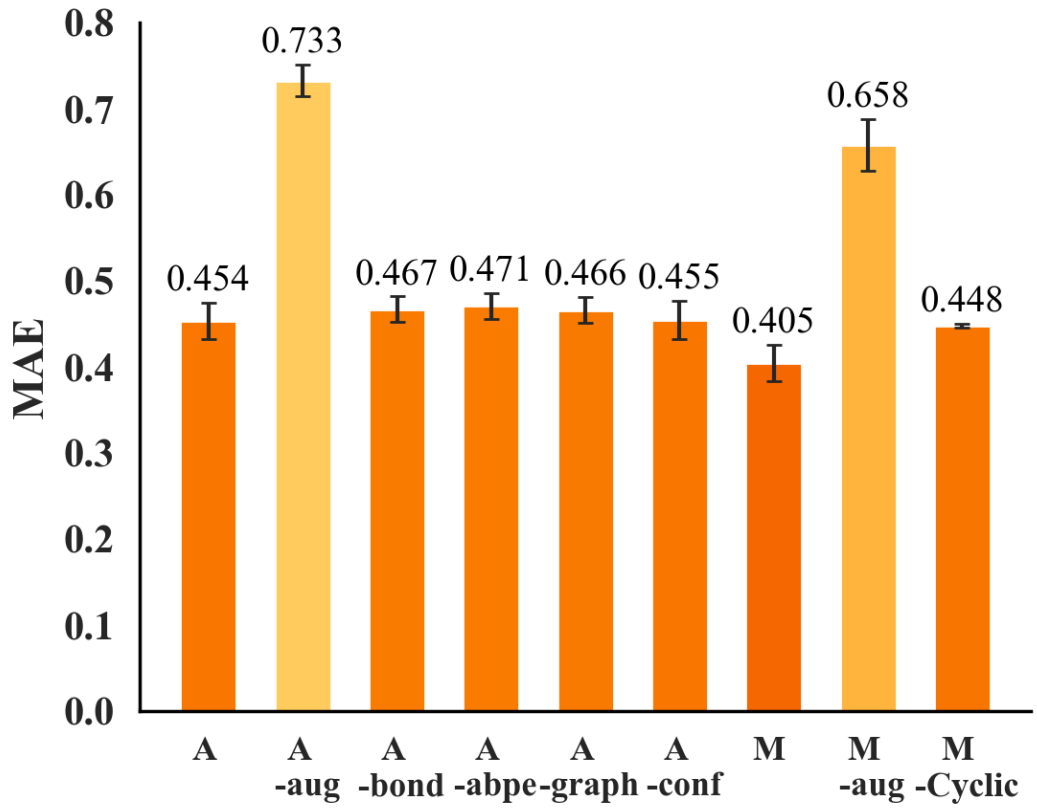


Figure 4.13: Ablation results (MAE) for the atom and monomer models using the test set.

4.4.5 Ablation study of atom and monomer models

We conducted ablation studies on atom and monomer models with complex architectures without altering the determined hyperparameters (Fig. 4.13). For the atom model, A is the original model (Fig. 3.6), and A–aug is the result without data augmentation. We measured the prediction accuracy when not using the *Bond* matrix (A–bond), using ordinary absolute positional encoding [81] instead of the *Bond* matrix (A–abpe), retaining only the *Conf* block (A–graph), or retaining only the graph block (A–conf). As shown in Fig. 4.13, the prediction accuracy of the atom model significantly improved by augmentation (A: 0.454, A–aug: 0.733). Regarding the architectural changes of the atom model, the original A showed the highest prediction accuracy, while the deletion of any element decreased the prediction accuracy. The relationship between atoms was captured more effectively using *Bond* (A: 0.454) than absolute positional encoding (A–abpe: 0.471), and the impact of removing *Graph* block (A–graph: 0.466) was greater than that of removing *Conf* block (A–conf: 0.455).

For the monomer model, M is the original model (Fig. 3.5) and M–aug is the result without data augmentation. We also measured the accuracy change when replacing the general 1D-CNN layers with CyclicConv layers (M–Cyclic, Fig. 3.4). Similar to the atom model results, the prediction accuracy of the monomer model significantly improved by augmentation (M: 0.405, M–aug: 0.658). These results showed that SMILES enumeration for the atom model and sequence arrangement for the monomer model effectively improved learning efficiency. Moreover, the augmentation technique is essential for learning the complex structure of cyclic peptides. Additionally, the 1D-CNN layer (M: 0.405) was superior to the CyclicConv layer (M–Cyclic: 0.448). This result indicates that even using the 1D-CNN layer, the circularity of cyclic peptides may be expressed to some extent through data augmentation.

4.4.6 Ablation study of the fusion model

The ablation study for the fusion model measured the influences of the number of replicas generated by augmentation and changes in architecture. Fig. 4.14 (A) shows the accuracy based on 1 (no augmentation), 5, 10, 20, 30, 40, 50, and 60 (CycPeptMP) replicas per peptide. We observed a significant improvement in prediction accuracy compared to that without augmentation (F–1: 0.456) even with five replicas (F–5: 0.394). However, over 20 replicas showed approximately the same prediction accuracy as the amount of training data increased. This may be due to the limitations of increased diversity caused by merely reordering the inputs and the lack of diversity in the generated conformations

In Fig. 4.14 (B), F is the original CycPeptMP model; F–aux is the model without auxiliary loss; F–atom, F–mono, and F–pep represent the models lacking the respective sub-models; F–3D is the model that did not use all 3D information ($Conf$ and 3D descriptors); and F–ensem represents the model with each sub-model allowed to directly predict membrane permeability and the average ensemble of three predictions was taken. The use of auxiliary loss insignificantly improved prediction accuracy (F–aux: 0.366). Furthermore, prediction accuracy decreased when any of the three sub-models were removed, indicating that the three levels of information were important to predicting membrane permeability. The peptide model had the greatest influence (F–pep: 0.388), followed by the monomer (F–mono: 0.387) and atom model (F–atom: 0.368). The use of 3D information improved prediction accuracy (F–3D: 0.38). Correctly addressing the possible conformational distribution of peptides appears important, and accuracy may be improved by generating conformations using a more rigorous method,

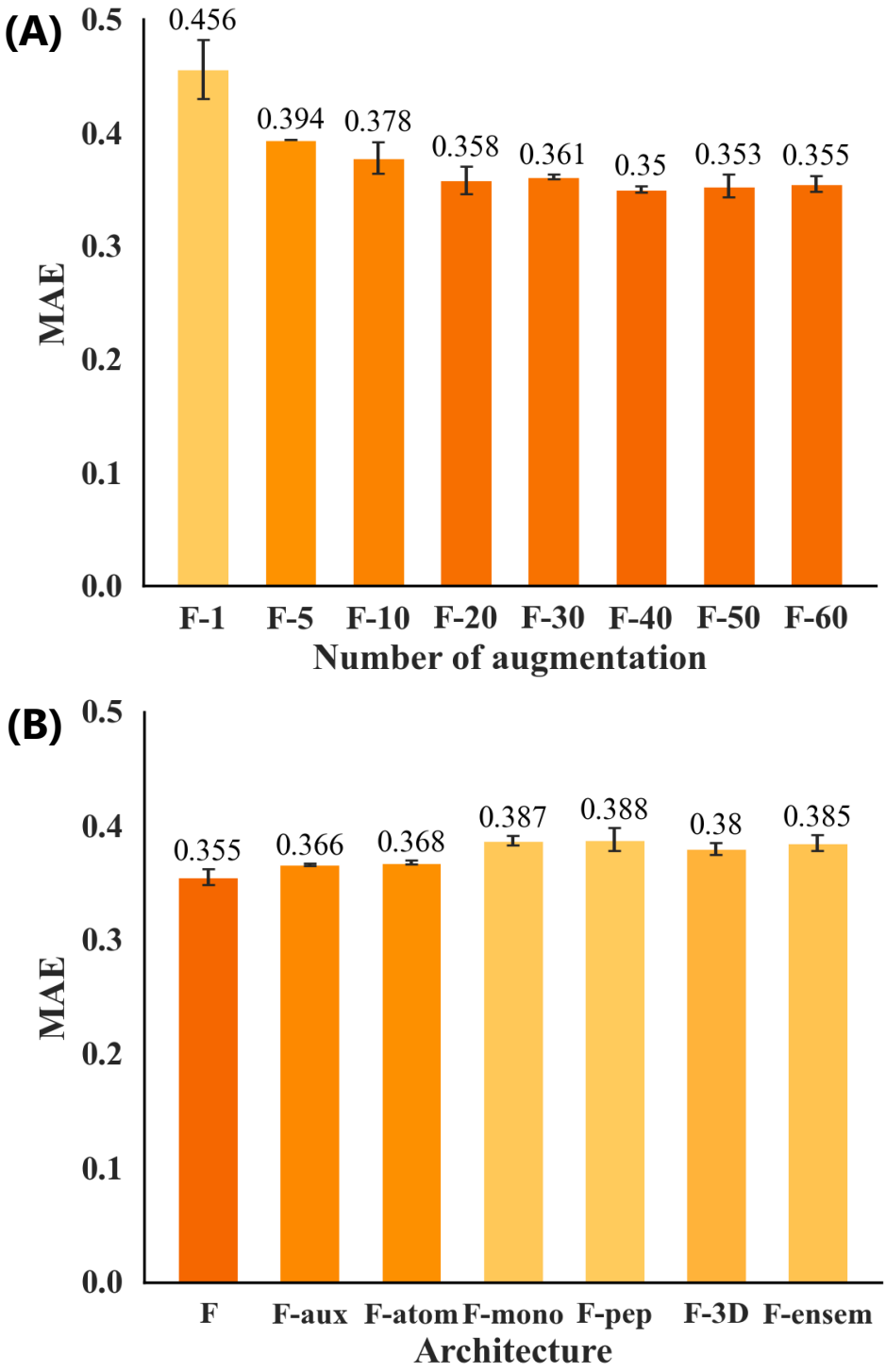


Figure 4.14: Ablation results (MAE) for the fusion model using the test set. (A) Different numbers of input replicas. (B) Different architectures.

Table 4.14: Prediction performance of CycPeptMP of the test set using 3D conformations regenerated by RDKit (five times with different seeds). The metrics are averages of three runs.

Augmentation	Trial	MAE	MSE	R	R^2
1 (Non Aug.)	t0 (original)	0.456 ± 0.026	0.373 ± 0.036	0.823 ± 0.018	0.664 ± 0.033
	t1	0.448 ± 0.026	0.361 ± 0.034	0.827 ± 0.017	0.674 ± 0.031
	t2	0.452 ± 0.026	0.365 ± 0.034	0.825 ± 0.017	0.671 ± 0.031
	t3	0.449 ± 0.026	0.359 ± 0.033	0.828 ± 0.016	0.676 ± 0.030
	t4	0.450 ± 0.025	0.364 ± 0.034	0.825 ± 0.017	0.672 ± 0.030
	t5	0.449 ± 0.025	0.361 ± 0.033	0.827 ± 0.017	0.674 ± 0.030
60 (CycPeptMP)	t0 (original)	0.355 ± 0.007	0.253 ± 0.013	0.883 ± 0.003	0.772 ± 0.011
	t1	0.355 ± 0.006	0.253 ± 0.012	0.882 ± 0.003	0.772 ± 0.011
	t2	0.355 ± 0.006	0.253 ± 0.012	0.882 ± 0.003	0.771 ± 0.011
	t3	0.355 ± 0.006	0.253 ± 0.012	0.882 ± 0.003	0.772 ± 0.011
	t4	0.355 ± 0.006	0.253 ± 0.012	0.882 ± 0.003	0.772 ± 0.011
	t5	0.355 ± 0.006	0.253 ± 0.012	0.882 ± 0.003	0.772 ± 0.011

such as MD simulations. Finally, the average prediction accuracy further decreased when using a sub-model ensemble (F–ensem: 0.385). Hence, it was better to extract latent features than having each sub-model directly predict permeability.

4.4.7 Performance using regenerated conformations

As shown by the results of the ablation study of the fusion model, utilizing information derived from the 3D conformation improved the prediction accuracy. Therefore, we measured the effect of the 3D conformations on the prediction accuracy. We regenerated the 3D conformations of the test set five times using the RDKit (with different seeds) to discuss how the prediction accuracy of CycPeptMP changed. As shown in Table 4.14, the prediction accuracy had a minor change (MAE = 0.448–0.456) using only one conformation per peptide (without augmentation), and the prediction accuracy did not change when using 60 times augmentation (MAE = 0.355).

4.4.8 Comparison with MD-based method

Cyclic peptides tend to exist in various conformations, resulting in slow conformational transitions relative to simulation time scales. Our group proposed the first MD-based large-scale prediction of cyclic peptide membrane permeability used steered MD and replica-exchange umbrella sampling to accelerate sampling and simulated the membrane permeation process of 100 six-residue and 56 eight-residue peptides through

Table 4.15: Peptides used in comparison with the MD-based method. The 23 peptides are included in the validation and test sets of this study. The AlogP and MD predicted values ($\log P_{ISMD\_mod}$) are reported by Sugita *et al* [17].

Source	MPDB ID	Compound Name	AlogP	Belongs to	Exp. Value	MD Pred. Value	CycPeptMP Pred. Value
2013_CHUGAI	536	DP-528	0.73	Test	−4.89	−4.91	−4.76
2013_CHUGAI	538	DP-530	1.20	Test	−4.96	−5.50	−4.87
2013_CHUGAI	677	DP-712	0.78	Test	−4.92	−5.08	−4.93
2016_Furukawa	1134	1.1-01	0.12	Valid-1	−7.13	−5.06	−6.98
2016_Furukawa	1146	1.1-13	3.06	Valid-2	−4.99	−4.71	−5.10
2016_Furukawa	1151	1.1-18	2.23	Valid-1	−6.14	−4.19	−6.04
2016_Furukawa	1200	1.2-27	3.11	Valid-2	−5.34	−4.34	−5.36
2016_Furukawa	1231	1.3-18	3.08	Valid-2	−5.29	−4.30	−5.23
2016_Furukawa	1240	1.3-27	3.42	Valid-2	−5.65	−4.00	−5.57
2016_Furukawa	1299	1.5-06	0.83	Valid-3	−6.64	−6.81	−6.75
2016_Furukawa	1308	1.5-15	3.15	Valid-3	−5.18	−4.86	−5.34
2016_Furukawa	1311	1.5-18	1.71	Test	−6.58	−5.85	−6.80
2016_Furukawa	1345	1.6-12	1.86	Valid-3	−6.44	−4.57	−6.29
2016_Furukawa	1371	1.6-38	6.24	Valid-1	−8.00	−4.66	−8.00
2016_Furukawa	1388	1.7-15	6.14	Valid-1	−8.00	−4.24	−7.75
2016_Furukawa	1409	1.7-36	5.93	Valid-2	−7.62	−4.81	−7.66
2016_Furukawa	1424	1.8-11	2.26	Valid-1	−6.23	−5.08	−6.20
2016_Furukawa	1425	1.8-12	0.97	Test	−7.04	−4.80	−6.87
2016_Furukawa	1449	1.8-36	4.47	Test	−7.42	−4.54	−7.29
2016_Furukawa	1451	1.8-38	5.35	Valid-3	−8.00	−4.32	−7.99
2016_Furukawa	1454	1.9-01	1.54	Test	−5.52	−4.51	−5.66
2016_Furukawa	1471	1.9-18	3.66	Valid-2	−5.16	−5.90	−5.28
2016_Furukawa	1489	1.9-36	4.89	Valid-3	−6.93	−5.31	−6.74

a lipid bilayer [17]. We compared their prediction results with the CycPeptMP results for 23 peptides (Table 4.15) included in three validation sets (16 peptides) and the test set (7 peptides).

While the MD-based method could not successfully predict the membrane permeability of these 23 peptides (MAE = 1.521), CycPeptMP accurately predicted them all (MAE = 0.107) (Fig. 4.15). Hydrophobic cyclic peptides have insufficient solubility, slowly diffuse in the unstirred water layer, and are likely adsorbed to the membrane. Therefore, these behaviors could not be reproduced using the inhomogeneous solubility-diffusion model (ISDM), which only considers direct membrane permeation processes [17]. They reported a prediction accuracy (R) of only 0.21 for all 100 six-residue peptides; however, the accuracy increased to 0.54 when 33 hydrophobic peptides (AlogP $\geq$ 4) were excluded. A similar trend was observed among the 23 peptides com-

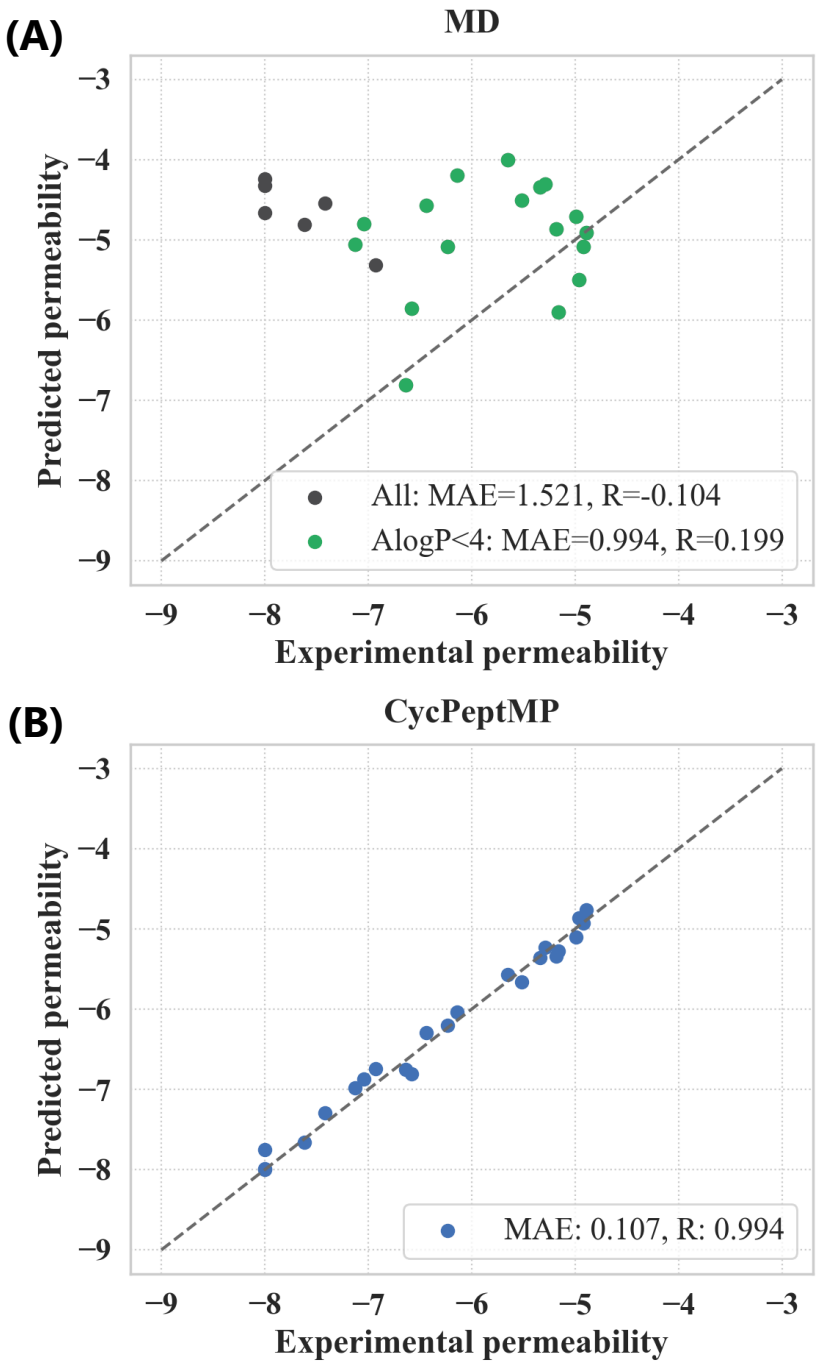


Figure 4.15: (A) Prediction results of the MD-based method. Black dots represent hydrophobic peptides with AlogP $\geq$ 4; green dots represent the remaining peptides with AlogP $<$ 4. (B) Prediction results of CycPeptMP.

pared in this study: the MAE of the MD-based method improved from 1.521 to 0.994 by excluding six peptides with AlogP ≥ 4. Overall, CycPeptMP can accurately and rapidly predict peptide permeability with superior performance compared to MD-based methods, representing a promising tool for cyclic peptide drug discovery.

4.5 Summary

In this chapter, we introduced CycPeptMP, a fusion model-based prediction model to predict the membrane permeability of cyclic peptides. To construct this model, we first curated a comprehensive dataset (CycPeptMPDB) containing 7,334 data points from over 40 publications. CycPeptMP outperformed several baseline models, including traditional ML-based cyclic peptide permeability prediction approaches, as well as state-of-the-art DL-based small molecule property prediction methods. Ablation studies confirmed the importance of this multi-level approach, as removing any component decreased prediction accuracy. Moreover, we confirmed that CycPeptMP accurately predicts the permeability of peptides with much lower computational costs where MD-based methods fail. With its ability to rapidly identify high-permeability peptides, CycPeptMP has the potential to significantly advance cyclic peptide drug discovery. It also paves the way for developing more effective DL-based techniques in related fields. Future studies should focus on improving prediction performance by generating 3D conformations using a more rigorous method.

Chapter 5

Development of a PPB Rate Prediction Model of Cyclic Peptides (CycPeptPPB)

5.1 Introduction

As we mentioned in Chapter 1 and Chapter 2, in contrast to the prediction of membrane permeability, few computational methods exist to predict the PPB rate for cyclic peptides due to the scarce available data. While both the peptide substructure and the overall conformational change significantly impact membrane permeability, the substructural information may be more important for PPB because the substructure forms a specific bond with plasma proteins (as shown in Fig. 1.11, the hydrocarbon side chain of dalbavancin is inserted deeply into the hydrophobic pocket of HSA).

In this chapter, we describe the development of a PPB rate prediction model of cyclic peptides, CycPeptPPB. We obtained 380 experimental data from collaborations with pharmaceutical companies and published literature, then applied the multi-level molecular features design and data augmentation methods proposed in Chapter 3 as well as the membrane permeability prediction. CycPeptPPB achieved excellent performance with the monomer-level feature, significantly improving prediction accuracy over existing methods.

5.2 Materials and Methods

5.2.1 Experimental dataset

Studies investigating the PPB of cyclic peptides have not significantly advanced, and there are less than 50 peptides available. Therefore, we collaborated with PeptiDream Inc., a leading company in cyclic peptide drug discovery, to provide us with many peptides with excellent PPB rates under a non-disclosure agreement. This allowed us to build a prediction model and to have a scientific discussion. In this study, private data containing 347 peptides were provided by PeptiDream Inc.; 16 synthesized peptide data and 17 approved peptide drug data with their experimentally determined PPB rates were collected from published literature. Fig. 5.1 show the distributions of experimentally determined PPB rate and molecular weight of cyclic peptides in each dataset.

PeptiDream (PD) dataset

The PD dataset is composed of 347 cyclic peptides designed and assayed by PeptiDream Inc. It covers a wide range of molecular weights (from 858.1 to 2247.7). Details of their structural information were confidential; however, many peptides have N-methylated residues, reducing the number of hydrogen bond donors and increasing lipophilicity. Since higher lipophilicity tends to enhance the PPB rate, peptides with %PPB $\geq$ 80% occupied approximately 80% (272) of the whole dataset.

Tajimi dataset

Tajimi *et al.* [111] designed 16 cyclic peptides composed only of natural amino acids and conducted %PPB measurement experiments using the equilibrium dialysis method. These peptides are relatively small, with rings consisting of seven to nine residues and molecular weights ranging from 809.0 to 1364.5. In this study, these 16 cyclic peptides were used as the Tajimi dataset (Table 5.1). All these cyclic peptides were cyclized by disulfide bonds between the N-terminal and C-terminal cysteine residues. Unlike the PD dataset, the peptides in the Tajimi dataset have not been optimized for enhanced lipophilicity, such as N-methylation. Therefore, these peptides had extremely low %PPB; no compound had more than 90%, and only two had more than 80%.

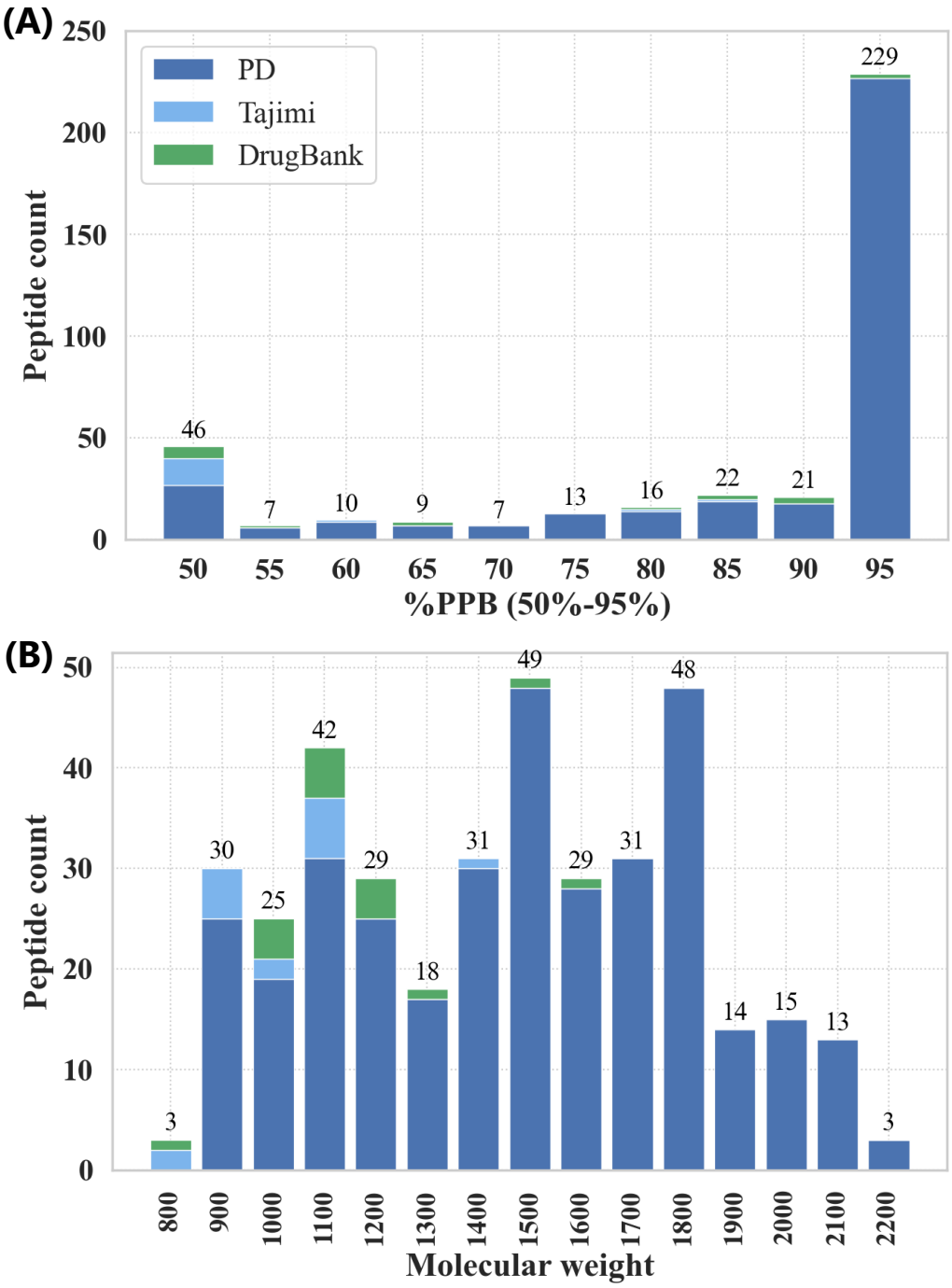


Figure 5.1: Distributions of experimental data. (A) Objective variable $\%PPB_{50-95}$. Data of $\leq 50\%$ is included in the leftmost bar, and data of $\geq 95\%$ is included in the rightmost bar. (B) Molecular weight (MolWt descriptor calculated by RDKit). Blue, orange, and green bars indicate PD (PeptiDream), Tajimi, and DrugBank datasets, respectively.

Table 5.1: Amino acid sequence and experimental %PPB of the Tajimi dataset.

Peptide	Amino acid sequence									%PPB
Pep.1	Cys	Tyr	Phe	Gln	Asn	Pro	Arg	Gly	Cys	24.2
Pep.2	Cys	Tyr	Ile	Gln	Asn	Pro	Leu	Gly	Cys	0.5
Pep.3	Cys	Ala	Trp	Lys	Val	Thr	Cys			0.04
Pep.4	Cys	Phe	Pro	Phe	Trp	Lys	Tyr	Cys		61.6
Pep.5	Cys	Trp	Arg	Pro	Arg	Val	Ala	Arg	Cys	0.0
Pep.6	Cys	Phe	Phe	Trp	Lys	Thr	Thr	Cys		26.3
Pep.7	Cys	Lys	Leu	Leu	Lys	Lys	Thr	Cys		0.0
Pep.8	Cys	Tyr	Tyr	Tyr	Tyr	Tyr	Tyr	Tyr	Cys	85.5
Pep.9	Cys	Ala	Gly	Leu	Val	Leu	Ala	Ala	Cys	0.0
Pep.10	Cys	Trp	Val	His	Pro	Gln	Phe	Glu	Cys	36.7
Pep.11	Cys	Asn	Gln	Pro	Trp	Gln	Cys			0.0
Pep.12	Cys	Ser	Phe	Asp	Asp	Trp	Leu	Ala	Cys	80.0
Pep.13	Cys	Tyr	Leu	Ala	Glu	Tyr	His	Gly	Cys	34.9
Pep.14	Cys	Ala	Pro	Ala	Trp	Ala	His	Gly	Cys	7.4
Pep.15	Cys	Phe	Val	Tyr	Ser	Ala	Val	Cys		15.3
Pep.16	Cys	Arg	Ile	Lys	Arg	Tyr	Cys			15.1

DrugBank dataset

Tajimi *et al.* [111] collected 24 cyclic peptide PPB data from the FDA-approved drug public database DrugBank [191]. From these data, we extracted 17 cases that are applicable to the other two datasets (containing one ring and composed of five residues or more). These 17 public drug data were used as the DrugBank dataset in this study. It is noted that there are data in which the PPB rate in the paper differs from the PPB rate in DrugBank, and there are data in which the PPB rate is not recorded in DrugBank. Therefore, we used the PPB rates from DrugBank for the data with PPB rates in DrugBank (accessed on 1/2/2021), and obtained PPB rates for the other data by survey (Table 5.2). The range of %PPB was wide, and eight peptides exceeded 80%. Unlike other datasets, some of these cyclic peptides contained fatty acid side chains; such peptides could exhibit high PPB rates even with low lipophilicity.

Division of training, validation, and test sets

Since the peptides of the DrugBank dataset are structurally diverse, we thought it was suitable for verifying the generalization performance of the prediction model and using it as an external test set. On the other hand, the PD and Tajimi datasets were

Table 5.2: %PPB reported in prior research and our surveyed %PPB (1/2/2021 accessed; experimental values used in this study) of the DrugBank dataset. If multiple %PPB values are listed, the average value is used, and the original range is shown in the parentheses.

Peptide	Survey source	Tajimi *et al.* [111] reported value	Our surveyed value
Acetyl-Daptomycin	Schneider *et al.* [13]	12%	12%
Anidulafungin	DrugBank	99%	84%
Caspofungin	DrugBank	97%	97%
Colistin	Couet *et al.* [192]	90.4%–92.9%	56% (55%–57%)
Cyclosporin A	DrugBank	90%	90%
Daptomycin	DrugBank	85%	91.5% (90%–93%)
Desmopressin	DrugBank	50%	17.3% (15.8%–18.8%)
Eptifibatide	DrugBank	25%	25%
Lanreotide	Medscape	79%–83%	81% (79%–83%)
Micafungin	DrugBank	99%	99%
Octreotide	DrugBank	65%	65%
Oxytocin	DrugBank	30%	30%
Pasireotide	DrugBank	88%	88%
Polymyxin B	DrugBank	55.9%	85.5% (79%–92%)
Quinupristin	Bearden [193]	55%–78%	66.5% (55%–78%)
Terlipressin	DrugBank	30%	30%
Vasopressin	DrugBank	1%	1%

relatively similar in structure; they were mixed and split into training, validation, and internal test sets. Similarly to membrane permeability predictions, we employed the KS algorithm to extract 10% of all data (36 peptides) as the internal test set based on 2048-bit Morgan FP (Algorithm 4.1). From the remaining data, we randomly extracted 10% validation sets (36 peptides) three times for parameter tuning, with no overlap between the three datasets. The distribution of each dataset in the PCA space (based on 2048-bit Morgan FP) is shown in Fig. 5.2. The average MAE (%), MSE, R, and R^2 from three repeated runs were used as evaluation metrics.

Preprocessing for objective variable

The objective variables used in PPB rate prediction studies for conventional small molecule compounds can be broadly divided into two types: the type where the binding ratio is used as it is [102, 103, 104] and the type where it is used after logarithmic conversion [101, 109]. In the case of conventional small molecules, the PPB rate tends

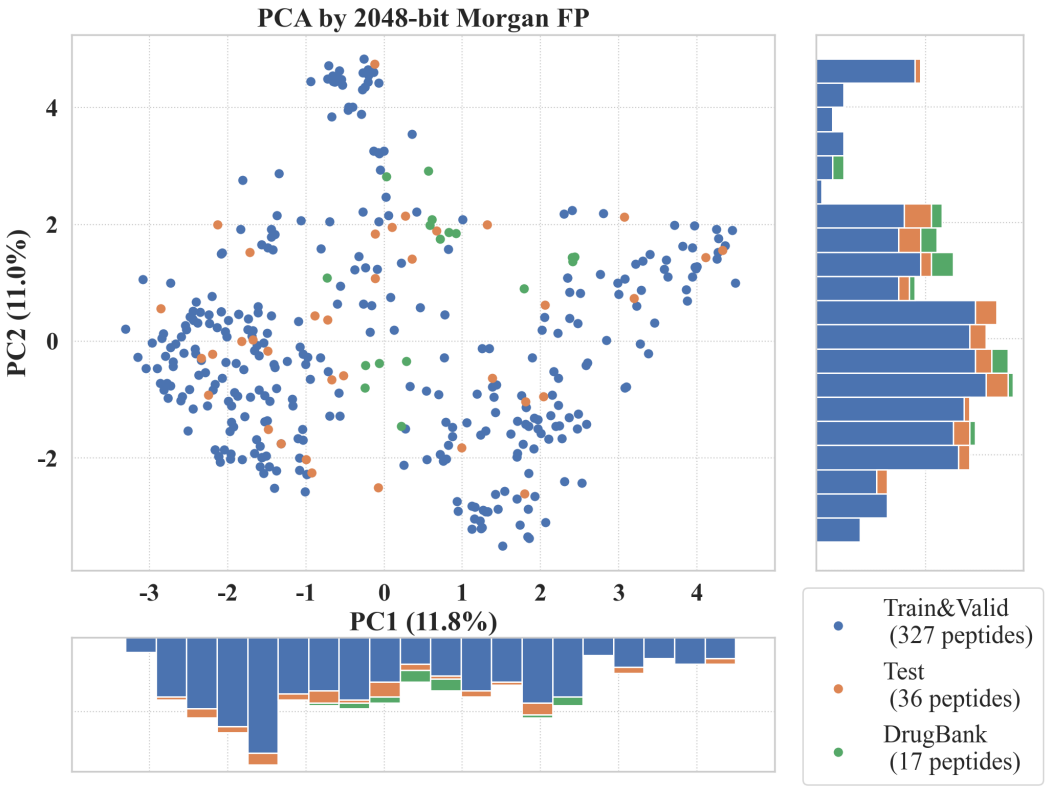


Figure 5.2: Experimental data distribution in PCA space, with the PC1 as the horizontal axis and the PC2 as the vertical axis; the contribution rates are shown in the parentheses of axes captions.

to be extremely high. Thus, logarithmic conversion, which has high resolution in the region of %PPB $>$ 90%, is considered suitable for the prediction of small molecules. However, the PPB rate of cyclic peptides is generally lower than small molecules. The prediction method for cyclic peptides should practically focus on peptides with %PPB ranging from 50% to 95%, and logarithmic conversion is not so effective in the range. Cyclic peptides with %PPB less than 50% are very unlikely to be effective as drugs, and there is no need to precisely quantify %PPB less than 50%. Therefore, in this study, we used the binding ratio as it is; an experimentally measured value of %PPB less than 50% was rounded to 50%, and %PPB higher than 95% was rounded to 95% ($\%PPB_{50-95}$; Fig. 5.1 (A)).

Table 5.3: Description of selected descriptors, arranged in order of RF feature importance.

Type	Name	Software	Description
2D	logP(o/w)	MOE	Log of the octanol/water partition coefficient
	AATS4se	Mordred	Averaged moreaubroto autocorrelation of lag 4 weighted by sanderson EN
	PEOE_VSA-1	MOE	Sum of van der Waals surface areas (calculated by a connection table approximation) for atoms within a certain range of partial charges
3D	vsurf_CW2	MOE	VolSurf capacity factor 2
	vsurf_CW3	MOE	VolSurf capacity factor 3

5.2.2 Descriptors selection

After the preprocessing of descriptors (Section 3.3.3), 387 (335 2D and 52 3D descriptors) peptide descriptors were selected. Subsequently, three 2D and two 3D peptide descriptors were selected based on the assigned feature importance from two RF models (one used 2D, and the other one used 3D descriptors). Fig. 5.3 shows the RF feature importance, Fig. 5.4 shown the heatmap of correlation matrix, and Table 5.3 shows the description of selected peptide descriptors.

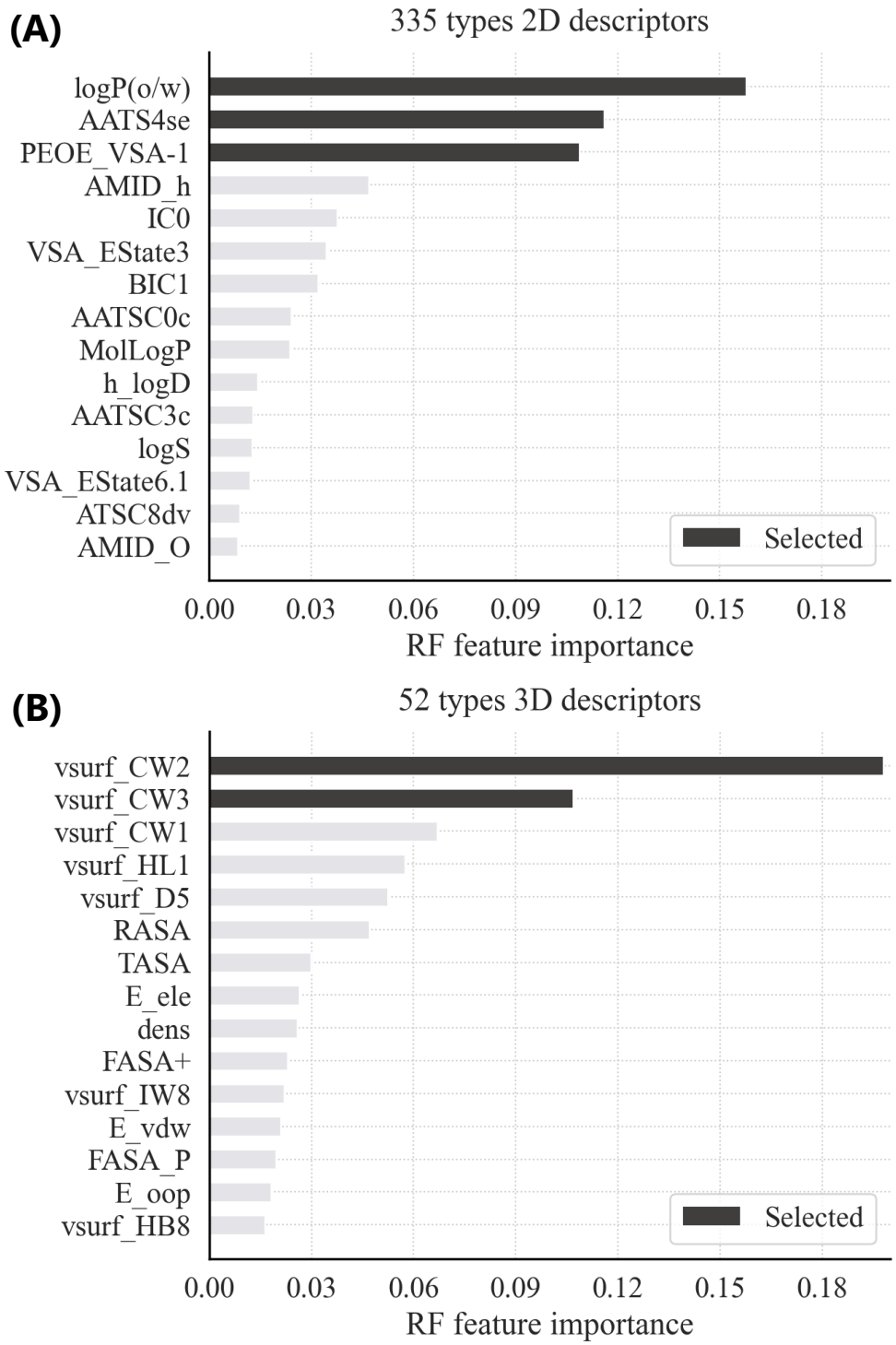


Figure 5.3: Top 15 peptide descriptors with the highest RF feature importance from (A) RF model with 2D descriptors and (B) RF model with 3D descriptors, respectively. Selected three 2D and two 3D peptide descriptors are shown in black.

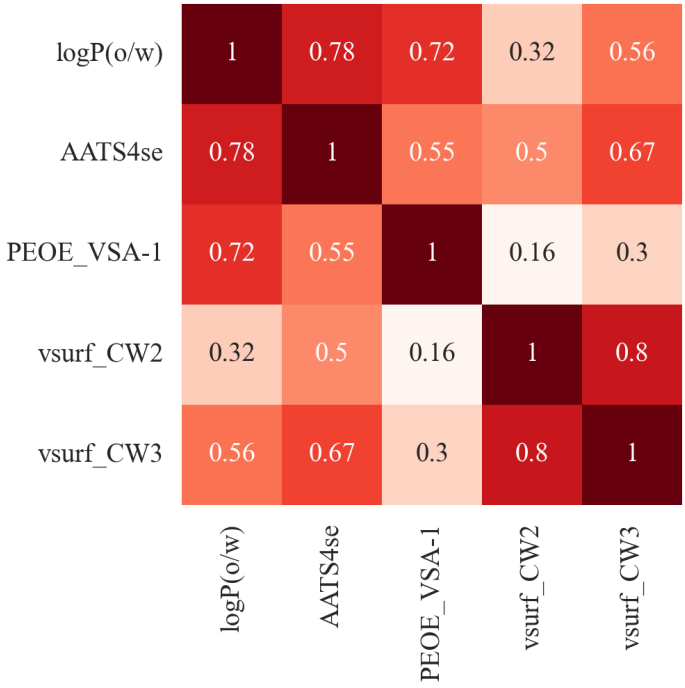


Figure 5.4: Heatmap of absolute correlation coefficient values for five selected peptide descriptors. The pair with the highest correlation is vsurf_CW2 and vsurf_CW3 ($|R| = 0.795$).

5.2.3 Hyperparameter search of proposed models

The partial structure of cyclic peptides heavily influences PPB; thus, it may be possible to make accurate predictions using a simpler model. Therefore, in addition to the fusion model, we also extracted a part of the fusion model to construct an atom model, a monomer model using 1D-CNN layers, and a monomer model using CyclicConv layers. Hyperparameters of these four models (the search range was the same as Table 4.5, only batch_size used $[32, 64, 128]$ due to the GPU memory limitation) were determined by 100 trials using Optuna software based on the average MSE of three runs. The search results are summarized in Table 5.4, and hyperparameters with high Optuna importance for fusion model search are shown in Fig. 5.5. As shown in Fig. 5.5, the type of activation function (ac) and the hyperparameters of the peptide model had a large effect. The activation function search results for all four proposed models were the same, GELU. Furthermore, the number of encoders (n_encoders) for the fusion mode and the atom model, and the number of convolutional layers (n_conv) for the fusion mode and two monomer models were the same.

Table 5.4: Hyperparameter search results of the four proposed models. "-" indicates that the hyperparameter does not apply to the respective model.

Objective	Hyperparameter	Fusion model	Atom model	Monomer model (1D-CNN)	Monomer model (CyclicConv)
Training	batch_size	128	128	128	64
	optimizer	NAdam	RAdam	NAdam	AdamW
	weight_deacy	1e-3	5e-4	5e-3	5e-5
All models	d_linear	256	128	256	256
	d_subout	16	1	1	1
	ac	GELU	GELU	GELU	GELU
Atom model	n_encoders	3	3	-	-
	dropout	0.1	0.1	-	-
	n_head	8	4	-	-
	d_model	32	64	-	-
	d_feedforward	512	128	-	-
	n_linears	2	1	-	-
	λ_g	0.2	0.7	-	-
Monomer model	n_conv	3	-	3	3
	conv_type	CyclicConv	-	1D-CNN	CyclicConv
	padding	0	-	1	0
	d_conv	[256, 64, 256]	-	[256, 256, 128]	[128, 128, 128]
	pooling	ave	-	ave	max
	n_linears	2	-	2	2
Peptide model	n_mlps	6	-	-	-
	dropout	0.2	-	-	-
	d_mlp	256	-	-	-
Shared layer	n_shared	1	-	-	-

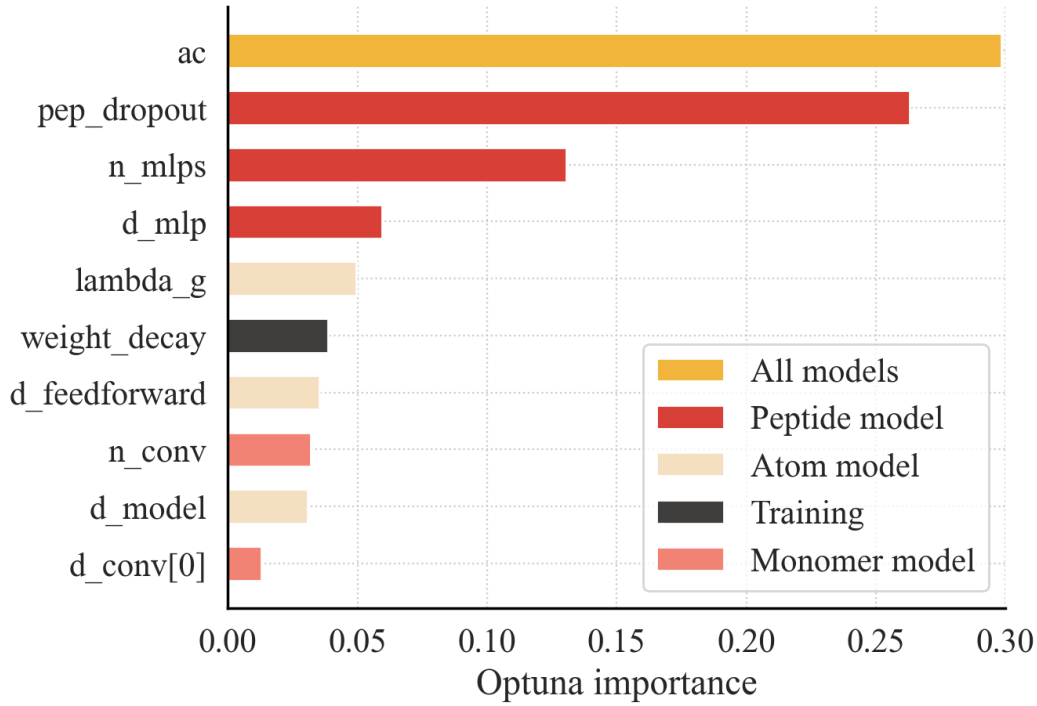


Figure 5.5: Top ten hyperparameters with an Optuna importance > 0.01 on the CycPeptMP hyperparameter search. Optuna importance is calculated based on the fANOVA hyperparameter importance evaluation algorithm [12]; the sum of the importance values is normalized to 1.0.

5.2.4 Baseline methods

In contrast to membrane permeability prediction, the training data for PPB rate prediction is limited to only 291 peptides. Furthermore, the maximum atom number of experimental data (162) is also larger than that of membrane permeability experimental data (128). Therefore, it is considered difficult to predict the PPB rate of cyclic peptides using DL-based small molecule property prediction methods, and such methods were not used for comparison. We validated the performance of four proposed models based on comparisons with four baseline methods.

- Small molecule PPB rate prediction commercial software: ADMET Predictor (version 10.0) [194] is a cheminformatics software for small molecules that can predict over 175 ADME-Tox properties, including PPB rates. The PPB rate prediction model was built based on 1,986 small molecule data collected from the literature. To compare with existing PPB rate prediction software, this study used ADMET Predictor as a comparison method.

- Three traditional baselines: As well as membrane permeability predictions, we constructed an RF model with 2048-bit Morgan FP, an SVM model with three 2D peptide descriptors, and an SVM model with five 2D and 3D peptide descriptors to represent traditional ML-based PPB rate prediction methods. The hyperparameters of the RF and SVM models were determined by a grid search (Table 5.5).

Table 5.5: Search range and results of the hyperparameter search (grid search) for the RF and two SVM models.

Objective	Hyperparameter	Description	Search range	Search result
RF model	n_estimators	Number of decision trees	50, 100, 200, 300, 500, 750, 1000	200
	max_depth	Maximum depth of each decision tree	None, 2, 5, 10, 20, 30	20
SVM model	kernel	Kernel function	-	Gaussian kernel (rbf)
	C	Penalty parameter	2^{-3}, 2^{-2}, 2^{-1}, 2^{0}, 2^{1}, 2^{2}, 2^{3}, 2^{4}, 2^{5}	2^{5} (SVM-2D), 2^{4} (SVM-2D3D)
	γ	Kernel coefficient of Gaussian kernel	2^{-6}, 2^{-5}, 2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}, 2^{0}	2^{-1} (SVM-2D), 2^{-5} (SVM-2D3D)

Table 5.6: Performance comparison between four baseline methods and four proposed models using the test set. Except for the ADMET Predictor, the metrics are the averaged values of three repeated runs; the best result for each metric is indicated in bold.

Metrics	ADMET Predictor	RF	SVM-2D	SVM-2D3D
MAE (%)	11.00	10.35 ± 0.36	4.41 ± 0.08	6.23 ± 0.13
MSE	211.7	183.1 ± 14.0	76.7 ± 2.1	97.6 ± 1.5
R	0.740	0.786 ± 0.017	0.902 ± 0.004	0.887 ± 0.001
R^2	0.476	0.547 ± 0.035	0.810 ± 0.005	0.758 ± 0.004
Metrics	Atom	Monomer (1D-CNN)	Monomer (CyclicConv)	Fusion
MAE (%)	5.41 ± 0.79	4.49 ± 0.39	3.90 ± 0.14	$\mathbf{2.44 \pm 0.29}$
MSE	105.2 ± 24.0	66.3 ± 8.9	45.9 ± 8.2	$\mathbf{28.2 \pm 4.7}$
R	0.882 ± 0.023	0.922 ± 0.014	0.951 ± 0.008	$\mathbf{0.973 \pm 0.004}$
R^2	0.739 ± 0.059	0.836 ± 0.022	0.886 ± 0.020	$\mathbf{0.930 \pm 0.012}$

5.3 Results and Discussion

5.3.1 Performance comparison

The prediction accuracy for the test and DrugBank sets by all eight models are shown in Table 5.6 and Table 5.7, respectively. In addition, the experimental %PPB and predicted %PPB for the test and DrugBank sets by all models are shown in Fig. 5.6 and Fig. 5.7, respectively.

Performance comparison for the test set

According to the prediction results shown in Table 5.6, ADMET Predictor exhibited the worst prediction accuracy for the test set (MAE = 11.00%, R = 0.740). This result showed that the conventional method for small molecules cannot accurately predict %PPB of cyclic peptides. RF model based on 2048-bit Morgan FP showed similar poor prediction accuracy as ADMET Predictor (MAE = 10.35%, R = 0.786), indicating that it may not fully capture the complexity of cyclic peptides. SVM-2D (MAE = 4.41%, R = 0.902) and SVM-2D3D (MAE = 6.23%, R = 0.887) models, using the peptide descriptors calculated from the entire cyclic peptide, exhibited higher prediction accuracy than the ADMET Predictor and RF model. These results indicated that the proposed peptide descriptor selection method worked correctly and could select appropriate descriptors for prediction; however, adding 3D descriptors did not improve prediction accuracy. For the four proposed models, the prediction accuracy of the fusion model was the highest among all eight methods (MAE = 2.44%, R = 0.973), showing

Table 5.7: Performance comparison between four baseline methods and four proposed models using the DrugBank set. Except for the ADMET Predictor, the metrics are the averaged values of three repeated runs; the best result for each metric is indicated in bold.

Metrics	ADMET Predictor	RF	SVM-2D	SVM-2D3D
MAE (%)	15.08	19.49 ± 0.13	13.36 ± 0.07	10.89 ± 0.16
MSE	280.4	483.3 ± 3.9	319.1 ± 2.7	246.4 ± 4.6
R	0.631	-0.236 ± 0.033	0.467 ± 0.005	0.581 ± 0.008
R^2	0.152	-0.461 ± 0.012	0.035 ± 0.008	0.255 ± 0.014
Metrics	Atom	Monomer (1D-CNN)	Monomer (CyclicConv)	Fusion
MAE (%)	12.90 ± 0.49	$\mathbf{4.40 \pm 0.31}$	7.69 ± 0.08	8.53 ± 0.47
MSE	242.3 ± 15.8	$\mathbf{39.3 \pm 7.0}$	110.5 ± 5.5	131.2 ± 13.8
R	0.580 ± 0.030	$\mathbf{0.947 \pm 0.010}$	0.835 ± 0.013	0.783 ± 0.025
R^2	0.267 ± 0.048	$\mathbf{0.881 \pm 0.021}$	0.666 ± 0.016	0.603 ± 0.042

that the multi-level feature design method was effective in predicting the test set. The monomer model using CyclicConv ranked second among all methods (MAE = 3.90%, R = 0.951), and the monomer model using 1D-CNN showed a high prediction accuracy (MAE = 4.49%, R = 0.922), although it was slightly lower than CyclicConv model. The prediction accuracy of the atom model was comparable to that of two SVM models and was the lowest among the proposed methods (MAE = 5.41%, R = 0.882). Overall, monomer-level information may be the most effective for predicting the PPB rate of cyclic peptides.

Performance comparison for the DrugBank set

For the external DrugBank set, in contrast to the internal test set, all comparison methods exhibited comparable poor accuracy (MAE = 10.89% to 19.49%, R = −0.236 to 0.631). Due to the difficulty in predicting PPB rates for peptides with large side chains in the DrugBank set, the two SVM models based on peptide descriptors calculated from whole molecules performed poorly compared to the test set and could not predict PPB rates. Among the proposed methods, the atom model still had the lowest prediction accuracy (MAE = 12.90%, R = 0.580). The fusion model showed lower performance compared to the two monomer models (MAE = 8.53%, R = 0.783). This drop may be due to the dependence on peptide-level information (as shown in Fig. 5.5, hyperparameters of the peptide model had a large effect) and overfitting to the training set, which does not contain peptides with large side chain structures. The monomer model with 1D-CNN exhibited the best performance (MAE = 4.40%, R = 0.947), and

the CyclicConv model was slightly lower than the 1D-CNN model (MAE = 7.69%, R = 0.835). These results demonstrated that our proposed monomer-level feature design method effectively captured the mechanistically important substructural information of the PPB of cyclic peptides. Therefore, the 1D-CNN monomer model was used as the representative model CycPeptPPB in this study, and its results were also used in the analysis of the contribution of substructures to the PPB rate prediction.

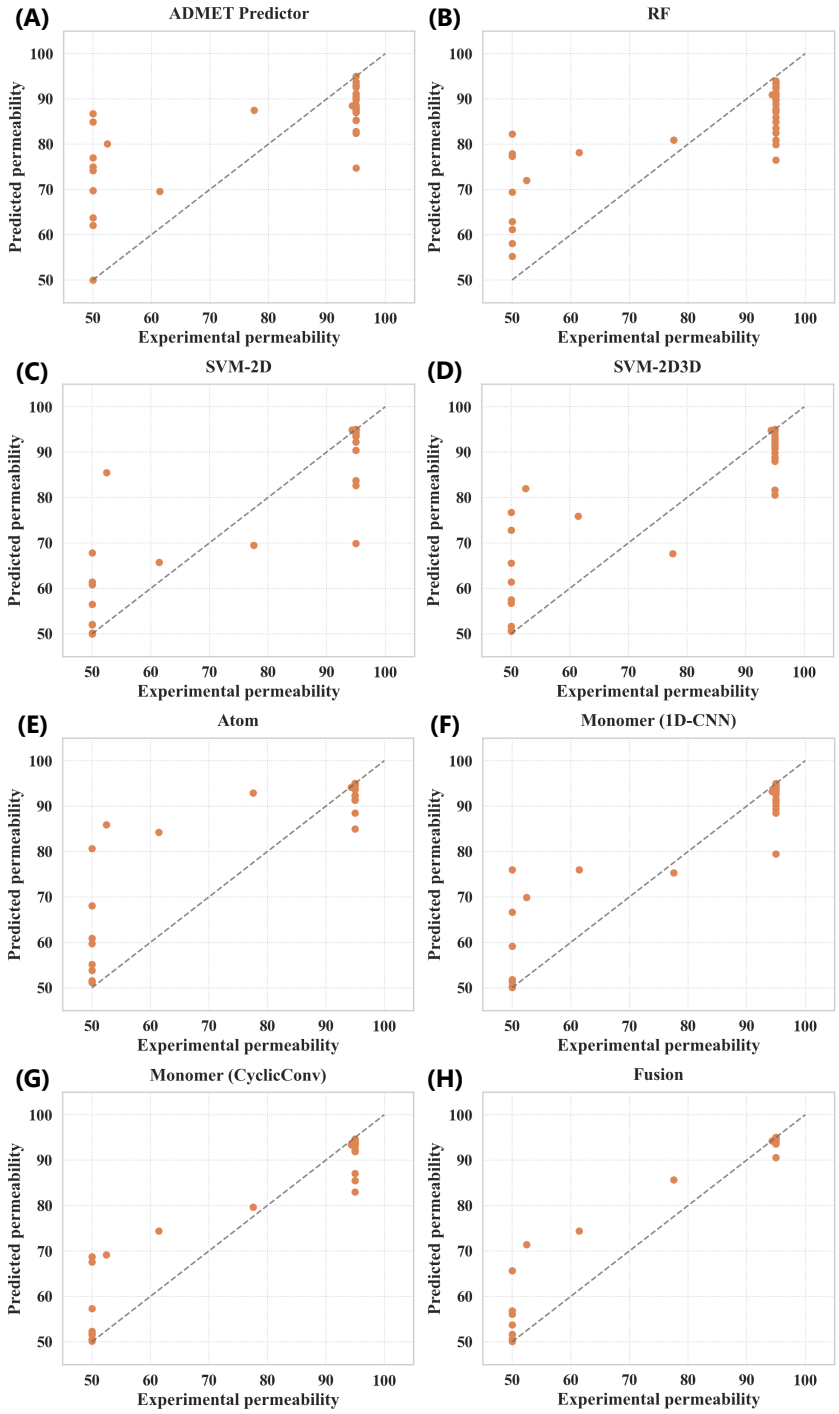


Figure 5.6: Prediction results of the test set by four baselines and four proposed methods. (A) ADMET Predictor, (B) RF model, (C) SVM-2D model, (D) SVM-2D3D model, (E) atom model, (F) monomer model (1D-CNN), (G) monomer model (CyclicConv), and (H) fusion model. Except for the ADMET Predictor, the predicted values are the average value of three runs.

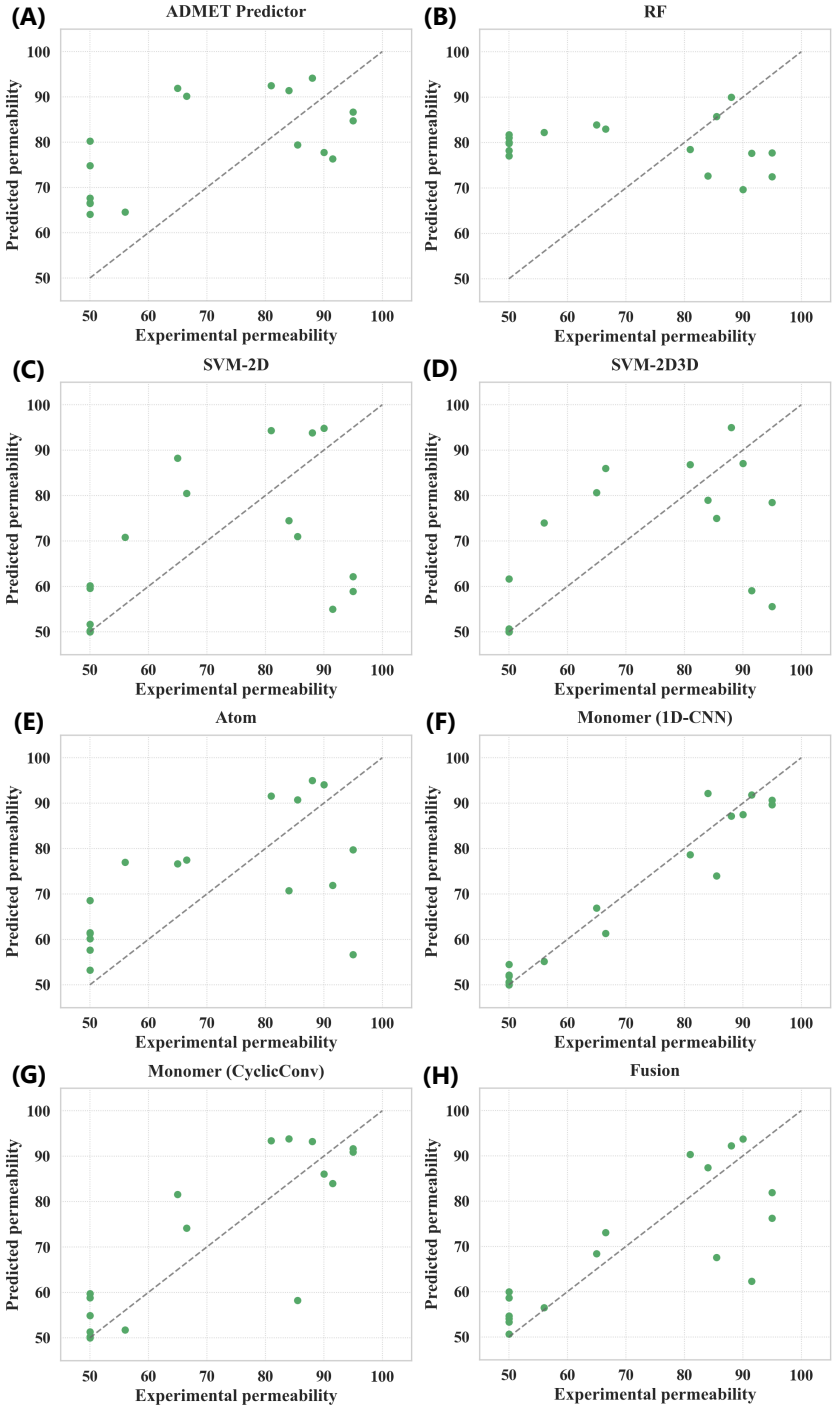


Figure 5.7: Prediction results of the DrugBank set by four baselines and four proposed methods. (A) ADMET Predictor, (B) RF model, (C) SVM-2D model, (D) SVM-2D3D model, (E) atom model, (F) monomer model (1D-CNN), (G) monomer model (CyclicConv), and (H) fusion model. Except for the ADMET Predictor, the predicted values are the average value of three runs.

5.3.2 Ablation study for the fusion and monomer models

We conducted ablation studies on the fusion and 1D-CNN monomer models. The ablation studies measured the influences of the number of replicas generated by augmentation and changes in model architecture. Fig. 5.8 shows the accuracy based on 1 (no augmentation), 10, 20, 30, 40, 50, and 60 replicas per peptide for fusion (start with F–) and monomer (start with M–) models. For the fusion model, augmentation improved the prediction accuracy for both the test set (F–1: 4.39%, F–60: 2.44%) and the DrugBank set (F–1: 11.11%, F–60: 8.53%). Note that, similar to the ablation results for permeability prediction (Fig. 4.14 (A)), over 30 replicas showed approximately the same prediction accuracy as the replica increased. For the monomer model, the use of augmentation significantly improved prediction accuracy for both the test set (M–1: 31.13%, M–60: 4.49%) and the DrugBank set (M–1: 20.44%, M–60: 4.40%). In contrast to the fusion model, whose prediction results were relatively stable without augmentation, the monomer model without augmentation had very low prediction accuracy, which increased as the replica increased. Overall, we were able to confirm the effectiveness of our proposed data augmentation method for cyclic peptides in improving the training efficiency of structurally complex cyclic peptides.

Fig. 5.9 shows the accuracy for different architectures for fusion (start with F–) and monomer (start with M–) models. F is the original fusion model; F–atom, F–mono, and F–pep represent the models lacking the respective sub-models; F–3D is the model that did not use all 3D information ($Conf$ and 3D descriptors). M is the original 1D-CNN monomer model, and M–3D is the model that did not use 3D monomer descriptors. For the fusion model, prediction accuracy decreased when atom or monomer models were removed on both two datasets. When the peptide model was removed, the accuracy on the test set decreased (F: 2.44%, F–pep: 3.82%); however, the accuracy on the DrugBank set increased (F: 8.53%, F–pep: 8.21%). The use of whole-molecule information was counterproductive when predicting the PPB rate of cyclic peptides with large side chains. The use of 3D information in the fusion model improved prediction accuracy (F–3D: 3.34%, 9.71%). For the monomer model, the use (M: 4.49%, 4.40%) or lack (M–3D: 4.54%, 4.53%) of 3D monomer descriptors had little effect.

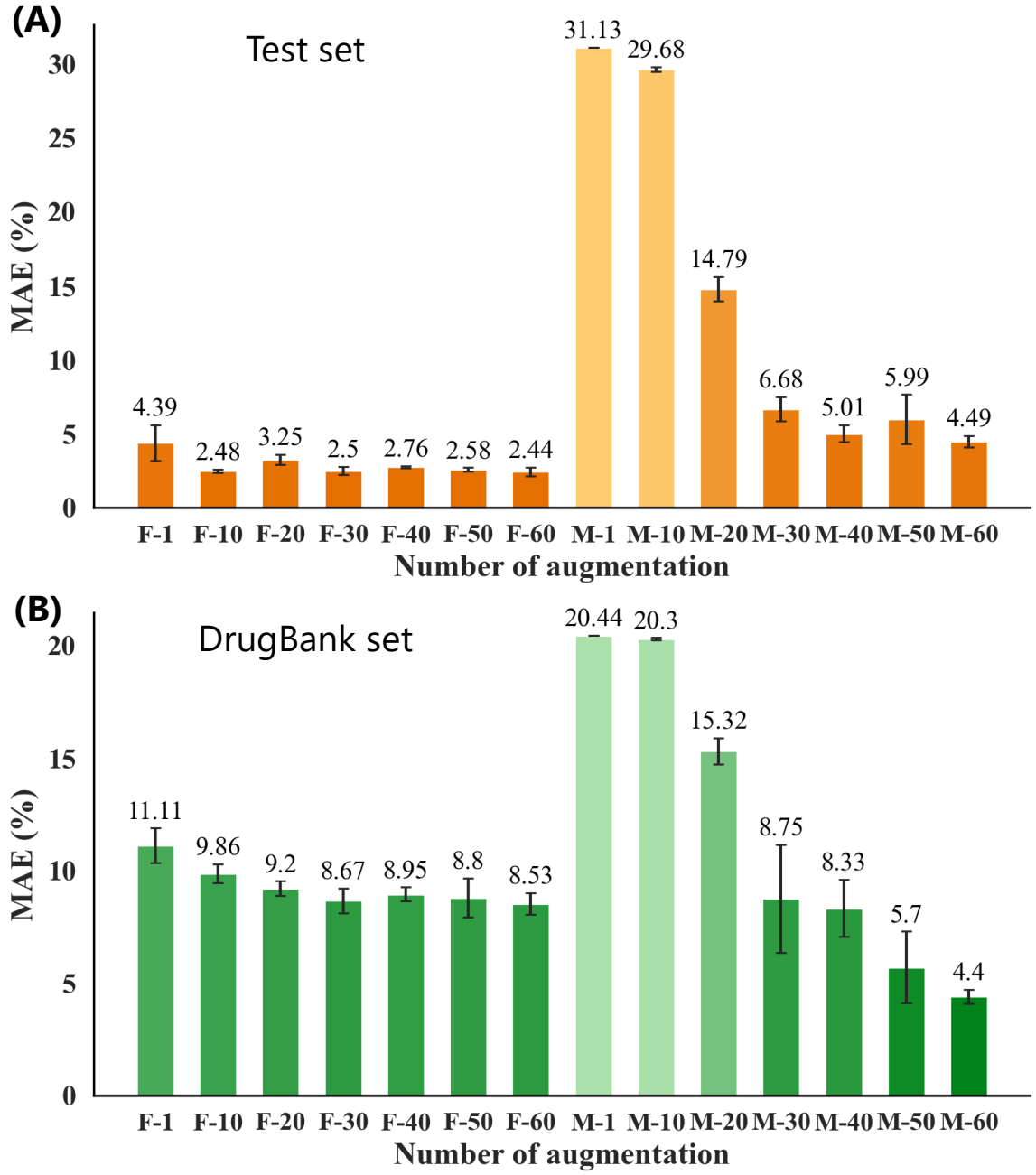


Figure 5.8: Ablation results (%MAE) for different input replica numbers for the fusion and 1D-CNN monomer models. (A) Results on the internal test set. (B) Results on the external test (DrugBank) set.

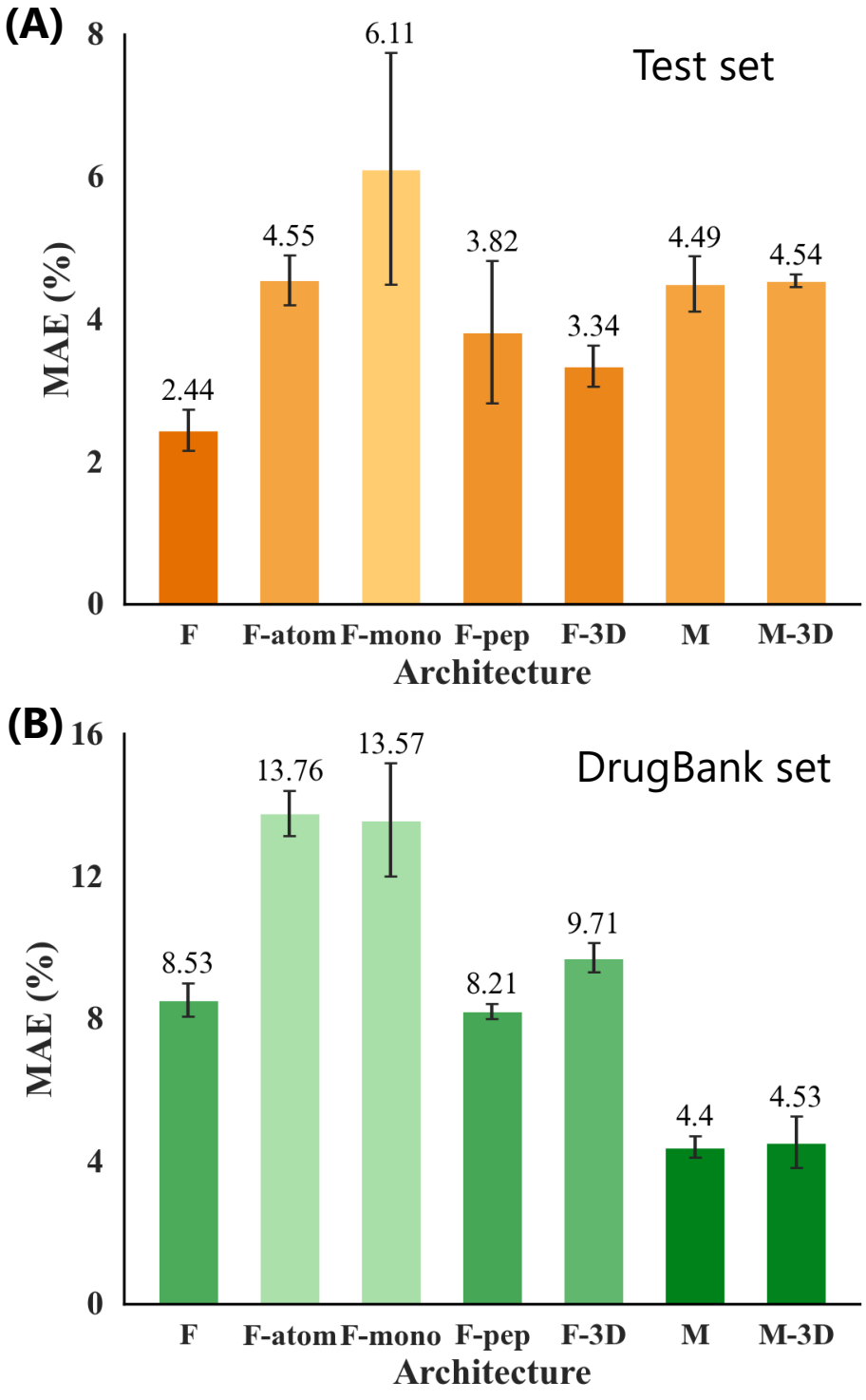


Figure 5.9: Ablation results (%MAE) for different architectures for the fusion and 1D-CNN monomer models. (A) Results on the internal test set. (B) Results on the external test (DrugBank) set.

5.3.3 Analysis of prediction results of peptide pairs in the DrugBank set with similar main structures

Cyclic peptides composed of natural amino acids of x residues have 20^x combinations, and it is computationally impossible to perform a trial-and-error based search of the entire space for a better PPB rate. Therefore, from the perspective of efficient design and optimization for the development of cyclic peptide drugs, determining which monomer enhances %PPB and which has no effect on %PPB is necessary. The saliency map was originally defined as a heatmap that estimates the parts of a visual image to which people pay attention when viewing it. Many methods have been reported to calculate the salience map of deep learning models [195, 196], and VanillaGrad [197] is one of the simplest methods. We analyzed the contribution of each monomer to the PPB rate prediction for the 1D-CNN monomer model using the saliency score calculated by VanillaGrad.

Calculation of saliency score

For a input (monomer descriptors) $\boldsymbol{x}_{ij}$ of replica j for peptide i, the saliency score $\boldsymbol{S}_{ij}^{k}$ based on the predicted PPB value $\hat{y}_{ij}^{k}$ of repeated runs k was calculated as:

$$\boldsymbol{S}_{ij}^{k} = \left| \frac{\partial \hat{y}_{ij}^{k}}{\partial \boldsymbol{x}_{ij}} \right| \tag{5.1}$$

The average saliency score across all three repeated runs was given by:

$$\bar{\boldsymbol{S}}_{ij} = \frac{1}{3} \sum_{k=1}^{3} \boldsymbol{S}_{ij}^{k} \tag{5.2}$$

After aligning the sequences, the saliency score averaged across all 60 replicas was calculated as:

$$\bar{\boldsymbol{S}}_{i} = \frac{1}{60} \sum_{j=1}^{60} \bar{\boldsymbol{S}}_{ij} \tag{5.3}$$

Finally, the normalized saliency score for peptide i was given by:

$$\boldsymbol{S}_{i} = (s_0, s_1, ..., s_m) = \frac{\bar{\boldsymbol{S}}_{i}}{\mathtt{Max}(\bar{\boldsymbol{S}}_{i})} \tag{5.4}$$

where m is the number of monomers of peptide i.

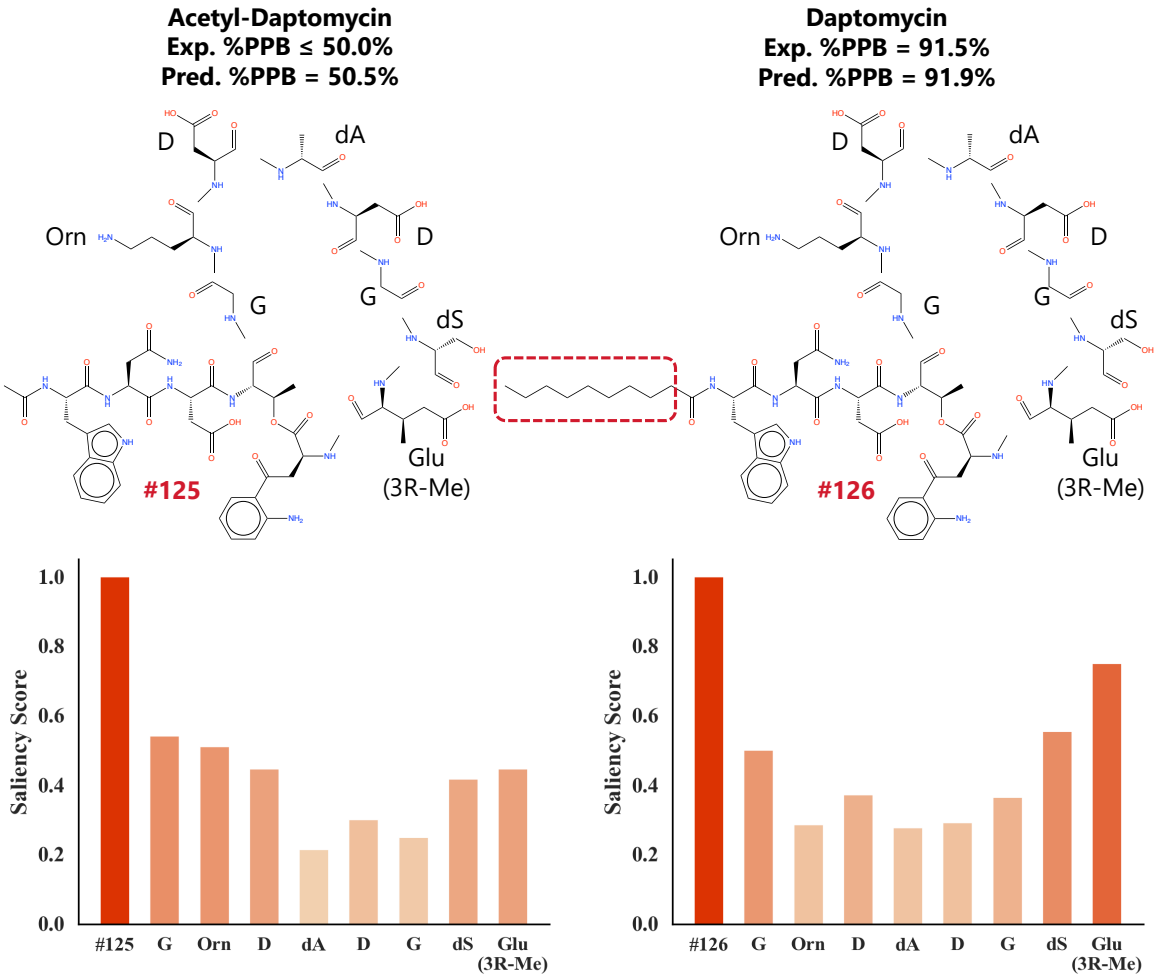


Figure 5.10: The monomers obtained by the decomposition procedure and these salience scores of acetyl-daptomycin (Exp.%PPB $\leq$ 50.0%, Pred.%PPB = 50.5%) and daptomycin (Exp.%PPB = 91.5%, Pred.%PPB = 91.9%). The salience score shows the average value of the salience score of all augmentation replicas across three repeated runs (normalized to a maximum value of 1).

Analysis of acetyl-daptomycin and daptomycin

DrugBank set contains some structurally similar peptide pairs. We analyzed the prediction results of acetyl-daptomycin and daptomycin (Fig. 5.10) among them, which differed only in the fatty acid side chain corresponding to the monomers #125 and #126 (a total of 126 monomers were obtained from the PPB experimental data, and #125 and #126 are the monomers with the second and first highest molecular weights, respectively). According to Schneider *et al.* [13], the N-terminal fatty acid side chain of daptomycin (a part of monomer #126) specifically binds to HSA by deeply pierc-

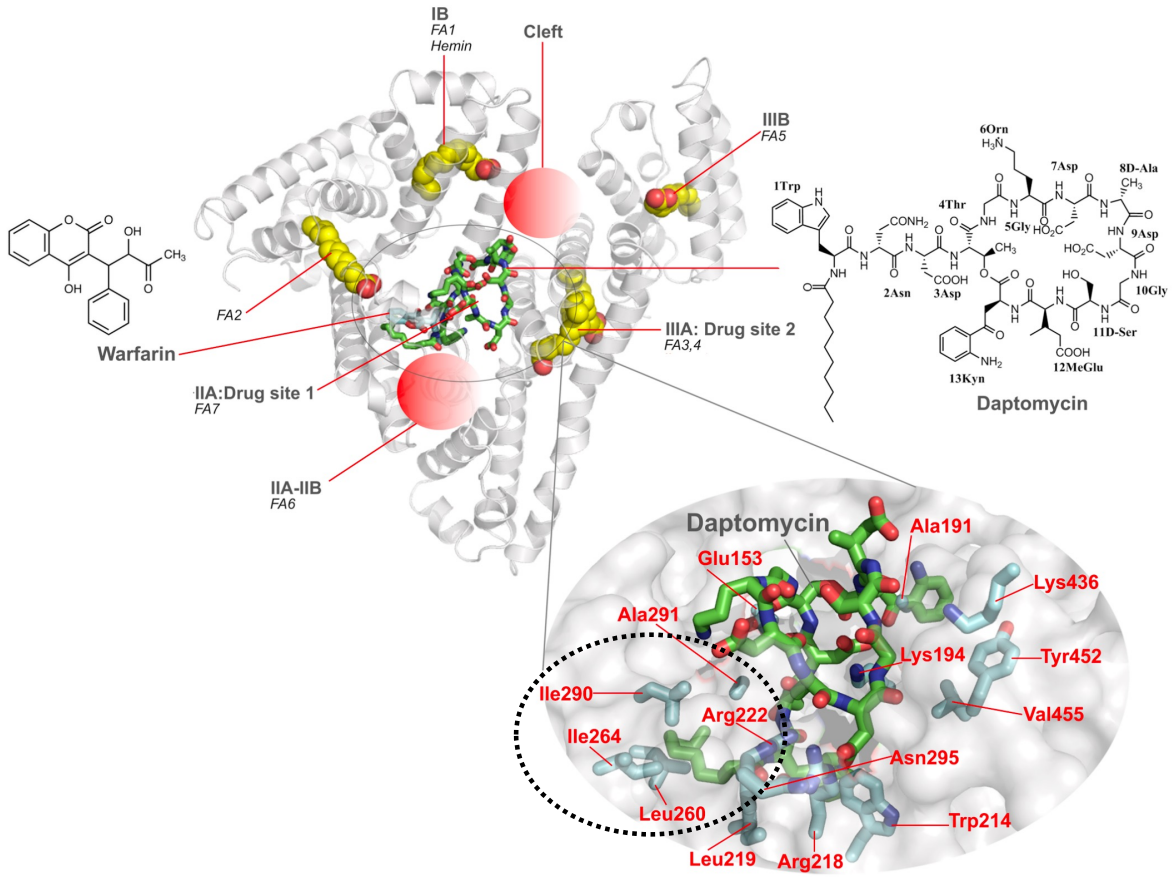


Figure 5.11: Docking results of daptomycin with Site 1 of HSA (cited from [13], added black dotted circle line). The N-terminal fatty acid side chain of daptomycin penetrates deeply into the binding pocket Site 1 of HSA and has extensive and strong hydrophobic interactions with the HSA residues Leu219, Leu260, Ile264, Ile290, and Ala291.

ing the binding pocket Site 1 (Fig. 5.11), resulting in the high PPB rate (91.5%). In contrast, acetyl-daptomycin does not have the fatty acid side chain, resulting in a low PPB rate (12.0%, and was rounded to 50.0% in this study). These differences, however, were hardly reflected in peptide descriptors; thus, Tajimi *et al.* [111] reported that it was difficult to distinguish them (predicted %PPB value by Tajimi *et al.* of acetyl-daptomycin: 18%; daptomycin: 49%). The same tendency was obtained with our peptide descriptor-based models (two SVM models, predicted $\%PPB_{50-95}$ value of acetyl-daptomycin: 59.6%, 50.0%; daptomycin: 55.0%, 59.1%). In contrast to these methods, our CycPeptPPB model (1D-CNN monomer model), which is based on monomer descriptors, correctly predicted their PPB rates (predicted $\%PPB_{50-95}$ value of acetyl-daptomycin: 50.5%; daptomycin: 91.9%). As shown in Fig. 5.10, both

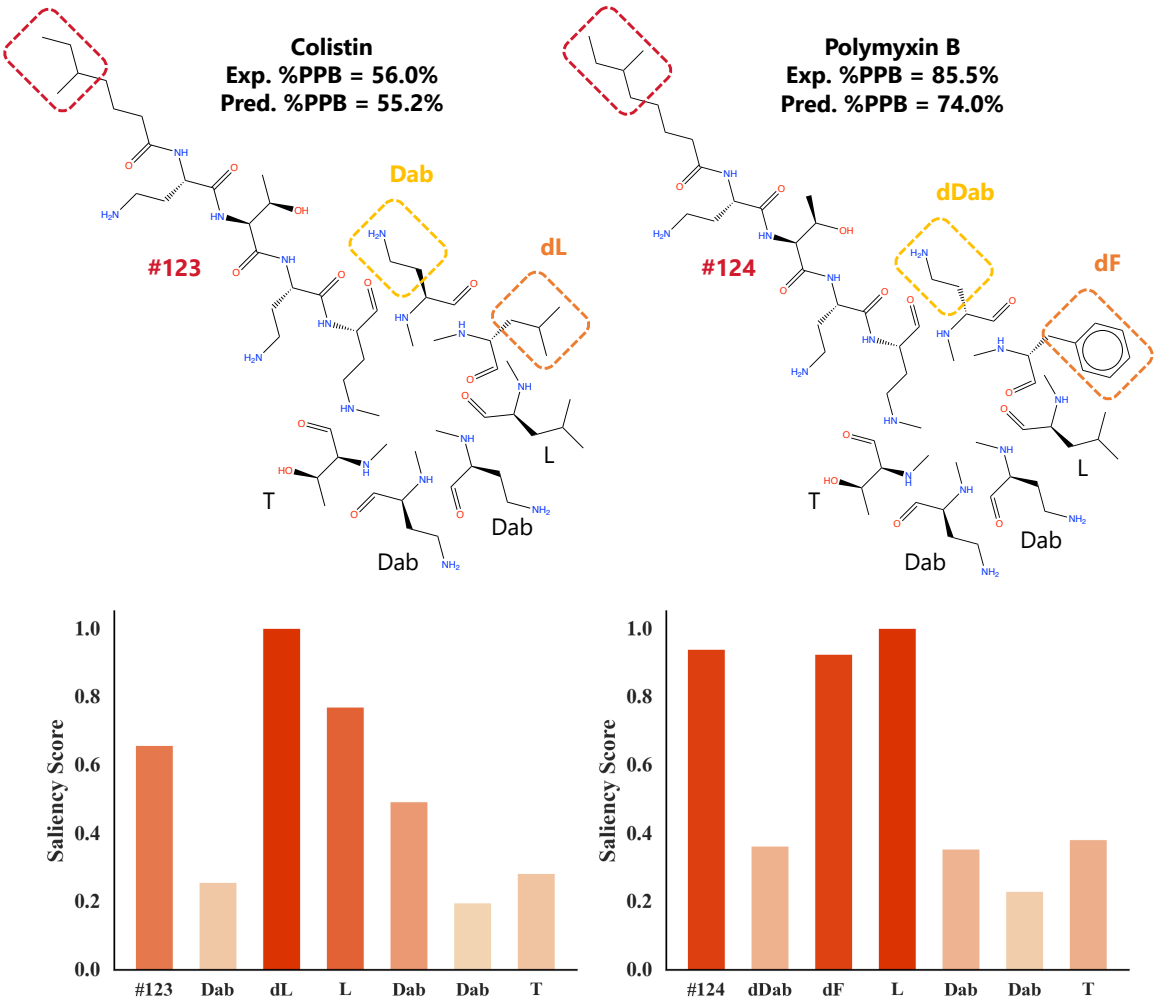


Figure 5.12: The monomers obtained by the decomposition procedure and these salience scores of colistin (Exp.%PPB = 56.0%, Pred.%PPB = 55.2%) and polymyxin b (Exp.%PPB = 85.5%, Pred.%PPB = 74.0%). The salience score shows the average value of the salience score of all augmentation replicas across three repeated runs (normalized to a maximum value of 1).

#125 and #126 attracted the most attention from the prediction model, and the model distinguished the structural change even though these two structures were not utilized in model training. The prediction values and analysis suggested that the proposed method with monomer descriptors was effective. It also revealed that the saliency score could detect important side chains.

Colistin and polymyxin b

Colistin and polymyxin b are peptide pairs that differ in three monomers (#123 and #124, Dab and dDab, and dL and dF) included in the DrugBank set with significantly different PPB rates (colistin: 56.0%; polymyxin b: 85.5%). Similar to the results in acetyl-daptomycin and daptomycin, two SVM models were unable to distinguish colistin and polymyxin b (predicted $\%PPB_{50-95}$ value of colistin: 70.8%, 74.0%; polymyxin b: 71.0%, 75.0%). The 1D-CNN monomer model was able to distinguish them (predicted $\%PPB_{50-95}$ value of colistin: 55.2%; polymyxin b: 74.0%) again. On the other hand, dL of colistin received the most attention from the model, while polymyxin b had high saliency scores on #124, dF, and L. All of these were monomers with relatively high lipophilicity. Little attention was paid to the difference in chirality between Dab and dDab. #123 already has a large hydrophobic side chain, and the difference between it and #124 was just one carbon atom. Thus, the high attention paid to dL and dF was probably the key to the model's successful distinction of colistin from polymyxin b.

5.4 PPB Rate Prediction of Cyclic Peptides by Docking Simulation

In this section, we describe the PPB rate prediction of cyclic peptides based on docking simulations. We performed docking simulations using Glide software (version 2019.1) [198] on cyclic peptides consisting of 6–12 residues in the PD and Tajimi datasets, except for the DrugBank set, which contains peptides with relatively large side chains, to attempt to predict the PPB rate of cyclic peptides by a structure-based method.

5.4.1 3D structures of HSAs and cyclic peptides

As we mentioned in Chapter 2, small molecule docking-based methods predict PPB rates using information such as docking scores obtained from docking simulations of plasma proteins (basically HSA) and ligand molecules. In PPB, induced fitting occurs when the conformation of the protein changes upon compound binding. Therefore, it is important to use multiple HSA structures to represent their flexibility. Chen *et al.* [110] calculated the RMSD (root mean square deviation) of each structure pair for a total of 62 HSA structures and classified them into four groups by K-means.

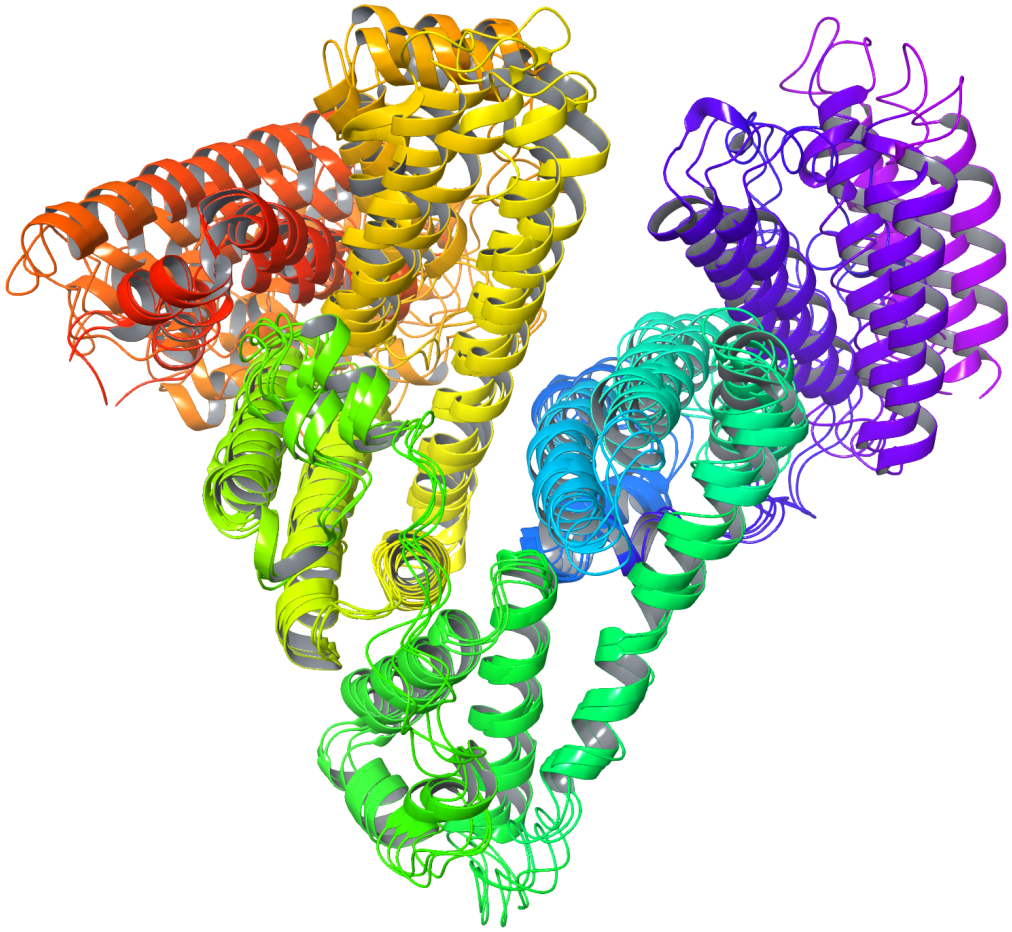


Figure 5.13: Results of the superposition of four different HSA structures used.

PDB IDs: 1E7A_A (2.2 Å), 1N5U (1.9 Å), 1O9X (3.2 Å), and 1UOR (2.8 Å) were used as representative structures for each group (the numbers in parentheses indicate the resolution). These representative structures are highly similar to all structures in their respective groups (RMSD < 2Å). In this study, we also obtained these four HSA structures from PDB and aligned the other three structures using Protein Preparation Wizard software based on 1N5U (1.9 Å), which has the highest resolution, to fill the missing side chain structures. Fig. 5.13 shows the result of the superposition of four HSA structures.

For cyclic peptides, we generated initial 3D structures with ionization states at

pH 7.0 ± 2.0 using LigPrep software (version 2019.1) [199] and then performed the conformational search using MacroModel software (version 2019.1) [200]. The enhanced mode torsional sampling method (Monte Carlo Molecular Mechanics, MCMM) (`Torsion sampling options:Enhanced`) was used for peptide conformational search. Considering the size of the peptide, the maximum number of searches was set to 15,000, an Energy Window of 100.0 kJ/mol was used, and the cutoff threshold for redundant conformations was set to 8.0 Å.

5.4.2 Overview of docking simulation

Site 1 to Site 6 of the HSA were used as binding pockets for docking. Site 1 and Site 2 are binding pockets for many small molecule compounds, and we used their coordinates reported by Lexa *et al* [107]. For Site 3–6, the average coordinates of ligands and fatty acids in four HSA structures aligned to 1N5U were used. The SP-Peptide mode of Glide was used for the docking simulation, and (15 Å)^3 for the internal box and (35 Å)^3 for the external box were used as the grid box size of the docking pocket. Fig. 5.14 shows the coordinates and grid boxes of Site 1 to Site 6 of 1N5U (green: internal box, purple: external box). Intel Xeon E5-2667 v4 CPUs (3.20 GHz, 8 cores) were used for the docking simulation calculations.

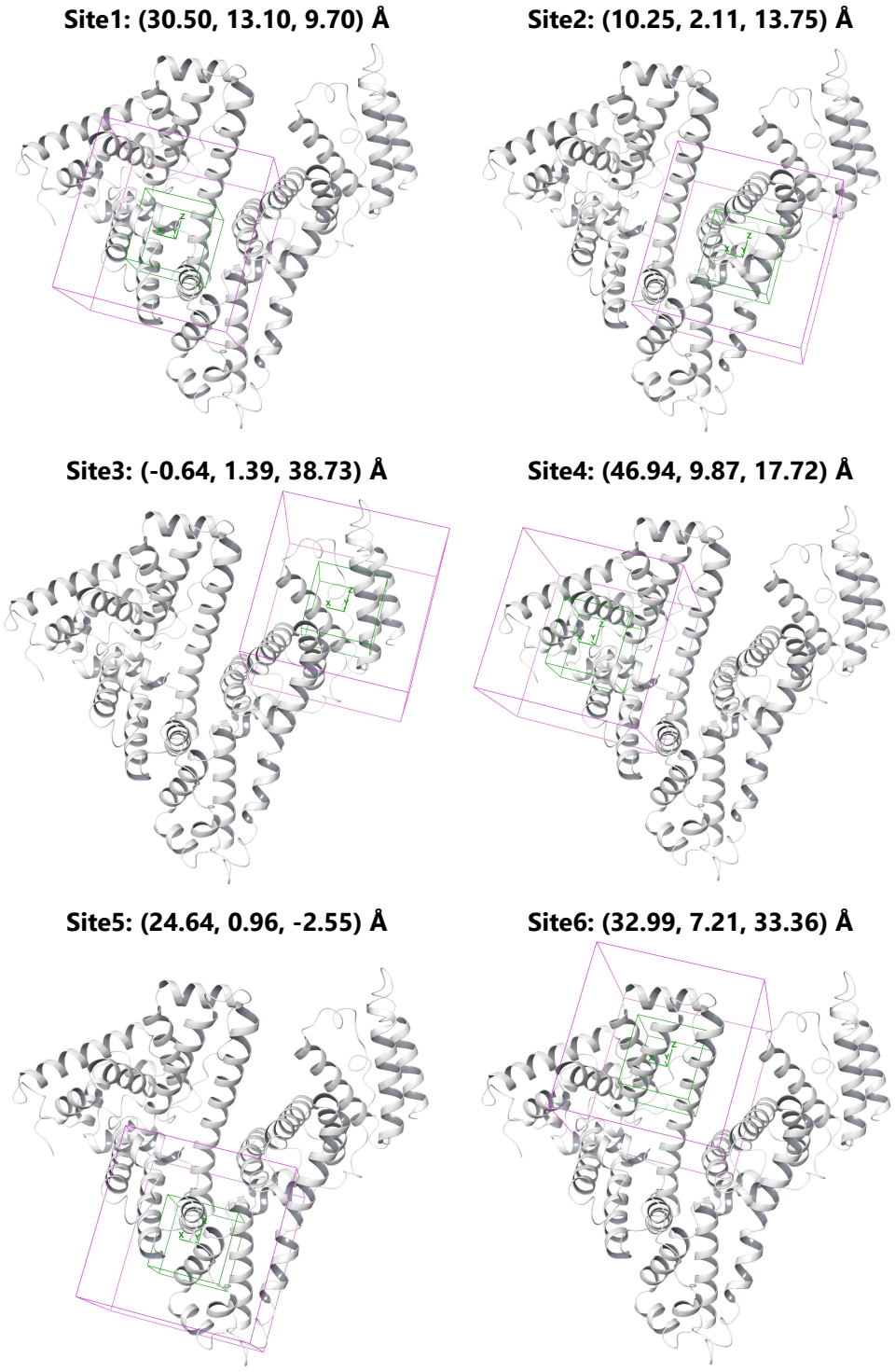


Figure 5.14: Coordinates and grid boxes of Site 1 to Site 6 of 1N5U (green: internal box, purple: external box).

Table 5.8: Number of binding poses obtained from each site docking of 1N5U for each number of residues (numbers in parentheses indicate the number of peptides).

Number of residues	Site 1	Site 2	Site 3	Site 4	Site 5	Site 6	Per peptide
6–9mer (131)	44,047	26,112	9,738	14,538	33,507	21,426	1,140
10mer (48)	15,559	1,849	46	144	799	525	394
11mer (32)	8,481	1,027	42	90	144	176	311
12mer (35)	3,721	147	4	20	49	67	115
All (246)	71,808	29,135	9,830	14,792	34,499	22,194	741

5.4.3 Results of docking simulation

The number of binding poses obtained from each site docking of 1N5U for each number of residues is shown in Table 5.8, and the total calculation time (CPU core time) for each site docking of 1N5U for each number of residues is shown in Table 5.9. As shown in Table 5.9, the average computation time per peptide for the 6–12mer 246 cyclic peptides in the PD and Tajimi datasets was 1 day 10 hours 2 minutes. The calculation time did not increase when the residue number of the cyclic peptide increased. On the other hand, as shown in Table 5.8, as the number of residues increased, the number of binding poses obtained decreased. For more than ten residues, the number of binding poses obtained from Site 3–6 was small (four to 799). Therefore, the results from Site 3 to Site 6 were not used, and only the results from the remaining two sites were used. In addition, there were two peptides for which no binding poses were obtained at Site 2 with ten residues, two peptides at Site 2 with 11 residues, two peptides at Site 1 with 12 residues, and 12 peptides at Site 2 with 12 residue. These peptides were not used, and the remaining 230 will be used for subsequent discussion.

Table 5.9: Total calculation time (CPU core time) for each site docking of 1N5U for each number of residues (numbers in parentheses indicate the number of peptides).

Number of residues	Site 1	Site 2	Site 3	Site 4	Site 5	Site 6	Total	Per peptide
6–9mer (131)	68d 9h 3min	17d 21h 46min	14d 23h 1min	16d 19h 11min	17d 3h 15min	32d 15h 10min	167d 19h 26min	1d 6h 45min
10mer (48)	32d 1h 19min	9d 19h 53min	10d 13h 50min	7d 3h 32min	6d 5h 4min	15d 4h 32min	81d 0h 10min	1d 16h 30min
11mer (32)	19d 1h 37min	6d 11h 4min	6d 12h 16min	4d 9h 37min	3d 21h 21min	8d 7h 42min	48d 15h 37min	1d 12h 29min
12mer (35)	19d 4h 40min	6d 21h 18min	7d 13h 44min	5d 2h 46min	4d 22h 54min	7d 13h 53min	51d 7h 15min	1d 11h 11min
All (246)	138d 16h 39min	41d 2h 1min	39d 14h 51min	33d 11h 6min	32d 4h 34min	63d 17h 17min	348d 18h 28min	1d 10h 2min

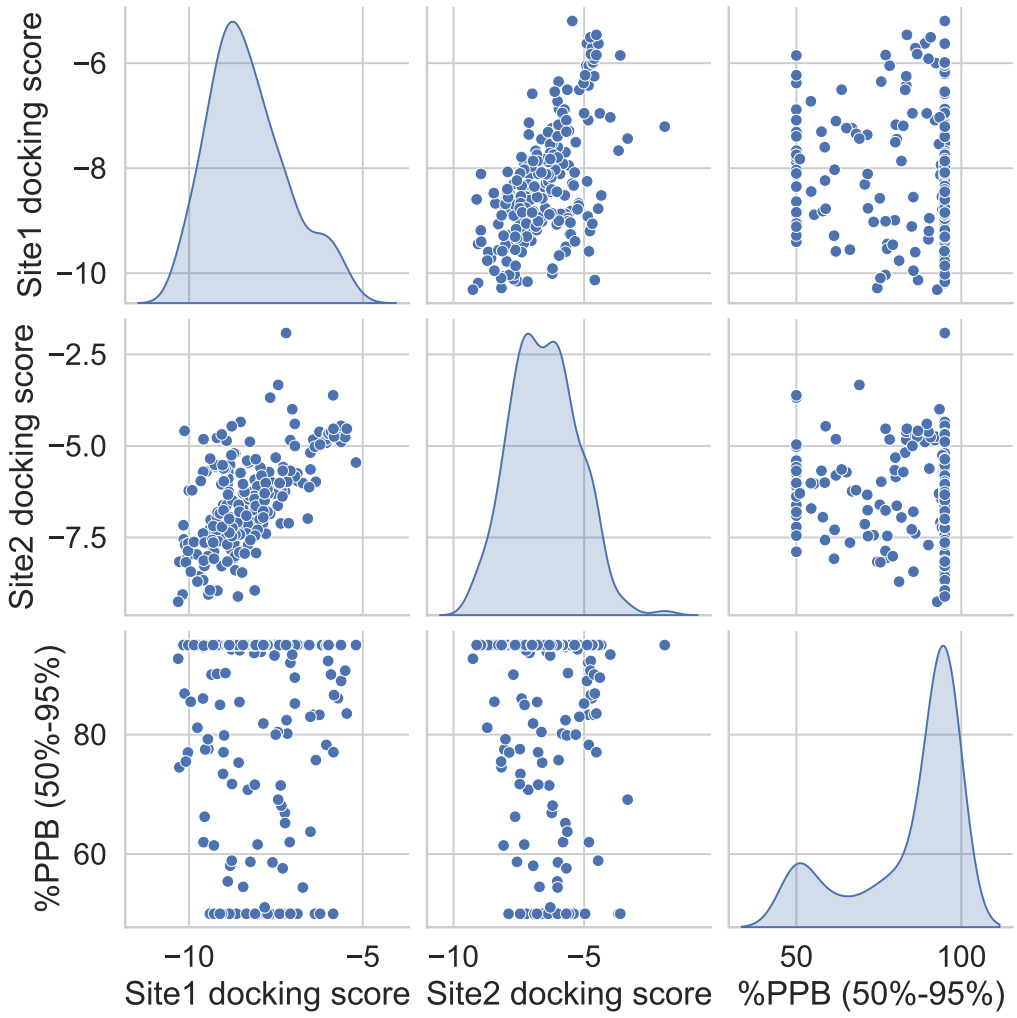


Figure 5.15: Distribution of Site 1, Site 2 docking scores (average of maximum five poses), and objective variables $\%PPB_{50-95}$.

From the docking results, the top five docking scores for each peptide were used (if the number of docking poses was less than five, all were used). The distribution of the mean docking scores of the docking poses of Site 1, Site 2, and objective variable $\%PPB_{50-95}$ is shown in Fig. 5.15. Although the docking scores of Site 1 and Site 2 showed some correlation (R = 0.61), the correlation coefficient between Site 1 docking score and $\%PPB_{50-95}$ was −0.16, and that between Site 2 docking score and $\%PPB_{50-95}$ was −0.15. Overall, predicting PPB rates for cyclic peptides based solely on docking scores derived from traditional docking software for small molecules has not been successful.

Table 5.10: Summary of used 12 glide descriptors.

Descriptor	Description
Glide gscore	GlideScore
Glide lipo	Lipophilic contact plus phobic attractive term in the GlideScore
Glide hbond	Hydrogen-bonding term in the GlideScore
Glide metal	Metal-binding term in the GlideScore
Glide rewards	Various reward or penalty terms
Glide evdw	Van der Waals energy
Glide ecoul	Coulomb energy
Glide erotb	Penalty for freezing rotatable bonds in the GlideScore
Glide esite	Term in the GlideScore for polar interactions in the active site
Glide emodel	Model energy, Emodel
Glide energy	Modified Coulomb-van der Waals interaction energy
Glide einternal	Internal torsional energy

Table 5.11: Performance comparison between three baseline methods and two glide models using the glide test set. The metrics of three baseline methods are calculated from the averaged prediction values of three repeated runs; the best result for each metric is indicated in bold.

Metrics	RF-FP	SVM-2D	SVM-2D3D	RF-Glide	SVM-Glide
MAE (%)	9.35	**4.82**	7.25	12.39	10.08
MSE	173.9	**67.8**	112.9	238.8	202.5
R	0.859	**0.926**	0.889	0.713	0.809
R^2	0.610	**0.848**	0.746	0.464	0.545

5.4.4 Prediction based on glide descriptors

To incorporate more docking information than the docking score, we calculated 12 types of glide descriptors (Table 5.10), including the van der Waals energy and Coulomb energy of the binding poses that can be calculated from glide. We used the average of the glide descriptors for the top five binding poses for each peptide (if the number of binding poses was less than five, all were used). There were 12 different glide descriptors for each of Site 1 and Site 2, and for each of these glide descriptors, the mean, maximum, and minimum values were taken. Finally, $3 \times 12 = 36$ representative values of glide descriptors were used. Three glide descriptors with 0 values across all peptides were deleted, and Z-score standardization was performed for the remaining 33 glide descriptors.

For 230 peptides with docking poses of Site 1 and Site 2, 20 peptides are included in the original test dataset (36 peptides). These were used as the glide test set, and the remaining 210 data were used as the glide training set to build an RF model and an SVM model. The hyperparameters of the two models were determined through 10-fold cross-validation of the training data within the same search range shown in Table 5.5 (RF: max_depth = 10, n_estimators = 750; SVM: C = 32, $\gamma = 2^{-6}$). The prediction accuracy for the glide test set is shown in Table 5.11. The prediction accuracy was significantly improved compared to prediction based on docking scores alone, and the SVM-Glide model (MAE = 10.08%, R = 0.809) was superior to the RF-Glide model (MAE = 12.39%, R = 0.713). However, there was still a large difference between the predictive capabilities of these two glide-based models and two peptide descriptor-based SVM models (SVM-2D: MAE = 4.82%, R = 0.926; SVM-2D3D: MAE = 7.25%, R = 0.889). Overall, by using glide descriptors, which contain more information than the docking score, it was possible to predict the PPB rate of cyclic peptides to some extent. However, compared with the ligand-based method, the structure-based method using docking simulation is much more computationally expensive, and prediction is difficult based on docking software for small molecules.

5.5 Summary

In this chapter, we introduced CycPeptPPB, a prediction model for the PPB rate of cyclic peptides. By leveraging experimental data from collaborations with pharmaceutical companies and published literature, we applied multi-level molecular features and data augmentation techniques. The model demonstrated superior prediction accuracy, particularly with monomer-level features, significantly outperforming conventional methods. The fusion model combining atom, monomer, and peptide level features achieved the best performance on the internal test set, while the monomer model with 1D-CNN layers showed strong generalization on the external DrugBank dataset. Furthermore, we also introduced the concept of saliency scores to analyze the contribution of each monomer to the PPB rate prediction of cyclic peptides. We confirmed that our CycPeptPPB (1D-CNN monomer) model accurately identified key monomers that significantly influence PPB rates. The ability to identify such important monomers has significant implications for drug development, as it enables researchers to pinpoint the structural features that enhance or reduce PPB rates. Additionally, we attempted to predict PPB rates using docking simulations, but results indicated that ligand-based methods were more effective for predicting cyclic peptide PPB rates.

Future studies should focus on expanding the dataset by incorporating more cyclic peptides with diverse structures, particularly those with larger side chains, which would improve the model's generalization capability.

Chapter 6

Overall Discussion

In this chapter, we provide a common discussion on applying the proposed multi-level feature design method to predict membrane permeability and PPB rate.

6.1 Descriptor Selection

As mentioned in Section 3.3.3, after calculating the descriptors, we performed pre-processing to remove redundant peptide descriptors and selected 2D and 3D peptide descriptors by two RF models, respectively. However, there are several variations in descriptor selection methods. Hence, we explored different selection methods to discuss how they could affect the model's performance.

6.1.1 Common important descriptors in membrane permeability and PPB rate prediction

In predicting membrane permeability and PPB rate of cyclic peptides, several peptide descriptors have been identified as important for both properties, highlighting their shared relevance in describing the chemical and physical features of cyclic peptides.

The only descriptor selected for both membrane permeability and PPB rate prediction was logP(o/w) (the logarithm of the octanol/water partition coefficient), a 2D descriptor calculated by MOE software (Table 6.1). LogP measures a molecule's lipophilicity, and as we mentioned in Section 2.2.2 and Section 2.3.1, it is generally considered the most important descriptor for predicting both permeability and PPB rate. Typically, peptides with appropriate lipophilicity are more likely to traverse the lipid bilayer of cell membranes [15] and have stronger hydrophobic interactions with

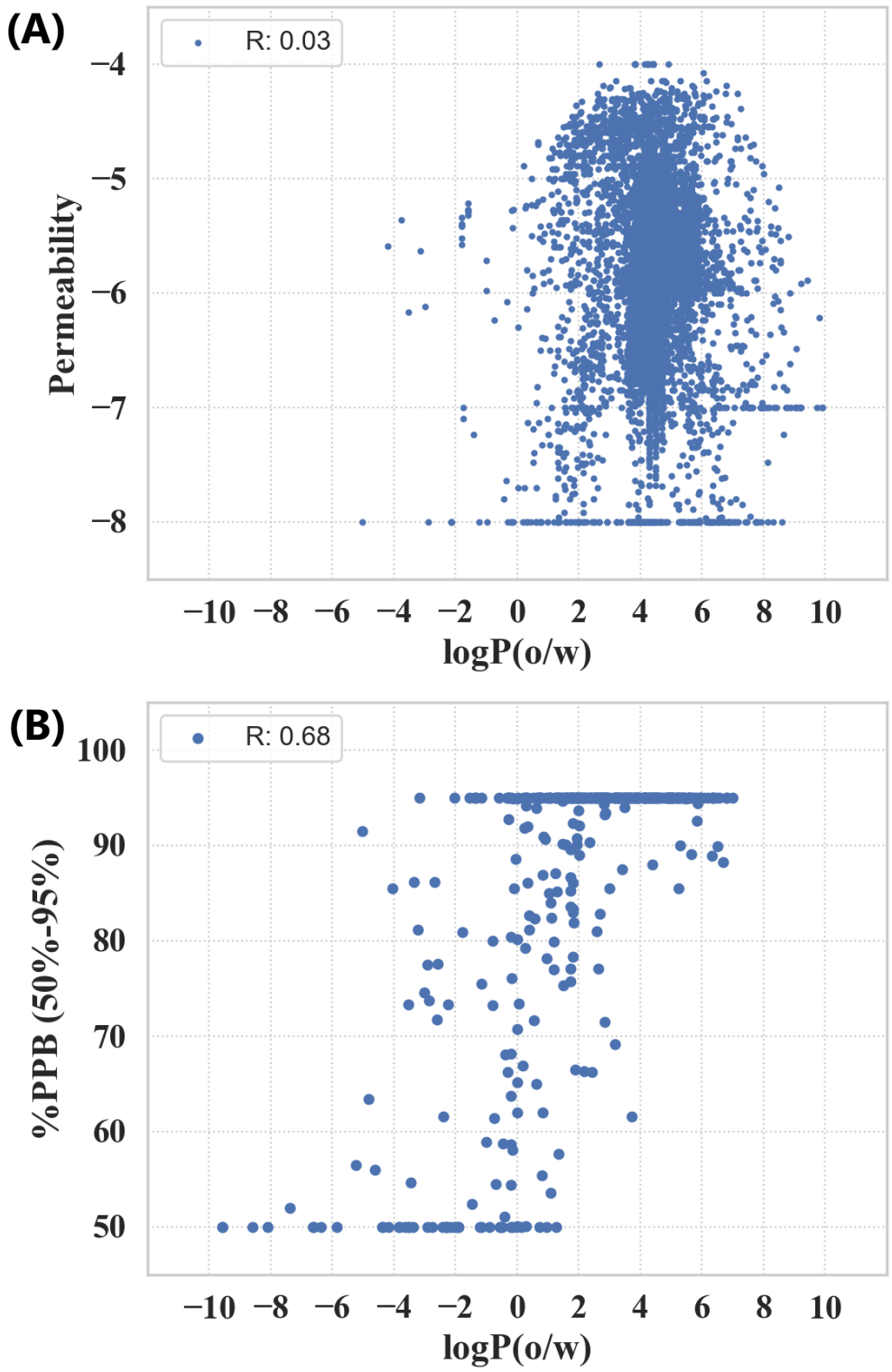


Figure 6.1: Distribution of logP(o/w) with (A) permeability LogP_{exp} and (B) PPB rate $\%\text{PPB}_{50-95}$.

Table 6.1: Selected descriptors and their RF feature importance for permeability and PPB rate prediction.

Type	Permeability descriptors	Feature importance	PPB descriptors	Feature importance
2D	VSA_EState9	0.184	logP(o/w)	0.158
	density	0.097	AATS4se	0.116
	MolLogP	0.085	PEOE_VSA-1	0.109
	fr_Al_OH	0.067	-	-
	logP(o/w)	0.023	-	-
	lip_violation	0.014	-	-
	h_logD	0.014	-	-
3D	dens	0.090	vsurf_CW2	0.198
	FNSA4	0.069	vsurf_CW3	0.107
	RNCS	0.048	-	-
	FASA-	0.042	-	-
	FCASA+	0.042	-	-
	FASA_P	0.041	-	-
	FNSA2	0.032	-	-
	FNSA5	0.024	-	-
	vsurf_Wp2	0.022	-	-

binding sites on HSA [4]. As shown in Fig. 6.1, although no correlation was found between logP(o/w) and membrane permeability (R = 0.03), all peptides with very high permeability ($LogP_{exp} \geq -5$) had logP(o/w) in the range from approximately 0 to 8. In particular, a parabolic relationship between logP(o/w) and membrane permeability is observed in data from several papers (Fig. 6.2). On the other hand, some correlation was observed between logP(o/w) and PPB rates (R = 0.68).

For permeability prediction, descriptors related to the van der Waals surface area showed the highest feature importance, such as VSA_EState9 (van der Waals surface area using EState indices and surface area contribution), density (molecular weight divided by approximated van der Waals volume, provides a rough estimated density based on atomic connectivity), and dens (molecular weight divided by 3D van der Waals volume, accounts for the actual 3D molecular shape and conformation). These descriptors are essential for capturing the overall molecular shape of the cyclic peptide during membrane permeation. Similarly, in the case of PPB rate prediction, PEOE_VSA-1 (sum of van der Waals surface areas for atoms within a certain range of partial charges), another van der Waals surface area-related descriptor, emerged as an

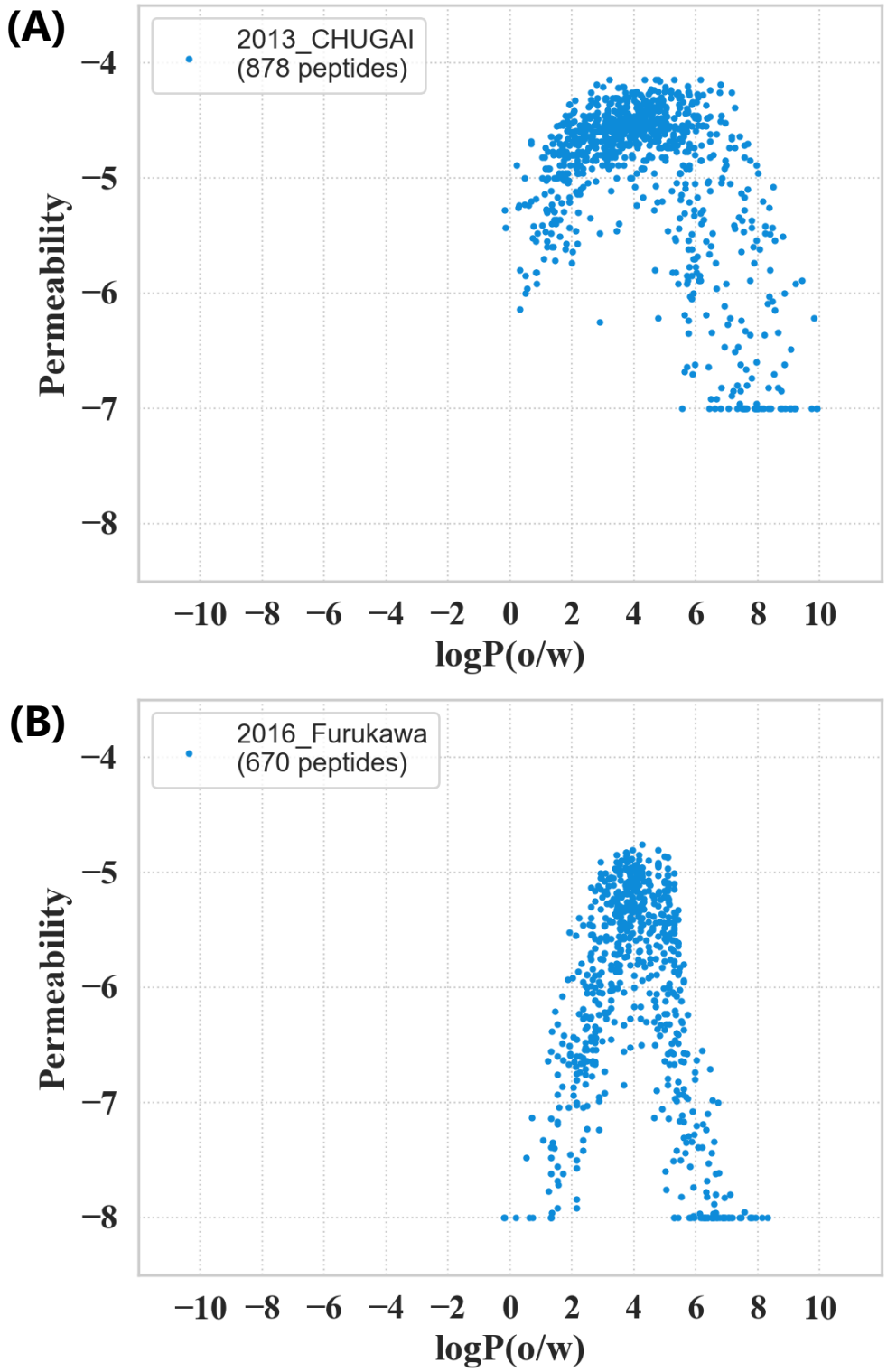


Figure 6.2: Distribution of logP(o/w) with permeability LogP_{exp} on the data from (A) 2013_CHUGAI [14] and (B) 2016_Furukawa [15].

important descriptor.

6.1.2 Integrated selection of 2D and 3D descriptors

In this section, we explore the impact of selecting 2D and 3D descriptors together using a single RF model for feature importance evaluation. Previously, we selected the descriptors separately using two independent RF models, one for 2D and the other for 3D descriptors. However, combining 2D and 3D descriptors in a single model may yield different results by allowing the model to consider the combined effects of both types of descriptors. As shown in Fig. 6.3, 3D descriptors tend to be less important than 2D descriptors in both membrane permeability and PPB rate prediction. However, interestingly, in the case of membrane permeability prediction, models using both 2D and 3D descriptors showed slightly improved accuracy compared to those using 2D descriptors alone (Table 4.10, Fig. 4.14). This observation highlights that the feature importance scores in a combined model do not necessarily directly correlate with the contribution to prediction accuracy. For both tasks, the dominant 2D descriptors remained unchanged when selected together (permeability: VSA_EState9, density, MolLogP, and fr_Al_OH; PPB: logP(o/w), AATS4se, and PEOE_VSA-1). In the case of permeability prediction, 3D descriptors played a relatively minor role, with six of the top 16 descriptors being 3D descriptors. On the other hand, for PPB rate prediction, 3D descriptors were even less influential, with only two 3D descriptors in the top 16 descriptors. This observation suggests that 3D information might be more crucial for predicting membrane permeability, where molecular conformation has a significant influence, compared to predicting PPB rates, where 2D molecular structure descriptors seem to dominate the prediction task.

To further assess the effectiveness of selecting 2D and 3D descriptors together, we constructed an SVM-mix model, which uses the newly selected descriptors. The number of selected descriptors was kept the same as the number of descriptors chosen in the separate models (membrane permeability: all 16 types shown in Fig. 6.3 (A); PPB rate: top five types shown in Fig. 6.3 (B), all were 2D descriptors) The hyperparameters were determined by a grid search (search range was the same as Table 4.7). The performance comparison between SVM-2D, SVM-2D3D, and SVM-mix models is shown in Table 6.2. For permeability prediction, the SVM-2D3D model consistently performed the best (MAE = 0.418, R = 0.834). Interestingly, the SVM-mix model (MAE = 0.504, R = 0.766) performed even worse than the SVM-2D model (MAE = 0.488, R = 0.781), indicating that combining 2D and 3D descriptors in this way did not enhance the

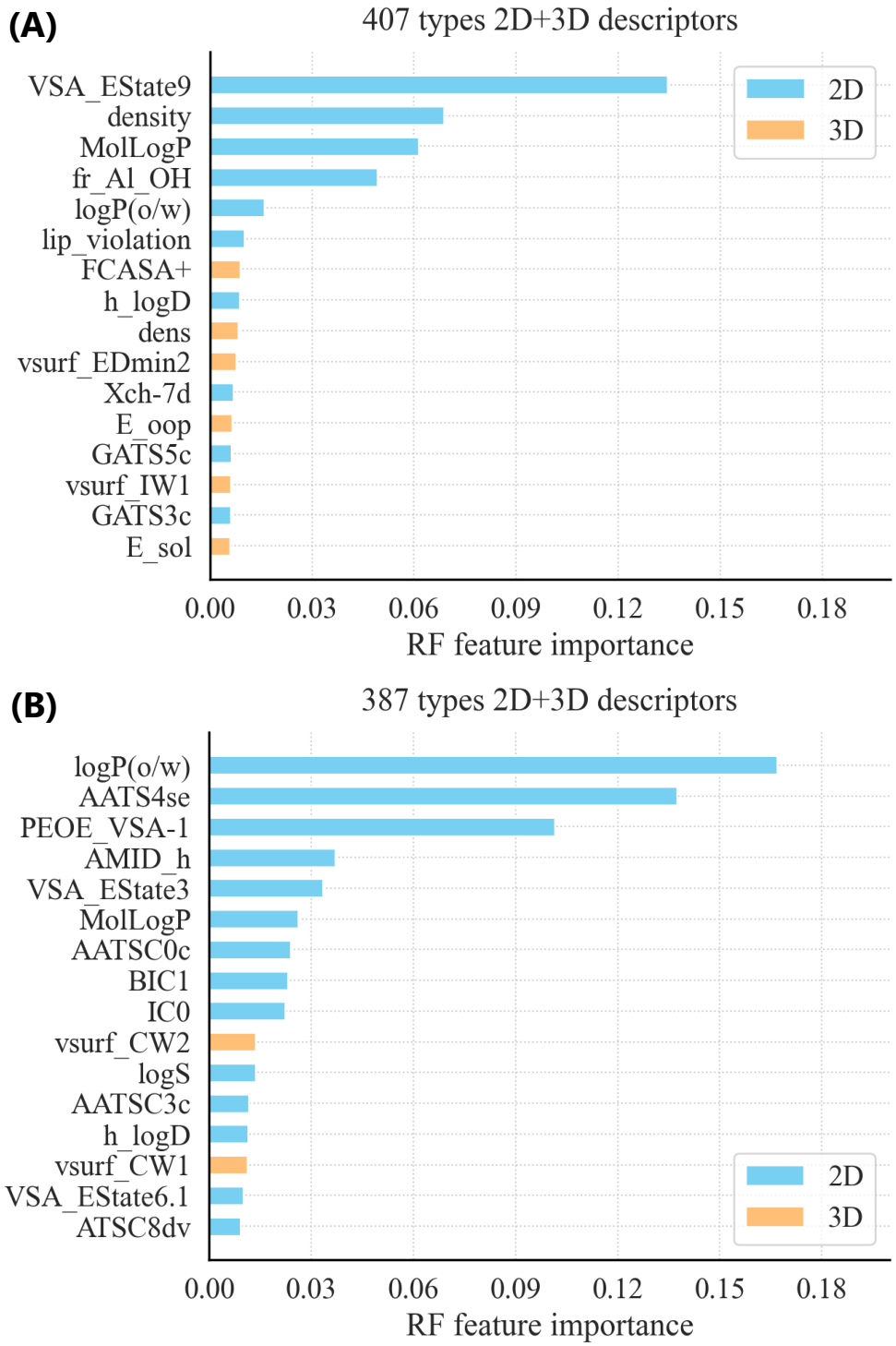


Figure 6.3: Top 16 peptide descriptors with the highest RF feature importance for (A) permeability prediction and (B) PPB rate prediction. The 2D and 3D descriptors are shown in light blue and orange, respectively.

Table 6.2: Performance comparison between three SVM models for permeability (using the test set) and PPB rate prediction (using the test set and DrugBank set). The metrics are the averaged values of three repeated runs; the best result for each metric is indicated in bold.

Task	Metrics	SVM-2D	SVM-2D3D	SVM-mix
Permeability (test set)	MAE	0.488 ± 0.005	$\mathbf{0.418 \pm 0.001}$	0.504 ± 0.001
	R	0.781 ± 0.007	$\mathbf{0.834 \pm 0.001}$	0.766 ± 0.003
Task	Metrics	SVM-2D	SVM-2D3D	SVM-mix
PPB (test set)	MAE (%)	$\mathbf{4.41 \pm 0.08}$	6.23 ± 0.13	6.94 ± 0.27
	R	$\mathbf{0.902 \pm 0.004}$	0.887 ± 0.001	0.884 ± 0.014
Task	Metrics	SVM-2D	SVM-2D3D	SVM-mix
PPB (DrugBank set)	MAE (%)	13.36 ± 0.07	$\mathbf{10.89 \pm 0.16}$	12.78 ± 0.41
	R	0.467 ± 0.005	0.581 ± 0.008	$\mathbf{0.637 \pm 0.007}$

predictive performance. When 2D and 3D descriptors are selected together, the 3D descriptors act as auxiliary features to the dominant 2D descriptors, which can lead to overfitting of the training data. This overfitting may limit the model's ability to capture the structural diversity in the test data effectively. This result supports our previous approach of selecting 2D and 3D descriptors separately as the optimal method for permeability prediction. In the case of PPB rate prediction using the test set, the SVM-2D model still achieved the best performance (MAE = 4.41%, R = 0.902). The inclusion of unnecessary 2D descriptors in the SVM-mix model actually led to a decrease in accuracy (MAE = 6.94%, R = 0.884), demonstrating that the simplicity of fewer descriptors might offer better generalization for this task. For the PPB rate prediction on the DrugBank set, although the SVM-mix model (MAE = 12.78%, R = 0.637) improved the accuracy compared to the SVM-2D model (MAE = 13.36%, R = 0.467), its MAE did not fall below 10%.

Algorithm 6.2 Details of the Bolasso algorithm for descriptor selection

Input:

Peptide descriptors: $\boldsymbol{X}$
Objective variable: $\boldsymbol{Y}$
Number of cross validations: k
Number of bootstrap replicates: m
Appearance threshold: t ▷ k was set to 3, m was set to 20, and t was set to 20

Split data ($\boldsymbol{X}$, $\boldsymbol{Y}$) for k-fold cross validation
for $i = 1$ to k **do**
    Select training data ($\boldsymbol{X}_i$, $\boldsymbol{Y}_i$)
    for $j = 1$ to m **do**
        Choose a random n from the range of 10%–90% of the training data size ▷ n is the size of one replica dataset
        Allow duplicates and randomly select n data ($\boldsymbol{X}_{ij}$, $\boldsymbol{Y}_{ij}$) from ($\boldsymbol{X}_i$, $\boldsymbol{Y}_i$)
        Perform Lasso estimate $\hat{\boldsymbol{w}}_{ij}$ for data ($\boldsymbol{X}_{ij}$, $\boldsymbol{Y}_{ij}$)
        Calculate signal $\boldsymbol{S}_{ij} = \{s, \hat{w}_{ij}^s \neq 0\}$
    end for
end for
Compute sum $\boldsymbol{S} = \sum_i^k \sum_j^m \boldsymbol{S}_{ij}$
Choose signal above threshold $\boldsymbol{S} = \{s, s \geq t\}$
Select $\hat{\boldsymbol{w}}^S$ from ($\boldsymbol{X}^S$, $\boldsymbol{Y}$)

6.1.3 Alternative algorithms for descriptor selection

While Random Forest (RF) was used in our study for final descriptor selection, various other algorithms can also be applied to select important molecular descriptors. One widely used algorithm is Lasso (Least Absolute Shrinkage and Selection Operator) [201], which is particularly known for enforcing sparsity in the model by reducing the coefficients of less important features to zero. Various improvements and extensions of Lasso have been reported [202, 203]. Bolasso [204] is an enhanced method that integrates bootstrap techniques into Lasso for feature selection, providing a more robust way to extract high-importance features. In this section, we also performed descriptor selection using Bolasso (Algorithm 6.2) as an alternative to RF and constructed SVM models based on Bolasso-selected descriptors.

Table 6.3: Descriptor selection results for permeability prediction by Bolasso when changing the Lasso hyperparameter α. Selected 2D and 3D descriptors and their frequency (up to $k \times m = 60$ times) are shown, respectively. Finally, we used six 2D descriptors (MATS2i, Xch-7d, VSA_EState9, fr_piperdine, AATSC4c, and BCUT_PEOE_0) and seven 3D descriptors (FASA-, FNSA4, vsurf_Wp4 FASA+, FNSA3, vsurf_R, and dens).

	$\alpha = 0.09$		$\alpha = 0.10$		$\alpha = 0.11$		$\alpha = 0.12$		$\alpha = 0.13$		$\alpha = 0.14$	
2D	MATS2i	60	MATS2i	60	MATS2i	60	MATS2i	60	MATS2i	60	MATS2i	60
	Xch-7d	54	Xch-7d	54	Xch-7d	55	Xch-7d	56	Xch-7d	56	Xch-7d	54
	fr_piperdine	50	fr_piperdine	48	fr_piperdine	46	fr_piperdine	42	VSA_EState9	45	VSA_EState9	47
	n10FRing	45	n10FRing	40	BCUT_PEOE_0	37	VSA_EState9	42	fr_piperdine	39	fr_piperdine	32
	opr_violation	42	BCUT_PEOE_0	39	n10FRing	37	AATSC4c	33	AATSC4c	31	AATSC4c	24
	BCUT_PEOE_0	40	AATSC4c	36	VSA_EState9	37	BCUT_PEOE_0	31	BCUT_PEOE_0	27	BCUT_PEOE_0	22
	AATSC4c	39	GATS3c	35	AATSC4c	34	n10FRing	30	n10FRing	24	-	-
	GATS3c	36	AATSC4s	34	GATS3c	32	GATS3c	27	density	22	-	-
	AATSC4s	35	VSA_EState9	33	AATSC4s	29	AATSC4s	25	GATS3c	22	-	-
	density	33	density	30	density	27	density	23	-	-	-	-
	$\alpha = 0.03$		$\alpha = 0.04$		$\alpha = 0.05$		$\alpha = 0.06$		$\alpha = 0.07$		$\alpha = 0.08$	
3D	dens	60	dens	60	FASA-	60	FASA-	60	FASA-	60	FASA-	60
	FASA-	60	FASA-	60	FNSA4	58	FNSA4	59	FNSA4	59	FNSA4	59
	vsurf_R	58	FNSA4	57	dens	57	dens	51	vsurf_Wp4	50	vsurf_Wp4	46
	FASA+	56	vsurf_R	56	FASA+	52	FASA+	50	FASA+	46	FASA+	36
	npr2	56	FASA+	55	vsurf_R	52	vsurf_Wp4	49	dens	37	FNSA3	34
	FNSA4	56	npr2	51	vsurf_Wp4	47	vsurf_R	44	FNSA3	37	vsurf_R	21
	RNCS	54	vsurf_ID7	51	npr2	45	FNSA3	36	vsurf_R	30	dens	20
	vsurf_ID7	53	RNCS	51	vsurf_ID7	43	vsurf_ID7	36	RNCS	25	-	-
	E_vdw	49	vsurf_Wp4	48	RNCS	43	npr2	35	npr2	24	-	-
	FPSA3	47	FPSA3	38	FNSA3	35	RNCS	34	vsurf_ID7	23	-	-

Table 6.4: Descriptor selection results for PPB rate prediction by Bolasso when changing the Lasso hyperparameter α. Selected 2D and 3D descriptors and their frequency (up to $k \times m = 60$ times) are shown, respectively. Finally, we used four 2D descriptors (logP(o/w), PEOE_VSA-1, EState_VSA3.1, and AATSC4se) and three 3D descriptors (vsurf_D5, vsurf_CW3, and vsurf_CW1).

	$\alpha = 1.0$		$\alpha = 1.2$		$\alpha = 1.4$		$\alpha = 1.6$		$\alpha = 1.8$		$\alpha = 2.0$	
2D	EState_VSA3.1	44	EState_VSA3.1	45	EState_VSA3.1	45	EState_VSA3.1	43	logP(o/w)	41	logP(o/w)	42
	logP(o/w)	37	logP(o/w)	37	logP(o/w)	38	logP(o/w)	42	EState_VSA3.1	38	PEOE_VSA-1	39
	PEOE_VSA-1	33	PEOE_VSA-1	34	PEOE_VSA-1	36	PEOE_VSA-1	37	PEOE_VSA-1	37	EState_VSA3.1	36
	AATSC4se	31	AATSC4se	30	AATSC4se	29	AATSC4se	24	AATSC4se	22	AATSC4se	23
	EState_VSA4.1	23	BCUT_SLOGP_3	22	BCUT_SLOGP_3	20	BCUT_SLOGP_3	20	BCUT_SLOGP_3	20	-	-
	BCUT_SLOGP_3	23	h_pKb	20	-	-	-	-	-	-	-	-
	h_pKb	23	-	-	-	-	-	-	-	-	-	-
	-	-	-	-	-	-	-	-	-	-	-	-
	-	-	-	-	-	-	-	-	-	-	-	-
	-	-	-	-	-	-	-	-	-	-	-	-

	$\alpha = 2.0$		$\alpha = 2.5$		$\alpha = 3.0$		$\alpha = 3.5$		$\alpha = 4.0$		$\alpha = 4.5$	
3D	vsurf_D5	49	vsurf_D5	48	vsurf_D5	47	vsurf_D5	47	vsurf_D5	47	vsurf_D5	44
	vsurf_CW3	45	vsurf_CW3	43	vsurf_CW3	39	vsurf_CW3	35	vsurf_CW3	32	vsurf_CW3	30
	vsurf_CW1	32	vsurf_CW1	35	vsurf_CW1	33	vsurf_CW1	29	vsurf_CW1	29	vsurf_CW1	27
	dens	26	vsurf_HL1	25	dens	22	dens	22	dens	21	-	-
	vsurf_HL1	24	dens	24	vsurf_HL1	21	-	-	-	-	-	-
	FASA+	23	vsurf_CW2	21	-	-	-	-	-	-	-	-
	vsurf_CW2	23	-	-	-	-	-	-	-	-	-	-
	-	-	-	-	-	-	-	-	-	-	-	-
	-	-	-	-	-	-	-	-	-	-	-	-
	-	-	-	-	-	-	-	-	-	-	-	-

Descriptor selection results for permeability and PPB rate prediction by Bolasso are shown in Table 6.3 and Table 6.4, respectively. For permeability prediction, the 2D descriptors with a high appearance frequency of Bolasso changed significantly from the results selected by RF, with only VSA_EState9 remaining in the top positions, and LogP-related descriptors (MolLogP, logP(o/w), and h_logD) disappeared. For 3D descriptors, FASA-, FNSA4, and dens remained in the top positions. Finally, we used six 2D descriptors (MATS2i, Xch-7d, VSA_EState9, fr_piperdine, AATSC4c, and BCUT_PEOE_0) and seven 3D descriptors (FASA-, FNSA4, vsurf_Wp4 FASA+, FNSA3, vsurf_R, and dens). For PPB rate prediction, the 2D descriptor selection results did not change much, with logP(o/w), PEOE_VSA-1, and AATSC4se remaining, and in addition, EState_VSA3.1 now appears at the top positions. Finally, we used four 2D descriptors (logP(o/w), PEOE_VSA-1, EState_VSA3.1, and AATSC4se) and three 3D descriptors (vsurf_D5, vsurf_CW3, and vsurf_CW1). These descriptors were used to build the SVM-Bo2D (only using 2D descriptors) and SVM-Bo2D3D (using both 2D and 3D descriptors) models. The hyperparameters were determined by a grid search (search range was the same as Table 4.7).

The performance comparison between SVM-2D, SVM-2D3D, SVM-Bo2D, and SVM-Bo2D3D models is shown in Table 6.5. For permeability prediction, the descriptors selected by Bolasso differed significantly from those selected by RF, particularly with the exclusion of LogP-related descriptors. As a result, both the SVM-Bo2D (MAE = 0.577, R = 0.668) and SVM-Bo2D3D (MAE = 0.595, R = 0.680) models demonstrated a noticeable decrease in prediction accuracy. This indicates that for permeability prediction, the descriptors selected by RF were more effective in capturing the molecular properties that affect permeability. On the other hand, for PPB rate prediction, both the test set and DrugBank set results showed a slightly improved prediction accuracy using Bolasso-selected descriptors. Notably, the SVM-Bo2D3D model achieved the lowest %MAE (test set: MAE = 4.22%; DrugBank set: MAE = 9.11%), suggesting that in the case of PPB rate prediction, the combination of 2D and 3D descriptors selected by Bolasso could more effectively capture the critical factors influencing PPB rates.

Overall, different feature selection algorithms can lead to significantly different results. Therefore, it is crucial to consider domain knowledge when interpreting the output of these algorithms and determining which descriptors to use. For example, LogP-related features are essential as they highly influence membrane permeability and PPB rate.

Table 6.5: Performance comparison between four SVM models for permeability (using the test set) and PPB rate prediction (using the test set and DrugBank set). The metrics are the averaged values of three repeated runs; the best result for each metric is indicated in bold.

Task	Metrics	SVM-2D	SVM-2D3D	SVM-Bo2D	SVM-Bo2D3D
Permeability (test set)	MAE	0.488 ± 0.005	$\mathbf{0.418 \pm 0.001}$	0.577 ± 0.006	0.595 ± 0.003
	R	0.781 ± 0.007	$\mathbf{0.834 \pm 0.001}$	0.668 ± 0.004	0.680 ± 0.008
Task	Metrics	SVM-2D	SVM-2D3D	SVM-Bo2D	SVM-Bo2D3D
PPB (test set)	MAE (%)	4.41 ± 0.08	6.23 ± 0.13	4.28 ± 0.07	$\mathbf{4.22 \pm 0.03}$
	R	0.902 ± 0.004	0.887 ± 0.001	0.899 ± 0.002	$\mathbf{0.911 \pm 0.002}$
Task	Metrics	SVM-2D	SVM-2D3D	SVM-Bo2D	SVM-Bo2D3D
PPB (DrugBank set)	MAE (%)	13.36 ± 0.07	10.89 ± 0.16	10.07 ± 0.07	$\mathbf{9.11 \pm 0.10}$
	R	0.467 ± 0.005	0.581 ± 0.008	$\mathbf{0.686 \pm 0.004}$	0.675 ± 0.005

6.2 Conformation Generation of Cyclic Peptides by MD Simulation

Considering the computational cost, we utilized RDKit to generate relatively simple low-energy cyclic peptide conformations (approximately three hours per peptide). However, based on the results from the permeability prediction (Section 4.4.1 and 4.4.6), incorporating 3D structural information significantly enhanced the prediction performance. Performing MD simulations to generate more accurate conformations could further improve prediction accuracy. Therefore, we developed a relatively simple MD simulation protocol that allows the search for sufficiently diverse conformations and generates conformations of cyclic peptides. These conformations were then compared with the conformations generated by RDKit.

6.2.1 Simulation details

Closed conformations, expected to maximize intramolecular interactions, are critical for the membrane permeability of cyclic peptides [56, 57, 58]. If stable and diverse conformations can be reliably explored, vacuum simulations are suitable for generating closed conformations. Therefore, given the computational costs of MD simulations in explicit water or lipid environments, we employed a more straightforward approach by performing MD calculations in a vacuum. The simulations used Amber software (version 20) [205], setting the relative permittivity to 4.8 (chloroform) to mimic a lipid-like environment. The peptides were parameterized using the Amber10 force field: Extended Huckel Theory (EHT) parameter set in MOE software [125]. To ensure adequate sampling, we employed a replica-exchange method [88] with the following eight temperature parameters: 300.0 K, 336.0 K, 375.0 K, 419.0 K, 470.0 K, 530.0 K, 600.0 K, and 681.0 K. The simulation was conducted with a time step of 2 fs, replicas were exchanged every 10 ps using the replica-exchange method based on the metropolis method for 10,000 exchanges, and an overall simulation time was 100 ns. From the last 24 ns of simulation, 60 conformations were extracted at equal intervals, yielding 480 conformations in total across all eight temperature replicas. The simulations were conducted on the TSUBAME3.0 supercomputer of the Tokyo Institute of Technology (currently Institute of Science Tokyo). To reduce the total computation time, we utilized two f-nodes of TSUBAME3.0 for parallel calculation for each cyclic peptide, each comprising two Intel Xeon E5-2680 v4 CPUs and four NVIDIA TESLA P100 GPUs (approximately one hour per peptide).

6.2.2 Analysis of descriptor differences between RDKit and MD conformations

We compared the MOE 3D descriptors calculated from conformations of six-monomer peptide 1NMe3 (CycPeptMPDB ID: 2328; Fig. 6.4) and 11-monomer Cyclosporin A (CycPeptMPDB ID: 7353; Fig. 6.5) generated using MD simulations and RDKit, while both cyclic peptides are known as high-permeability cyclic peptides. The descriptors analyzed include (A) potential energy (E), (B) molecular mass density (dens, molecular weight divided by 3D van der Waals volume), (C) fractional water accessible surface area of all atoms with negative partial charge (FASA-), (D) fractional positive charge weighted surface area (FCASA+), (E) fractional water accessible surface area of all polar atoms (FASA_P), and (F) VolSurf polar volume (vsurf_Wp2). Among these, the last five descriptors (B to F) are key 3D descriptors selected and utilized for CycPeptMP model construction (Table 4.4).

The potential energy of MD conformations for both cyclic peptides increases with temperature, confirming that the replica-exchange approach successfully samples diverse conformational states. For 1NMe3, the potential energy (E) of RDKit conformations lies between those of MD conformations at 470 K and 530 K (Fig. 6.4 (A)), while for Cyclosporin A, the potential energy of RDKit conformations is comparable to that of MD conformations at the highest temperature, 681 K (Fig. 6.5 (A)). For molecular density (dens; Fig. 6.4 (B), Fig. 6.5 (B)), the RDKit conformations for both cyclic peptides exhibit lower dens values compared to MD conformations, suggesting that the van der Waals volumes of RDKit conformations are larger. These results suggest that RDKit fails to generate low potential energy closed conformations, which are expected to have minimized potential energy due to maximized intramolecular interactions. This limitation highlights RDKit's inefficiency in exploring energetically favorable conformations critical for modeling cyclic peptide membrane permeability due to weak sampling algorithms. In contrast, our MD simulations could generate more compact and energetically favorable conformations through enhanced sampling. The fractional surface area descriptors (FASA-, FCASA+, FASA_P; Fig. 6.4 (C) to (E), Fig. 6.5 (C) to (E)) reveal minimal differences among MD conformations across varying temperatures, and their values are not significantly different from those of RDKit conformations. This stability suggests that these descriptors are relatively insensitive to the conformational diversity explored in MD simulations, and RDKit conformations may be sufficient for capturing these specific features in prediction models. For VolSurf polar volume (vsurf_Wp2; Fig. 6.4 (F), Fig. 6.5 (F)), MD conformations show

a slight upward trend with increasing temperature. Notably, for Cyclosporin A, the vsurf_Wp2 value of RDKit conformations is significantly higher than that of MD conformations. This discrepancy suggests that RDKit's methods may overestimate polar volume properties, particularly for more complex cyclic peptides.

Overall, these results highlight the advantages of MD simulations in generating biologically relevant conformations compared to RDKit. While RDKit can adequately capture certain descriptors, such as fractional surface areas, it shows limitations in properties like potential energy, molecular density, and polar volume. This suggests that RDKit conformations may be insufficient for tasks requiring detailed structural accuracy or compactness. Utilizing MD conformations could improve the predictive accuracy of models based on these 3D descriptors.

6.2.3 Comparison of RDKit and MD conformations in PCA space

To provide a more intuitive comparison of the overall differences in descriptors between RDKit and MD conformations, we conducted a PCA using 116 MOE 3D descriptors (removing one descriptor with a standard deviation of zero). Each descriptor was standardized using Z-scores, calculated based on the mean and standard deviation of the RDKit conformations for all cyclic peptides. As shown in Fig. 6.6, it is evident that although the contribution ratios of PC1 are relatively small (23.6% and 25.2%), the distributions of MD conformations and RDKit conformations of 1NMe3 and Cyclosporin A differ significantly in PC1. Furthermore, MD conformations exhibit a broader distribution in PC2. The differences between MD conformations across different temperatures are not very obvious.

Overall, using MD conformations, which provide greater structural diversity and potentially more biologically relevant conformations, may lead to differences in prediction results for membrane permeability.

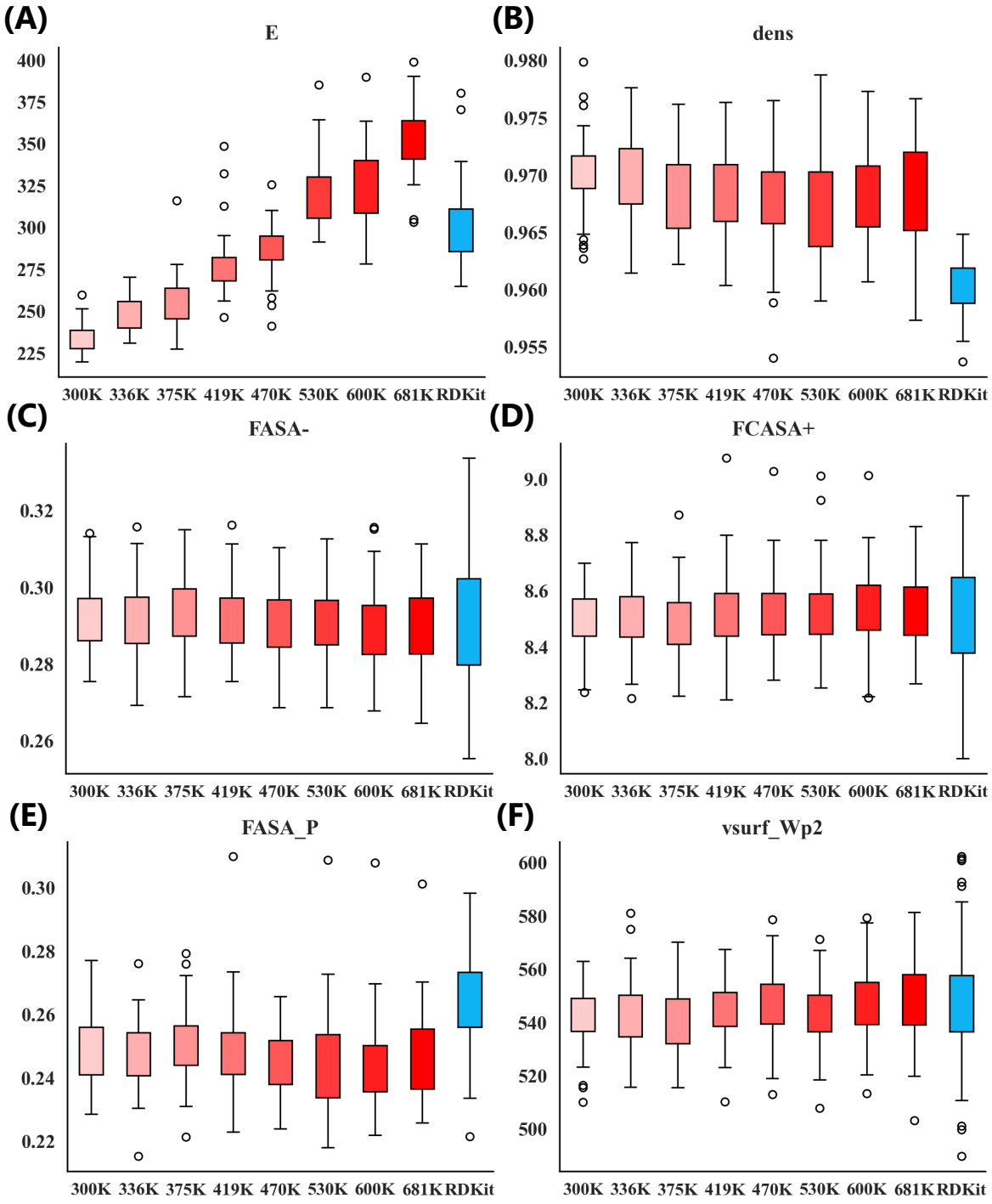


Figure 6.4: Comparison of six MOE 3D descriptors of 1NMe3 (CycPeptMPDB ID: 2328) calculated from MD simulation conformations (red) and RDKit conformations (light blue). (A) E, (B) dens, (C) FASA-, (D) FCASA+, (E) FASA_P, and (F) vsurf_Wp2.

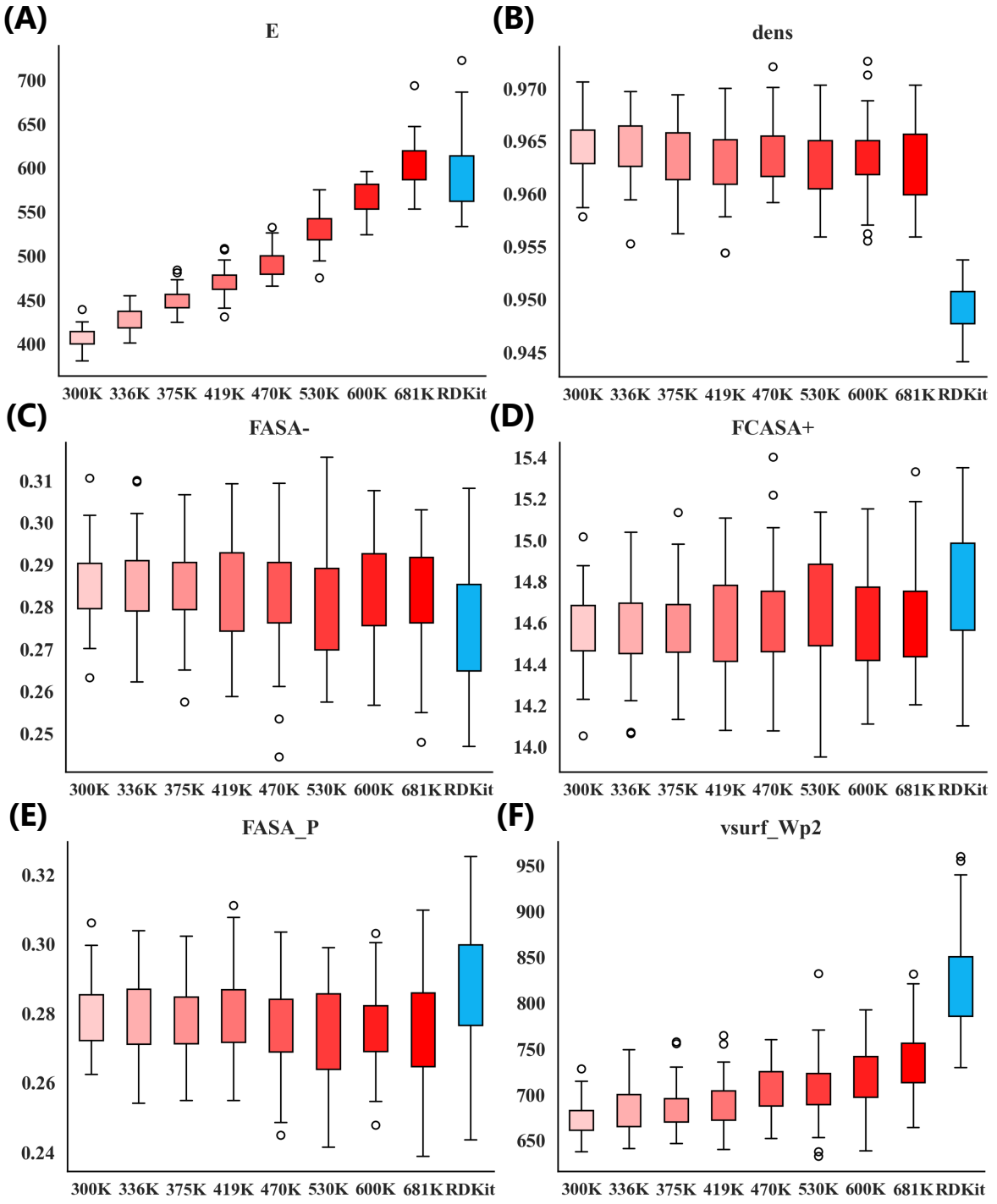


Figure 6.5: Comparison of six MOE 3D descriptors of Cyclosporin A (CycPeptMPDB ID: 7353) calculated from MD simulation conformations (red) and RDKit conformations (light blue). (A) E, (B) dens, (C) FASA-, (D) FCASA+, (E) FASA_P, and (F) vsurf_Wp2.

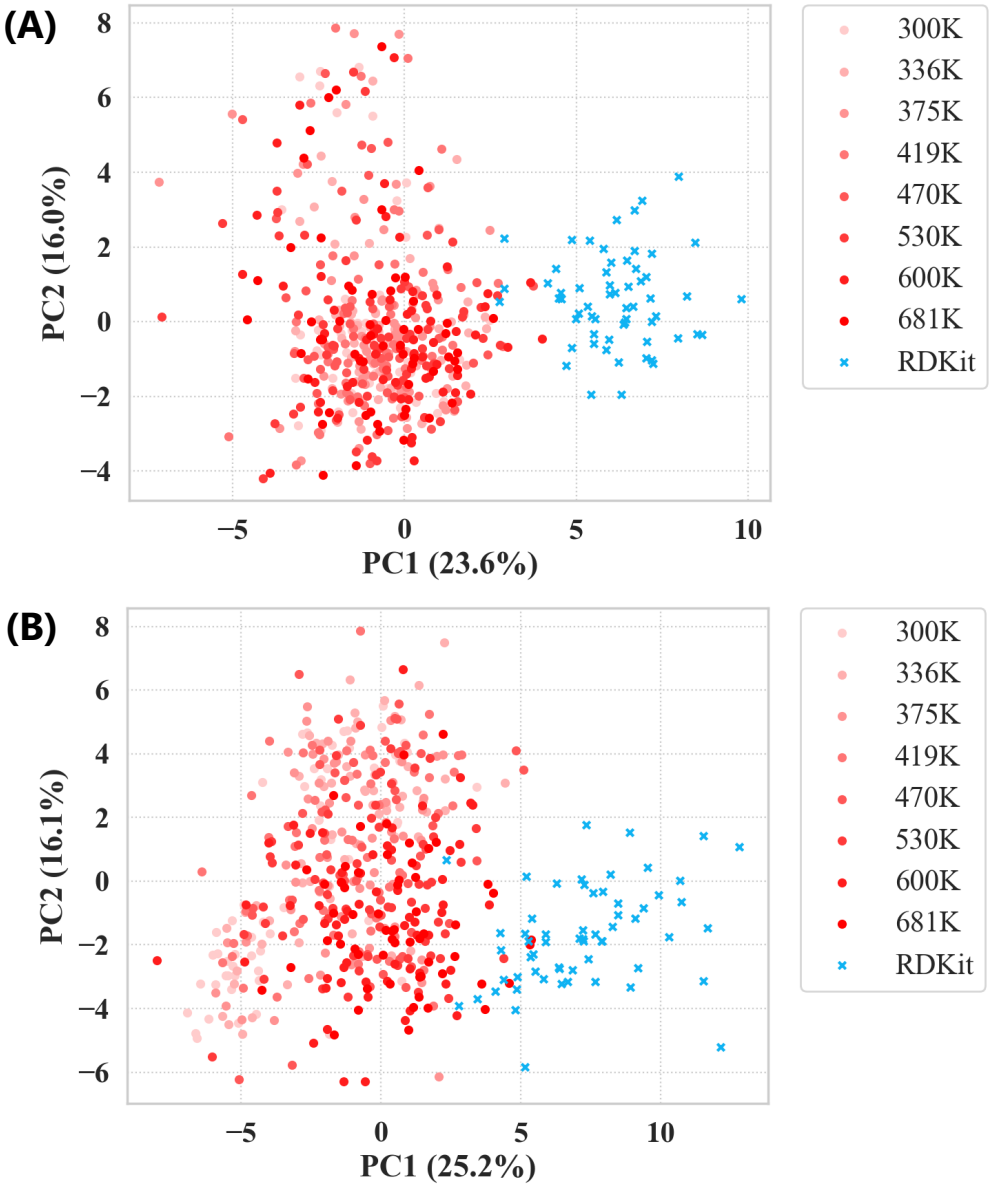


Figure 6.6: Comparison of RDKit and MD conformations of (A) 1NMe3 and (B) Cyclosporin A in PCA space, with the PC1 as the horizontal axis and the PC2 as the vertical axis; the contribution rates are shown in the parentheses of axes captions.

6.3 Conformation Generation of Monomers

To effectively capture the local structural information of cyclic peptides, we designed monomer-level features. Although other levels of features, such as fragment-level features commonly used in traditional small molecules, could also be considered for expressing local structures, we designed monomer-level features because monomers are the standard units in the chemical synthesis of cyclic peptides. When calculating the conformation of monomers, as described in Section 3.4.2, we did not extract the monomer conformations directly from the cyclic peptide conformation. Instead, we recalculated each monomer conformation individually. The reason is that the accuracy of the monomer conformation is expected to be higher than that of the cyclic peptide conformation calculated by RDKit (currently, most conformational calculation software has not been modified specifically for calculations of cyclic peptides). Furthermore, we wanted to focus on the characteristics of each monomer itself. If the monomer conformations were extracted from the cyclic peptide, the conformations of each monomer would be highly constrained by its neighbors. However, calculating monomer conformations independently might generate conformations that are not feasible within the actual cyclic peptide structure. To address this, larger capping groups, such as ACE-NME capping [206, 207, 208] (Fig. 6.7), widely used in simulating proteins and peptides, could be applied during monomer division to mimic the behavior of monomers within cyclic peptides more accurately.

To compare these two capping approaches, we selected four monomers from the membrane permeability data and compared the conformations of these monomers generated with the original capping approach with ACE-NME capping. These four monomers are the relatively small natural amino acid Leucine (L), the unnatural amino acid Pye (N, N-pyrrolidinyl glutamine), and two monomers, Sub25 and Sub27, which contain large side chain portions of the Lariat peptide. We first aligned the two types of conformations (60 conformations with original capping and 60 conformations with ACE-NME capping) of each monomer. Then, we selected the common portions of the two types of conformations and performed PCA based on the three-dimensional coordinates of these atoms. As shown in Fig. 6.8, neither monomer showed a clear difference between the two types of conformations. The results of this comparison indicate that the conformations generated with ACE-NME capping are generally consistent with those generated using the original capping approach, indicating that the original capping method is sufficient for capturing relatively accurate monomer conformations in most cases.

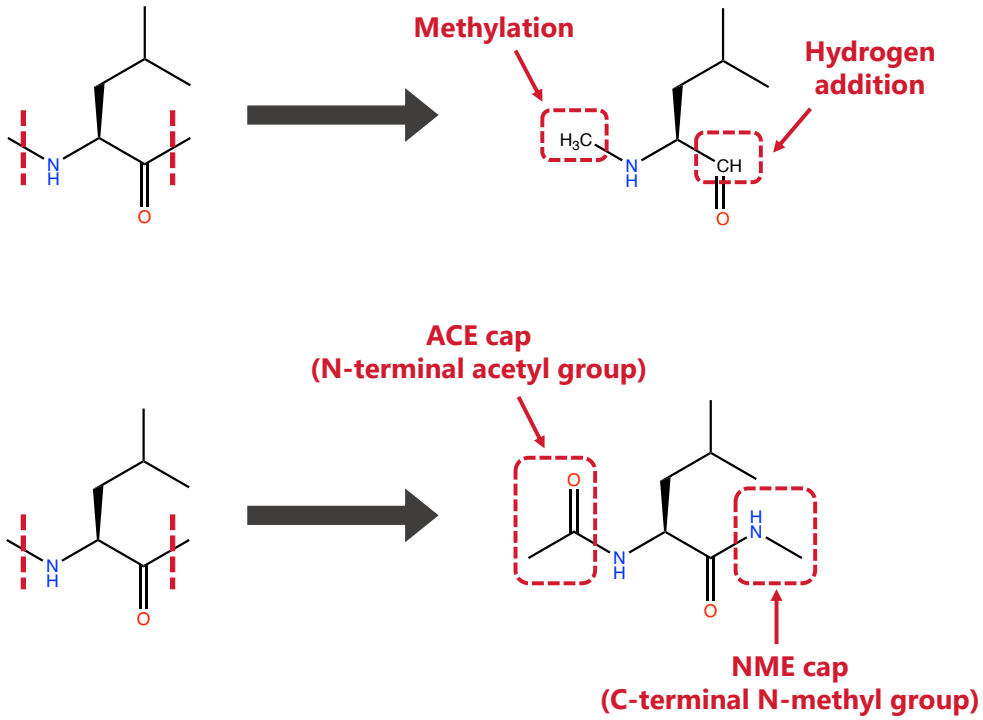


Figure 6.7: Comparison of current (top) and ACE-NME (bottom) capping methods when dividing leucine. Where ACE is an N-terminal acetyl group, and NME is a C-terminal N-methyl group.

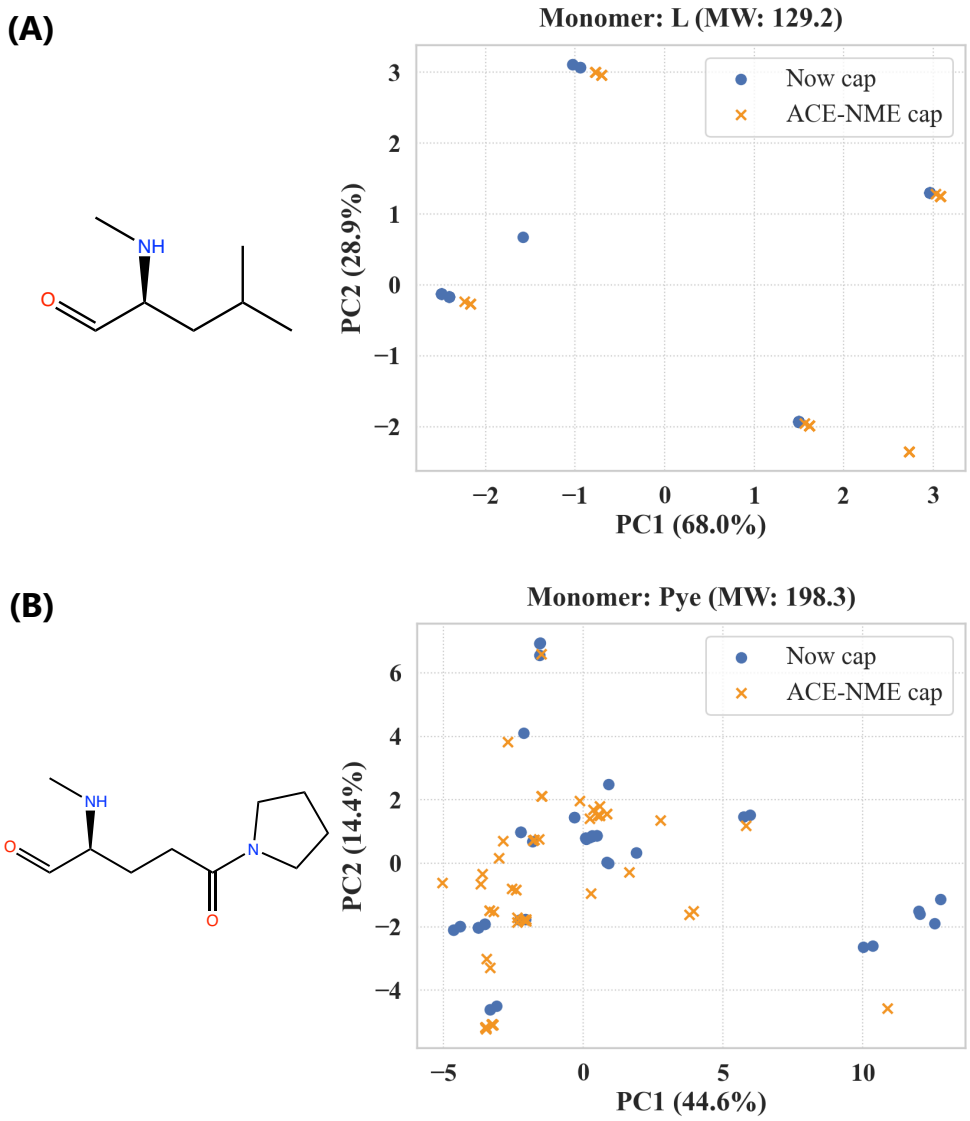


Figure 6.8: Structure and comparison of conformations generated with two capping methods in PCA space of (A) Leucine, (B) Pye, and two monomers, Sub25 (C) and Sub27 (D), which contain large side chain portions of the Lariat peptide.

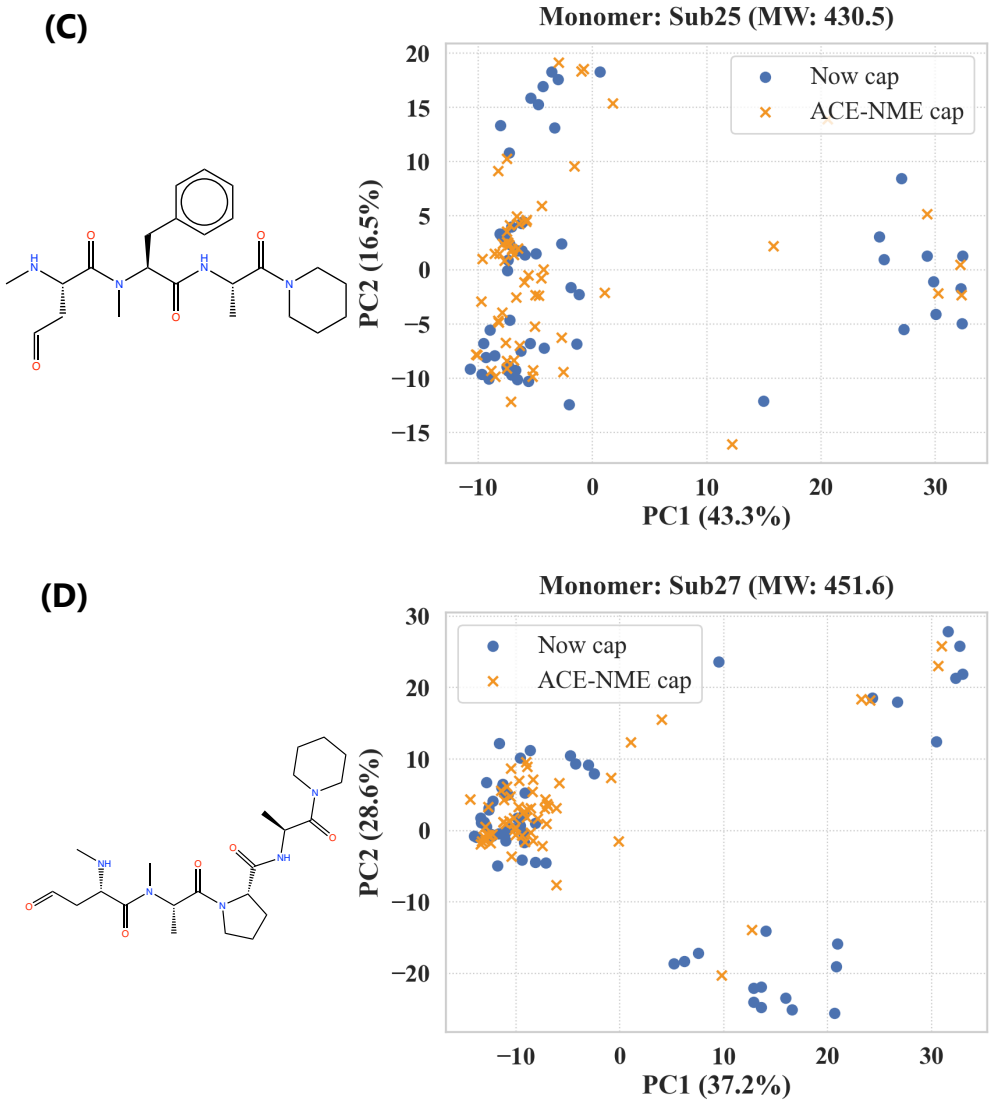


Figure 6.8: (continuation)

6.4 Effect of Atom-Level Features

Atom-level features were incorporated into the model and designed to capture fine-grained local interactions and provide detailed information on cyclic peptides. Unlike monomer- or peptide-level features, which reflect broader substructural or molecular properties, atom-level features focus on the individual atoms, their bonding environments, and spatial relative relation. These features allow the model to account for changes in local atomic environments, which could affect molecular binding to proteins, membrane passage, and other biological interactions. Furthermore, atom-level features, such as bond types, hybridization states, and distances between atoms, provide important context for understanding the global shape of the molecule and its ability to interact with external biological systems. For the prediction of membrane permeability, atom-level features had the lowest impact among the three levels of features of the fusion model (Fig. 4.14 (B); F: MAE = 0.355, F–atom: 0.368, F–mono: 0.387, and F–pep: 0.388). In contrast, for PPB rate prediction, the atom-level features were more informative (Fig. 5.9; on test set, F: MAE = 2.44%, F–atom: 4.55%, F–mono: 6.11%, and F–pep: 3.82%; on DrugBank set, F: 8.53%, F–atom: 13.76%, F–mono: 13.57%, and F–pep: 8.21%). This suggests that while atom-level information is less critical for predicting permeability, it still contributes meaningfully to predicting PPB rates, particularly in capturing certain local interactions and structures.

One possible explanation for the minimal influence of atom-level features may be that cyclic peptides are structurally larger and more complex than small molecules. For instance, in the permeability dataset, the maximum number of heavy atoms was 128, while in the PPB dataset, it was 162 (Fig. 6.9). Currently, when calculating atom-level features, we apply padding to cyclic peptides with fewer heavy atoms than the largest peptide in the dataset. Given the wide range and uneven distribution of heavy atom numbers in the dataset—particularly in the permeability dataset—many atom-level features end up being sparse, contributing less informative data to the model. This sparsity is likely one of the key factors that limit the utility of atom-level features for cyclic peptides. Moreover, we currently calculate the 3D Euclidean distances between atoms ($Conf$) based on 3D conformations generated using a relatively simple process by RDKit software, which may not provide sufficiently accurate structures. This limitation in the accuracy of the 3D conformations could further reduce the utility of atom-level features, especially when predicting permeabilities of complex cyclic peptides where precise structural information is crucial. Furthermore, the monomer- and peptide-level features rely on descriptors representing higher-order physicochemical

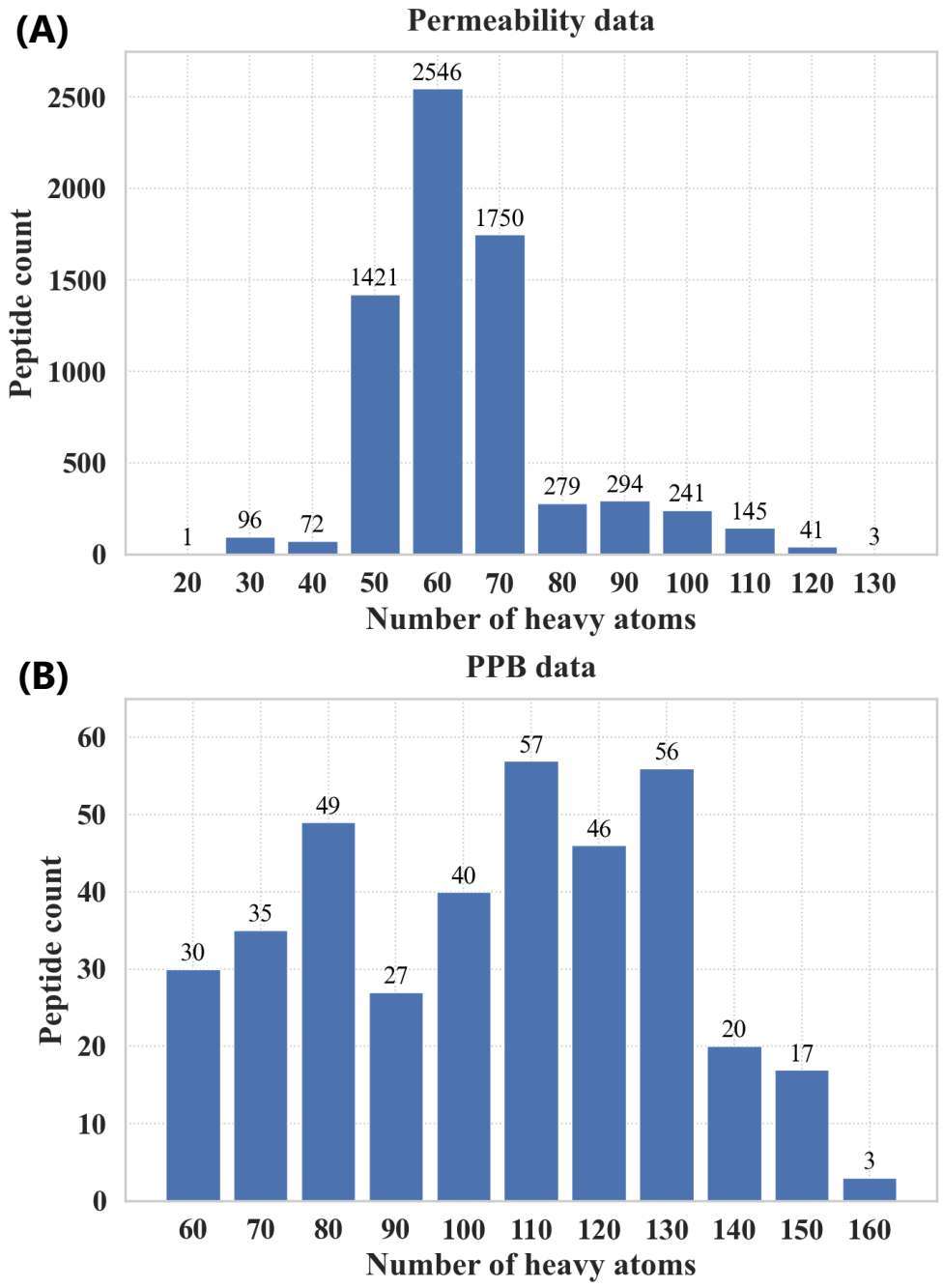


Figure 6.9: Distribution of the number of heavy atoms of (A) permeability and (B) PPB datasets.

properties, such as lipophilicity and polar surface area, which provide more advanced information. In contrast, the atom-level features primarily include basic information such as atom types and relative distances, which may not be sufficient to capture the complex properties that govern the interactions between cyclic peptides and biological systems. Regarding integrating the three levels of information (peptide, monomer, and atom), the current model concatenates the latent feature vectors from each sub-model. While this approach is straightforward, it may not effectively capture the unequal contributions of each feature level, particularly for tasks where certain types of features (such as monomer-level features for PPB rate prediction) are more relevant. A more refined integration approach, such as attention mechanisms or weighted combination methods, could better balance the contributions of the three-level features.

6.5 Effect of Single-Level Data Augmentation

We implemented three different data augmentation techniques to perform peptide-, monomer-, and atom-level augmentation. These augmentation methods were designed based on the inherent complexity of cyclic peptide conformational changes, the nature of cyclic peptide sequences, and the flexibility of SMILES representations. Data augmentation significantly improved prediction accuracy for both membrane permeability and PPB rate predictions. Furthermore, we assessed model performance at various augmentation folds (Section 4.4.6 and Section 5.3.2). For example, when using 5-fold augmentation, each sub-model (peptide, monomer, and atom) employed five distinct input replicas, respectively. Additional experiments were performed on the fusion model of membrane permeability prediction to discuss which single-level augmentation would be most effective. For n-fold augmentation, we applied the augmentation only to one selected sub-model while using n repetitions of the same input for the other two sub-models. This approach allowed us to isolate the effect of each augmentation method and evaluate the individual contribution of peptide-, monomer-, and atom-level augmentation on the overall model performance.

The MAEs of the fusion model, when only peptide-, monomer-, and atom-level augmentations were applied, respectively, are shown in Fig. 6.10 (A) to (C). In this case, only monomer-level augmentation showed an improvement, with MAE decreasing up to 10-fold augmentation (M–1: 0.456, M–10: 0.384), beyond which no further changes were observed (Fig. 6.10 (B)). For peptide-level and atom-level augmentations, the MAE did not decrease as replicas increased, and the lowest MAE was observed when no augmentation was applied (P–1 and A–1: 0.456). A possible reason for this result is that the monomer-level feature is completely independent, while both peptide- and atom-level features rely on peptide conformation. For example, when peptide-level augmentation is applied, the values of 3D descriptors vary across different replicas, but at the same time, in the atom-level input, the 3D distance matrix ($Conf$) between atoms does not change across replicas, which may lead to a mismatch in correspondence.

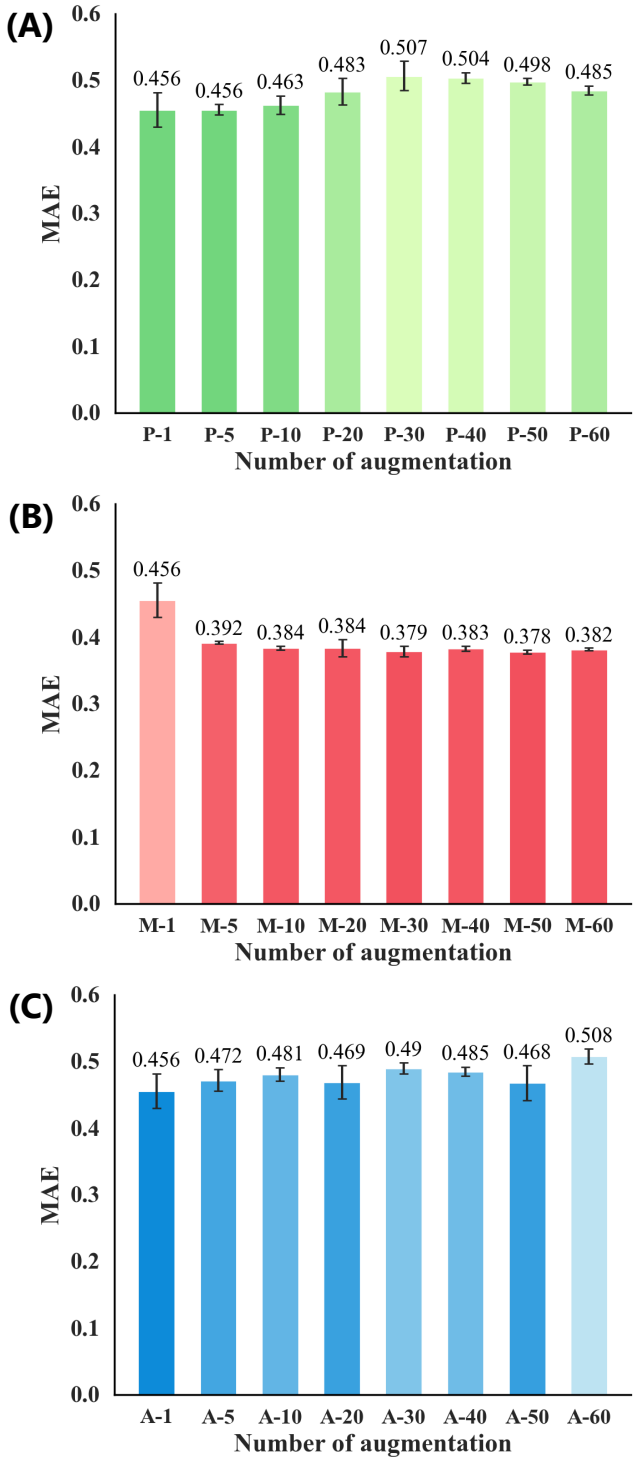


Figure 6.10: Ablation study results (MAE) for membrane permeability prediction with the fusion model on the test set, showing performance across varying numbers of input replicas at each augmentation level: (A) Peptide-level, (B) Monomer-level, and (C) Atom-level. Note that for each level, the other two levels use duplicated inputs.

Chapter 7

Conclusion

7.1 Conclusion

In this thesis, we describe the development of a multi-level molecular features design method and data augmentation strategy for accurately predicting the permeability and PPB rate of cyclic peptides. These novel approaches were implemented for two prediction models: fusion model-based CycPeptMP for cell membrane permeability prediction and monomer model-based CycPeptPPB for PPB rate prediction related to plasma stability. Both models demonstrated significant improvements over conventional methods (CycPeptMP: MAE = 0.355, R = 0.883; CycPeptPPB: MAE = 4.40%, R = 0.947). Below, we summarize the contributions provided by this work.

7.1.1 Contributions

- We proposed a multi-level molecular feature design method, which integrates peptide-, monomer-, and atom-level features to comprehensively capture the structural complexity of cyclic peptides (Chapter 3). It was inspired by the physicochemical knowledge of membrane permeability and PPB of cyclic peptides, aimed to concurrently capture both the critical local substructures and the overall peptide structures. Additionally, to address the scarcity of cyclic peptide data, we introduced various data augmentation strategies tailored to cyclic peptides, including generating multiple 3D conformations, sequence arrangements, and SMILES enumeration, to improve model training efficiency. These techniques effectively increased the diversity of the training dataset, considering the conformational flexibility and circularity of peptides.

- We constructed the world's first cyclic peptide membrane permeability database, CycPeptMPDB (Section 4.2, Appendix C). We conducted an extensive search through published papers and pharmaceutical company patent documents, compiling over 7,000 cyclic peptide structures along with their experimentally measured membrane permeabilities from 47 distinct sources. CycPeptMPDB also provides various valuable information, such as unified sequence representations that are essential for the development of cyclic peptide drugs. By addressing the limitations caused by insufficient data, CycPeptMPDB has significantly advanced the development of DL-based prediction methods for cyclic peptide membrane permeability, overcoming a major hurdle in the field.

- We developed CycPeptMP: an accurate and efficient cyclic peptide membrane permeability prediction method (Chapter 4). By leveraging the three-level molecular features, the fusion model-based CycPeptMP effectively captured critical local substructures and the global molecular conformation, which are essential for predicting membrane permeability in cyclic peptides. CycPeptMP outperformed diverse existing methods, including traditional machine learning approaches for cyclic peptide permeability prediction and state-of-the-art DL-based small molecule property prediction models. Moreover, CycPeptMP successfully predicted peptide permeabilities, which are challenging for MD-based methods, while significantly lower computational costs.

  The practical value of CycPeptMP in real-world cyclic peptide drug development lies in its capability to rapidly filter out candidate compounds with inadequate permeability during lead optimization. With MAE = 0.355, CycPeptMP achieves a high level of predictive accuracy; given the logarithmic nature of LogP_{exp}, a prediction difference of 0.35 translates to an approximate 2.25-fold difference in actual permeability speed. The collected data encompass a broad range of LogP_{exp} values, from -10 to -4, representing a six-log unit range in permeability speed. Within this range, a 2.25-fold discrepancy remains acceptable in practical drug development. Additionally, MAE = 0.355 aligns closely with experimental error in biochemical assays. For example, Cyclosporin A, a widely used cyclic peptide drug, and 1NMe3, a known high-permeability cyclic peptide, have 10 and 14 distinct measurement values (not limited to PAMPA) recorded in CycPeptMPDB. The LogP_{exp} values for Cyclosporin A vary from -5.01 to -6.20, with a standard deviation of 0.360, while 1NMe3 values range from -4.40 to -6.40, with a standard deviation of 0.690. CycPeptMP offers predictions comparable to assay-level

reliability, making it a viable alternative to traditional biochemical assays.

The current computational cost for predicting a single cyclic peptide is divided into three main stages: generating 3D conformations (approximately three hours), calculating molecular features (approximately ten minutes), and the final prediction stage (a few seconds on an NVIDIA RTX4090 GPU), comparable in speed to traditional ML-based methods. This cost is significantly lower than biochemical assays (requiring actual compound synthesis and up to 24 hours of measurement) and MD simulations (requiring extensive GPU calculations over several tens of hours). Furthermore, based on ablation studies, omitting 3D structural information yields a predictive accuracy of MAE = 0.380, which remains strong. Skipping the most time-consuming 3D conformation generation step reduces computational time to ten minutes. Finally, although we applied 60-fold data augmentation, achieving similar performance with only 20-fold augmentation (MAE = 0.358) could cut computational costs by two-thirds, reducing the prediction time to approximately one hour per peptide.

- We also developed CycPeptPPB: a highly accurate cyclic peptide PPB rate prediction model (Chapter 5). Monomer-level features played the most important role, and the monomer model-based CycPeptPPB outperformed various comparison methods, demonstrating its ability to effectively capture the essential substructural information critical for predicting the PPB rates of cyclic peptides. Furthermore, CycPeptPPB's saliency map analysis enabled the identification of important monomers that influence PPB, contributing to more efficient cyclic peptide drug design and optimization.

  CycPeptPPB demonstrates sufficient accuracy to selectively filter compounds with inadequate PPB rates during early drug development. With MAE = 4.40%, CycPeptPPB surpasses traditional ML-based models (MAE of 10–20%), providing a robust level of precision. The collected public datasets (Tajimi and DrugBank datasets) contain PPB rates distributed across a range from 0% to 100%. Given this variability, a prediction error of less than 5% is within acceptable limits for practical drug development. Moreover, through discussions with pharmaceutical development specialists at PeptiDream Inc. during the development of CycPeptPPB, we concluded that CycPeptPPB has practical applicability in real-world drug development.

  Regarding computational cost, while CycPeptPPB is based on the monomer model and thus more lightweight than the fusion model, generating 3D conforma-

tions remains the most computationally demanding step, with each cyclic peptide prediction requiring approximately half an hour. Compared to traditional PPB assays (requiring compound synthesis and up to 24 hours of measurement), CycPeptPPB offers a significantly reduced time investment. Notably, CycPeptPPB achieves MAE = 4.53% even without using 3D structural information, maintaining high predictive accuracy while reducing the prediction time for each peptide to just a few minutes.

The completion of this study marks the first time that large-scale cyclic peptide membrane permeability and PPB rate prediction, which cannot be evaluated by traditional screening methods, has been realized. This advancement enables the efficient design of candidate cyclic peptides and accelerates the development of cyclic peptide drugs, which are expected to revolutionize drug discovery. As a result, cyclic peptide-based medications will become available for treating more diseases, and with innovations such as oral administration, the market share of peptide drugs, which stood at only 5% in 2019 [18], is expected to expand significantly.

Furthermore, the multi-level feature design and data augmentation methods proposed in this study are not limited to cyclic peptides. For example, they can also be extended and applied to linear peptides, enabling a broader range of drug discovery and optimization efforts.

7.2 Future Works

This section discusses future directions from two perspectives: the technical improvements of the proposed methods and broader challenges in the cyclic peptide drug discovery field. The technical challenges focus on improving the predictive accuracy and generalization of the proposed models, while the broader challenges emphasize addressing the remaining issues in cyclic peptide drug development after achieving large-scale permeability and PPB rate predictions.

7.2.1 Technical improvements of proposed methods

Improvement of 3D conformations generation method

Currently, the 3D conformations are generated using a relatively simple approach through the RDKit software. For permeability prediction where 3D information plays a critical role, as discussed in Section 6.2, we have already utilized MD simulation

to generate cyclic peptide conformations. Comparison with RDKit conformations revealed significant differences in the structural features derived from MD. This suggests that training prediction models using MD conformations could potentially improve performance due to their greater structural diversity and higher accuracy. Notably, by employing two TSUBAME3.0 f-nodes with eight GPUs in parallel, the computation time for a single cyclic peptide was reduced to approximately one hour. For PPB rate prediction, leveraging 3D conformations obtained from docking poses may similarly enhance the predictive performance of the models.

Improvement of integration method of sub-level features

At present, the three sub-models extract latent features, which are then simply concatenated for the final prediction. Considering cases like PPB, where monomer-level features are particularly crucial, a weighted combination or alternative integration method, rather than straightforward concatenation, may yield better prediction accuracy by giving appropriate importance to each feature level.

Further data collection

The available PPB data is still quite limited. Although there is already a substantial amount of data for membrane permeability, the vast possible combinations of monomers make the existing data insufficient to cover all cases. Specially modified monomers not included in the current dataset may still pose challenges, and expanding the dataset to include such examples would improve model robustness.

7.2.2 Challenges in cyclic peptide drug discovery

Immunogenicity

Oligopeptides are generally considered to be poor immunogens [209]. However, notable examples, such as discontinuing late-stage clinical trials for linear peptides, reveal that immunogenicity can still pose significant issues [4]. For instance, medium-sized peptides (3–5 kDa), even those derived from human sequences like taspoglutide, have been shown to cause severe immune responses. The lack of extensive research on smaller macrocyclic peptides prevents definitive conclusions about their immunogenic potential. Further studies are essential to advance the therapeutic use of macrocyclic peptides, uncover the factors that drive immunogenic or non-immunogenic responses, and ensure their safety in clinical applications.

Clearance pathways prediction

The metabolism and elimination of drugs involve coordinated actions of various metabolic enzymes and transport proteins, collectively called clearance pathways. Accurately predicting clearance pathways is crucial for ensuring the safe and effective clinical use of therapeutic agents. However, predicting clearance pathways of cyclic peptides presents a significant challenge. The scarcity of relevant experimental data limits the feasibility of large-scale machine learning approaches, which rely heavily on extensive and diverse datasets for model training and validation.

Advancing docking techniques for cyclic peptides

While cyclic peptides with high target affinity can often be identified through display techniques, understanding their binding mechanisms or leveraging docking to predict PPB rates and PPI inhibition, etc., requires specialized, high-accuracy docking methods. However, due to several limitations, most currently available docking programs are not well-suited for protein-cyclic peptide docking [210]. First, the structural features of cyclic peptides, including closed loops and noncanonical amino acids, challenge the applicability of traditional docking programs. Second, the unique conformational flexibility of cyclic peptides often results in distributions that deviate from those predicted by classical force fields, which are typically optimized for linear peptides. This divergence significantly impacts the predictive accuracy of these docking programs. Existing docking programs capable of handling protein-cyclic peptide interactions, such as AutoDock CrankPep (ADCP) [211] and HADDOCK2.4 [212], usually extend methodologies developed for the protein-small molecule or protein-peptide docking. These tools typically generate cyclic peptide conformation ensembles using amino acid sequence data and perform docking with the resulting structures. For instance, ADCP employs a ring closure method with cyclization potentials or distance restraints during conformation search and demonstrates state-of-the-art sampling performance based on a dataset of 38 protein-cyclic peptide complexes. However, it is primarily limited to cyclic peptides composed of canonical amino acids and cyclized by a narrow range of bond types, restricting their broader applicability. Additionally, the scoring functions in programs like AutoDock, originally developed and calibrated for small molecules, often fail to describe protein-peptide interactions accurately [211]. Adapting these scoring functions or creating new ones tailored explicitly for cyclic peptides would significantly enhance pose ranking and overall docking accuracy. Overall, addressing these limitations and developing dedicated docking methods is crucial for advancing

cyclic peptide-based drug discovery and development.

Expanding computational methods to bicyclic peptides

Most existing computational approaches are designed for monocyclic peptides, overlooking the unique structural and functional characteristics of bicyclic peptides. However, bicyclic peptides, which naturally occur and exhibit significant biological activities, offer additional advantages over monocyclic peptides. These include increased conformational rigidity, enhanced target binding affinity and selectivity, improved metabolic stability, and better membrane permeability [213]. The lack of computational frameworks that effectively account for bicyclic peptides' structural and functional complexity limits our understanding and ability to design these molecules. Addressing this gap is essential for advancing the understanding and design of bicyclic peptides, broadening the scope of cyclic peptide-based drug application.

Appendix A

Full List of FDA-Approved Macrocyclic Drugs

Table A.1: Full therapeutic indications and target classification for the FDA-approved macrocyclic drugs dataset ($n = 72$) (cited from [5]). Five macrocycles (macrocycle names in bold) are duplicated because each is used in two therapeutic indications. The table is ordered by therapeutic indication (alphabetically) and then by target (alphabetically). Complete target names are reported. NA: Target not available.

Therapeutic indication	Target	Drug	Approval year	Origin	Absorption
Achondroplasia	Atrial natriuretic peptide receptor (NPR)	Vosoritide	2021	Natural product derivative	Parenteral
Acromegaly	Somatostatin receptor (SSTR)	**Lanreotide**	2007	Natural product derivative	Parenteral
Acute coronary syndrome	Integrin beta-3 (CD61)	Eptifibatide	1998	Natural product derivative	Parenteral
Antidiuretic	Vasopressin receptors (VR)	Desmopressin	1978	Natural product derivative	Oral
	Vasopressin receptors (VR)	Vasopressin	2014	Natural product	Parenteral
Autoimmune diseases	Cyclophilin (CyP), Calcium signal-modulating cyclophilin ligand (CAMLG), Calcineurin subunit B (CNB)	Voclosporin	2021	Natural product derivative	Oral
	Cyclophilin (CyP), Calcium signal-modulating cyclophilin ligand (CAMLG), Calcineurin subunit B (CNB)	**Cyclosporin**	1983	Natural product	Oral
	FKBP12, Calcineurin subunit B (CNB)	**Tacrolimus**	1994	Natural product	Oral
	FKBP12, Calcineurin subunit B (CNB)	Pimecrolimus	2001	Natural product derivative	Parenteral
Chronic Idiopathic Constipation (CIC)	Guanylate cyclase soluble subunit alpha-2 (GUCY1A2)	Plecanatide	2017	Natural product derivative	Parenteral
Chronic pain	Voltage-dependent N-type calcium channel subunit alpha-1B (CACNA1B)	Ziconotide	2004	Natural product	Parenteral
Cushing's disease	Somatostatin receptor (SSTR)	Pasireotide	2012	Natural product derivative	Parenteral
Genetic obesity	Melanocortin receptor (MCR)	Setmelanotide	2020	Natural product derivative	Parenteral
Heart failure	Atrial natriuretic peptide receptor (NPR)	Nesiritide	2001	Natural product	Parenteral

Table A.1: (continuation)

Therapeutic indication	Target	Drug	Approval year	Origin	Absorption
Immuno-suppressant	Cyclophilin (CyP), Calcium signal-modulating cyclophilin ligand (CAMLG), Calcineurin subunit B (CNB)	**Cyclosporin**	1983	Natural product	Oral
	FKBP12, Calcineurin subunit B (CNB)	**Tacrolimus**	1994	Natural product	Oral
	FKBP12, Serine/threonine-protein kinase mTOR	**Sirolimus**	1999	Natural product	Oral
	FKBP12, Serine/threonine-protein kinase mTOR	**Everolimus**	2009	Natural product derivative	Oral
Induction of labor	Oxytocin receptor (OXTR)	Oxytocin	1980	Natural product	Parenteral
Infection: Antibacterial	16S/23S rRNA (cytidine-2'-O)-methyltransferase TlyA	Capreomycin	1971	Natural product	Parenteral
	23S ribosomal RNA (50S)	Azithromycin	1991	Natural product derivative	Oral
	23S ribosomal RNA (50S)	Clarithromycin	1991	Natural product derivative	Oral
	23S ribosomal RNA (50S)	Dirithromycin	1995	Natural product derivative	Oral
	23S ribosomal RNA (50S)	Erythromycin	1964	Natural product	Oral
	23S ribosomal RNA (50S)	Telithromycin	2004	Natural product derivative	Oral
	Bacterial membrane	Colistimethate	1970	Natural product	Parenteral
	Bacterial membrane	Daptomycin	2003	Natural product	Parenteral
	Bacterial membrane	Polymyxin B	1951	Natural product	Parenteral
	C55-isoprenyl pyrophosphate	Bacitracin	1948	Natural product	Parenteral
	NAM/NAG peptide (D-Ala-D-Ala)	Dalbavancin	2014	Natural product derivative	Parenteral

Table A.1: (continuation)

Therapeutic indication	Target	Drug	Approval year	Origin	Absorption
	NAM/NAG peptide (D-Ala-D-Ala)	Oritavancin	2014	Natural product derivative	Parenteral
	NAM/NAG peptide (D-Ala-D-Ala)	Telavancin	2009	Natural product derivative	Parenteral
	NAM/NAG peptide (D-Ala-D-Ala)	Vancomycin	1958	Natural product	Parenteral
	RNA polymerase	Fidaxomicin	2011	Natural product	Parenteral
	RNA polymerase	Rifabutin	1992	Natural product derivative	Oral
	RNA polymerase	Rifampicin	1971	Natural product derivative	Oral
	RNA polymerase	Rifamycin	2018	Natural product	Parenteral
	RNA polymerase	Rifapentine	1998	Natural product derivative	Oral
	RNA polymerase	Rifaximin	2004	Natural product derivative	Parenteral
	Streptogramin A acetyltransferase	Dalfopristin	1999	Natural product derivative	Parenteral
Infection: Antifungal	1,3-beta-glucan synthase component (FKS1)	Anidulafungin	2006	Natural product derivative	Parenteral
	1,3-beta-glucan synthase component (FKS1)	Caspofungin	2001	Natural product derivative	Parenteral
	1,3-beta-glucan synthase component (FKS1)	Micafungin	2005	Natural product derivative	Parenteral
	Ergosterol	Amphotericin B	1966	Natural product	Parenteral
	Ergosterol	Natamycin	1978	Natural product	Parenteral
	Ergosterol	Nystatin	1964	Natural product	Parenteral
Infection: Antiparasitic	GABA-A gated chloride channel (GABAA)	Moxidectin	2018	Natural product	Oral
	Glutamate-gated chloride channel (GluCl), GABA-A gated chloride channel (GABAA)	Ivermectin	1996	Natural product	Oral
(Hepatitis C) Antiviral Infection:	HCV NS3/4A protease	Glecaprevir	2017	*De novo*	Oral
	HCV NS3/4A protease	Grazoprevir	2016	*De novo*	Oral
	HCV NS3/4A protease	Paritaprevir	2014	*De novo*	Oral
	HCV NS3/4A protease	Simeprevir	2013	*De novo*	Oral
	HCV NS3/4A protease	Voxilaprevir	2017	*De novo*	Oral
Macular degeneration	NA (Reactive oxygen species) (ROS)	Verteporfin	2000	Natural product derivative	Parenteral

Table A.1: (continuation)

Therapeutic indication	Target	Drug	Approval year	Origin	Absorption
Oncology	ALK receptor	Lorlatinib	2018	*De novo*	Oral
	CXCR4 chemokine receptor	Plerixafor	2008	*De novo*	Parenteral
	DNA	Dactinomycin	1964	Natural product	Parenteral
	Histone deacetylase 1,2 (HDAC)	Romidepsin	2009	Natural product	Parenteral
	NA (Reactive oxygen species) (ROS)	Porfimer sodium	1995	Natural product derivative	Parenteral
	Prostate-specific antigen (PSA)	Lutetium Lu-177 Vipivotide Tetraxetan	2022	Natural product derivative	Parenteral
	FKBP12, Serine/threonine-protein kinase mTOR	**Everolimus**	2009	Natural product derivative	Oral
	FKBP12, Serine/threonine-protein kinase mTOR	**Sirolimus**	1999	Natural product	Oral
	FKBP12, Serine/threonine-protein kinase mTOR	Temsirolimus	2007	Natural product derivative	Parenteral
	Somatostatin receptor (SSTR)	**Lanreotide**	2007	Natural product derivative	Parenteral
	Somatostatin receptor (SSTR)	Lutetium Lu 177 Dotatate	2018	Natural product derivative	Parenteral
	Somatostatin receptor (SSTR)	Octreotide	1988	Natural product derivative	Oral
	Tubulin	Eribulin	2010	Natural product derivative	Parenteral
	Tubulin	Ixabepilone	2007	Natural product derivative	Parenteral
	Tyrosine-protein kinase JAK2 Receptor-type tyrosine-protein kinase FLT3	Pacritinib	2022	*De novo*	Oral
Premenopausal women (with hypoactive sexual desire disorder)	Melanocortin receptor (MCR)	Bremelanotide	2019	Natural product derivative	Parenteral
Vitamin B12 deficiency	Methionine synthase (MS) Methylmalonyl-CoA mutase (MCM) Methionine synthase reductase (MTRR) (mitochondrial)	Cyanocobalamin	1942	Natural product	Oral
	Methionine synthase (MS) Methylmalonyl-CoA mutase (MCM) Methionine synthase reductase (MTRR) (mitochondrial)	Hydroxocobalamin	1975	Natural product	Oral

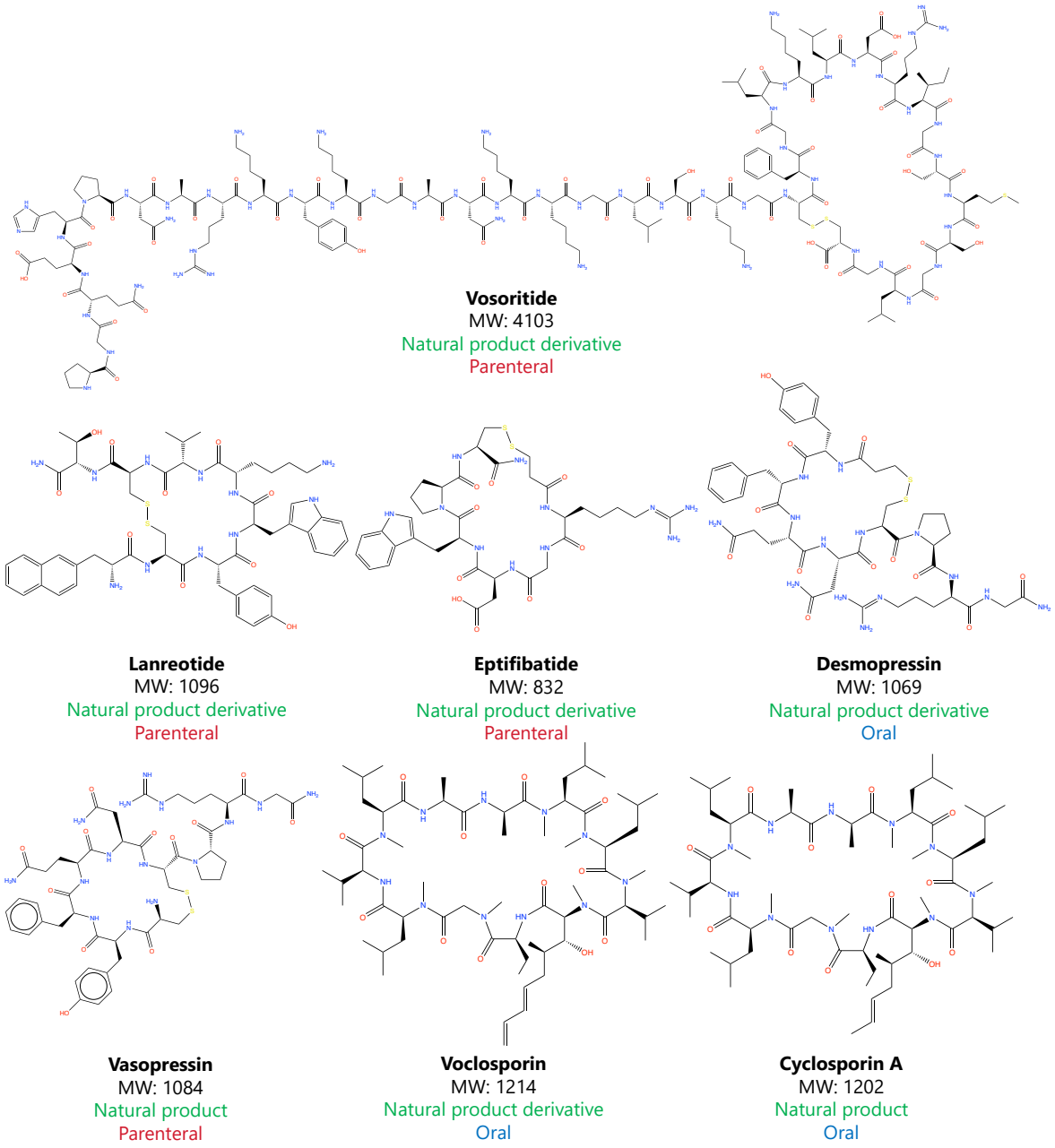


Figure A.1: Structure of all 67 FDA-approved macrocyclic drugs. Structures and molecular weights are from PubChem and ChEMBL databases.

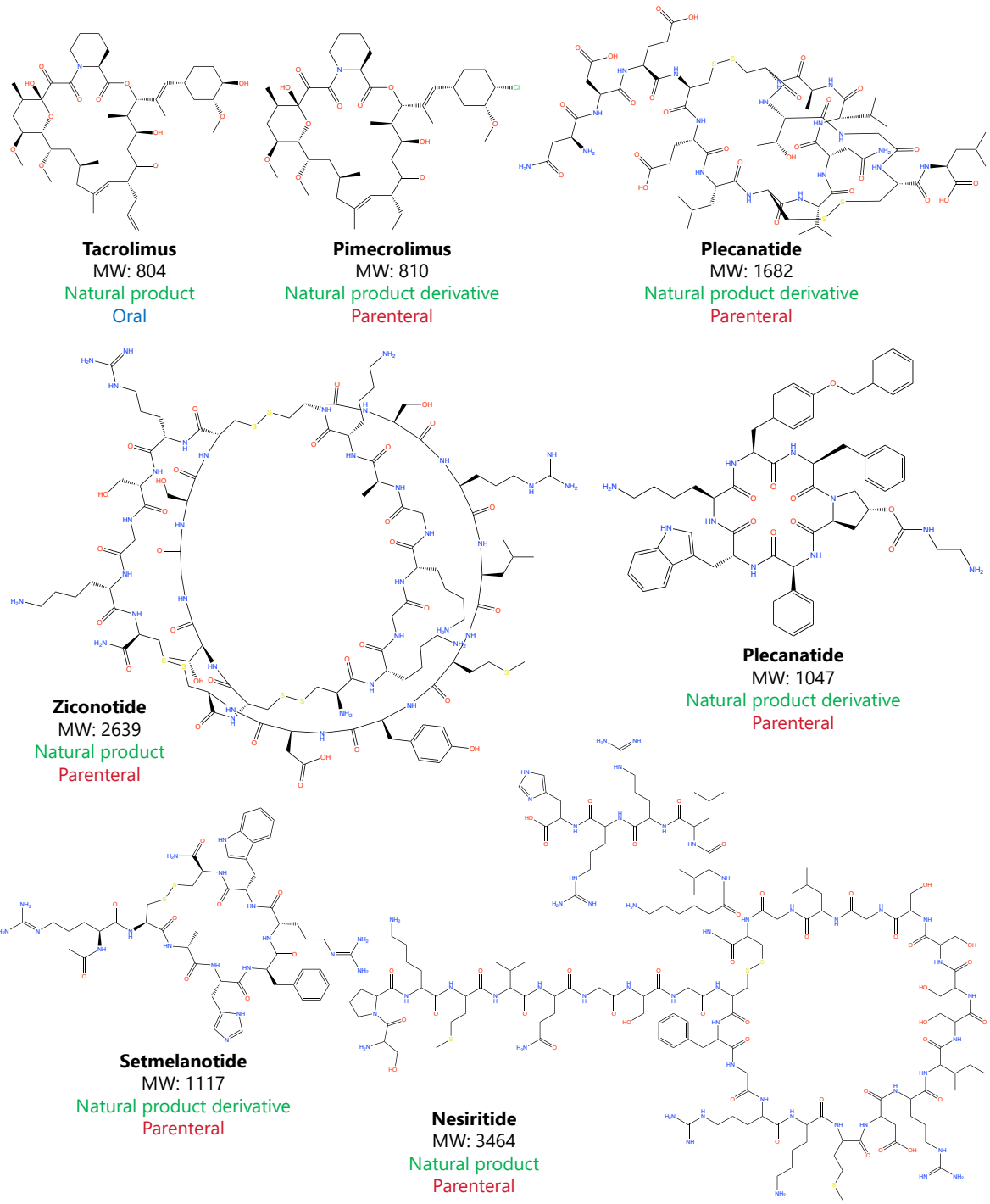
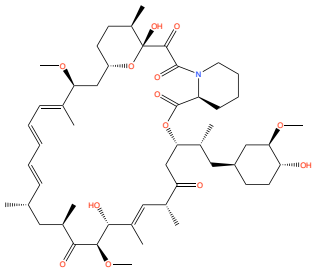
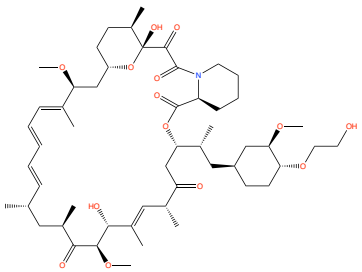


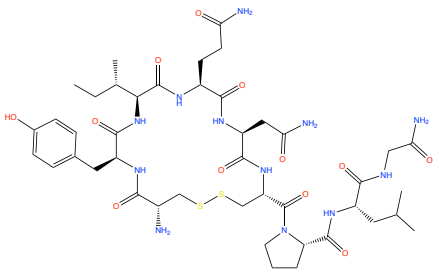
Figure A.1: (continuation)



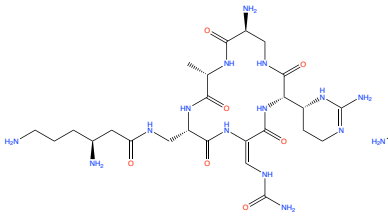
Sirolimus
MW: 914
Natural product
Oral



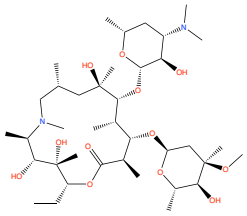
Everolimus
MW: 958
Natural product derivative
Oral



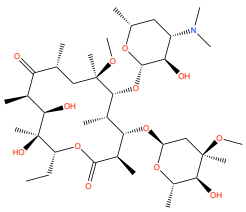
Oxytocin
MW: 1007
Natural product
Parenteral



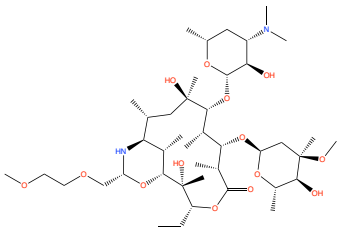
Capreomycin
MW: 1321
Natural product
Parenteral



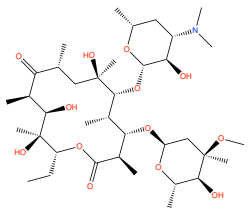
Azithromycin
MW: 749
Natural product derivative
Oral



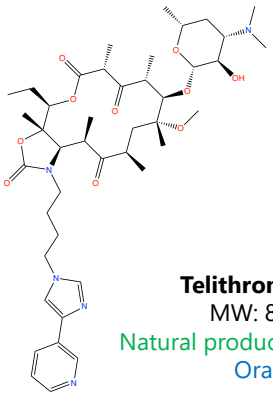
Clarithromycin
MW: 748
Natural product derivative
Oral



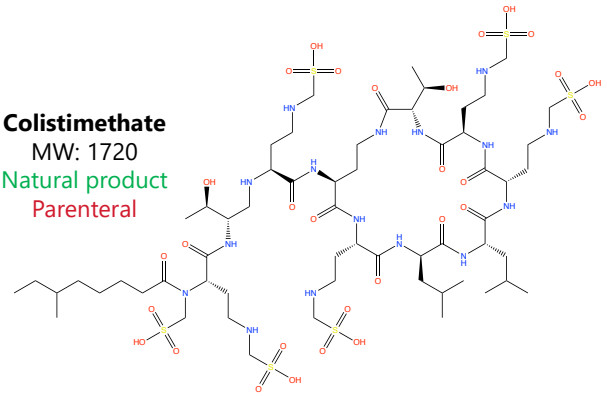
Dirithromycin
MW: 835
Natural product derivative
Oral



Erythromycin
MW: 733
Natural product
Oral



Telithromycin
MW: 812
Natural product derivative
Oral



Colistimethate
MW: 1720
Natural product
Parenteral

Figure A.1: (continuation)

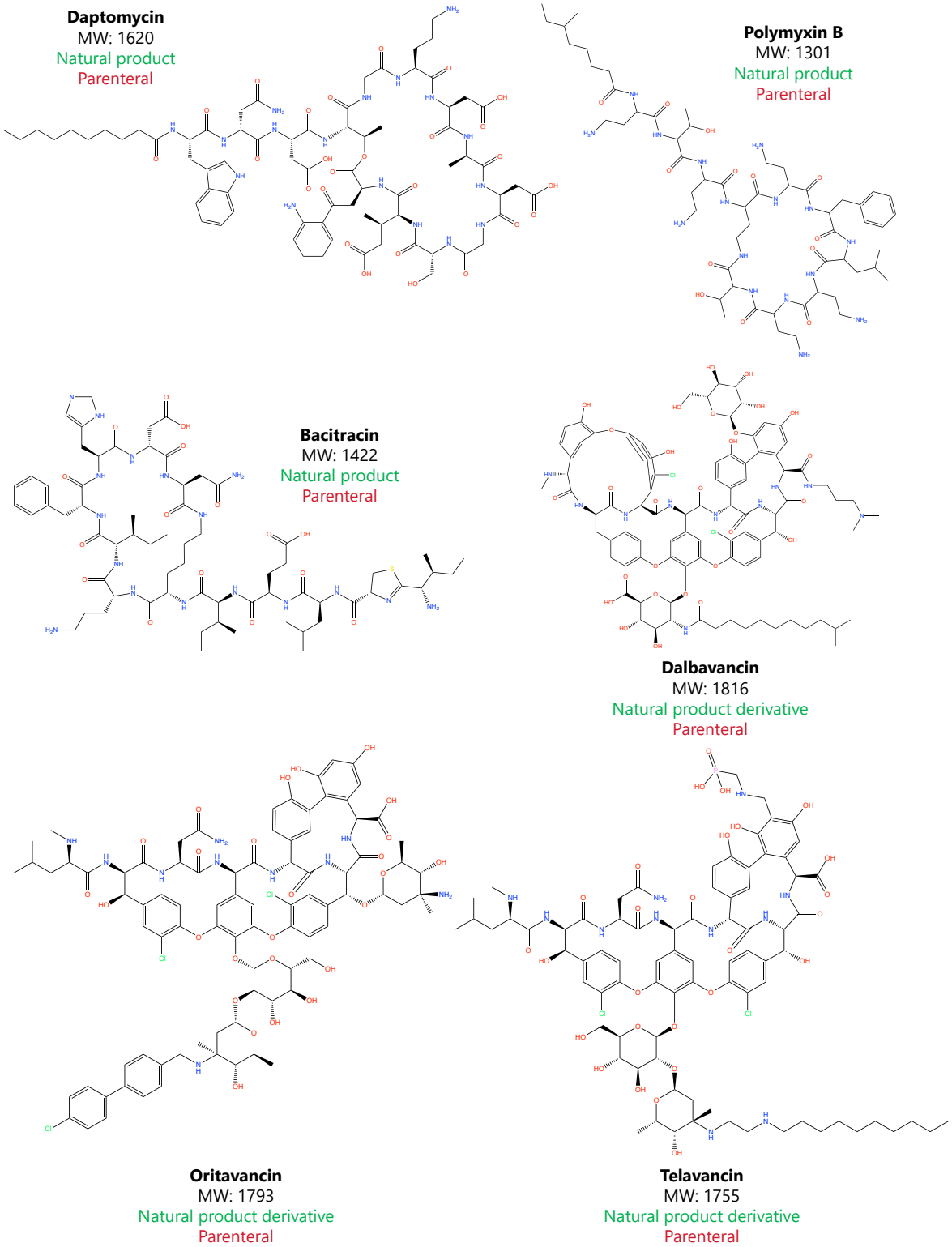


Figure A.1: (continuation)

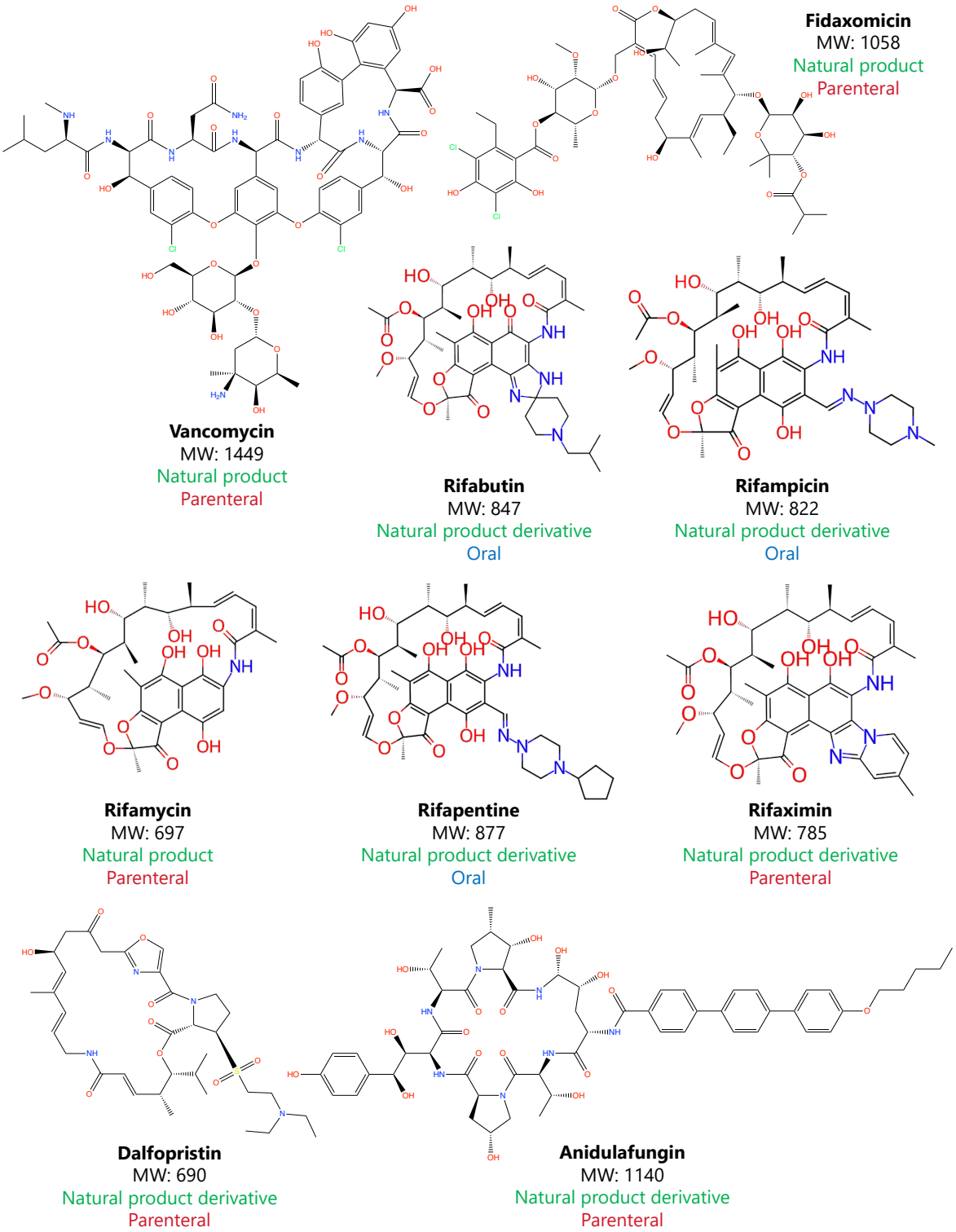


Figure A.1: (continuation)

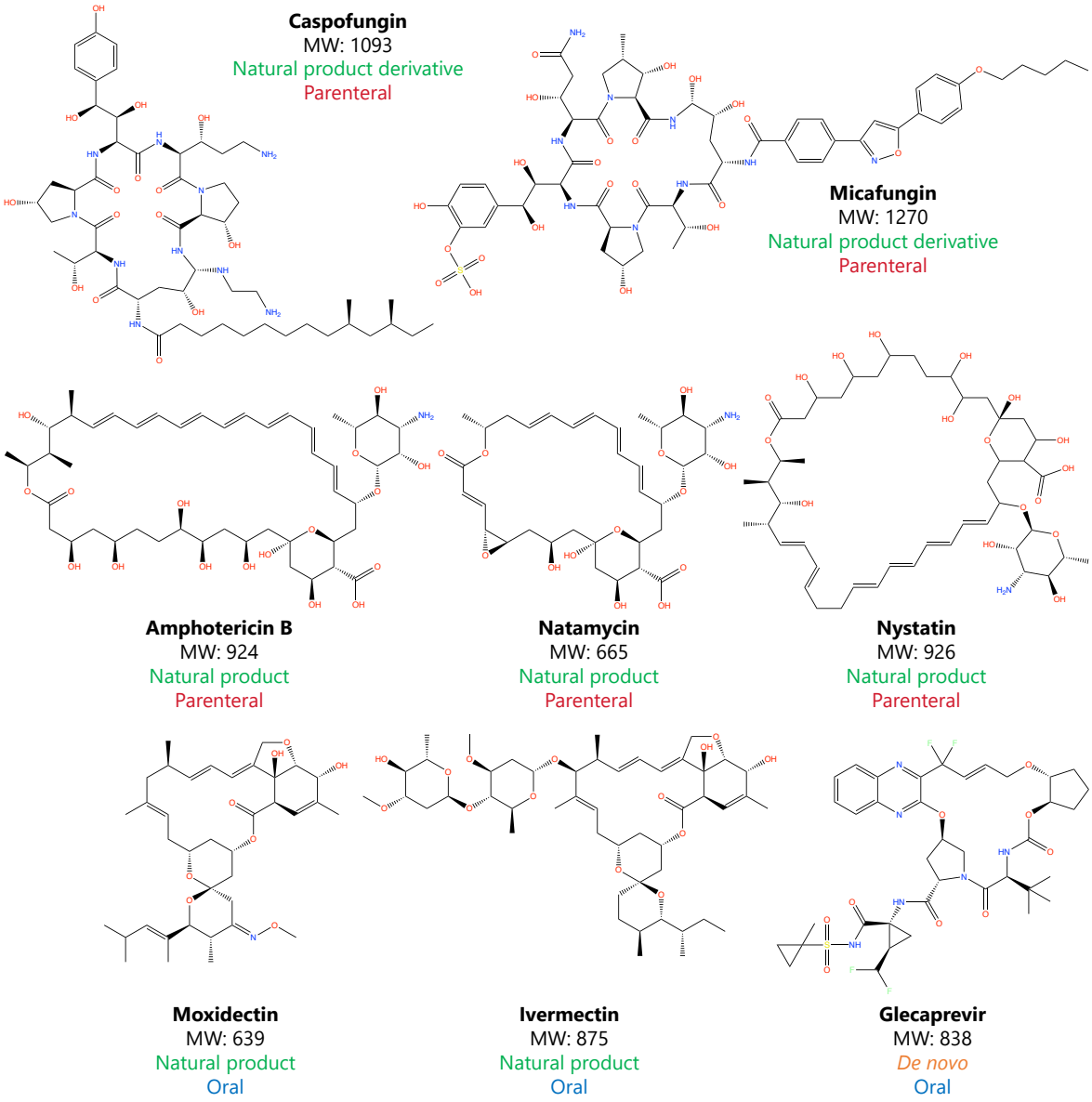
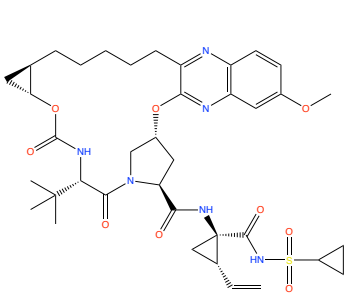
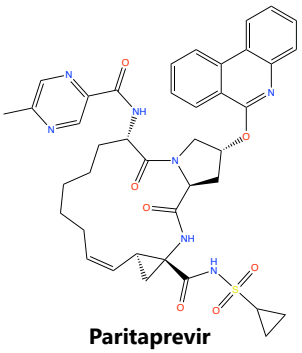


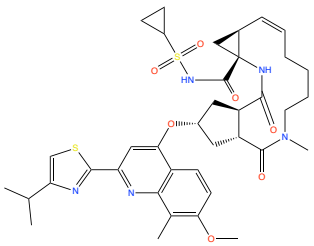
Figure A.1: (continuation)



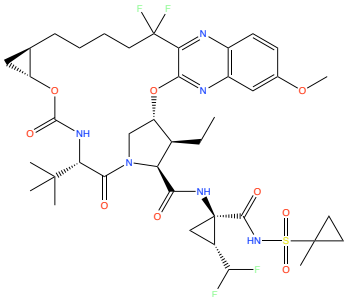
Grazoprevir
MW: 766
De novo
Oral



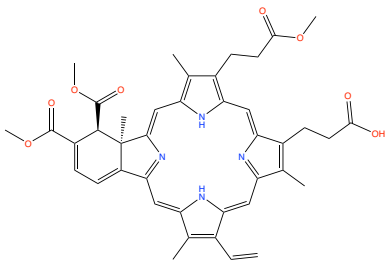
Paritaprevir
MW: 765
De novo
Oral



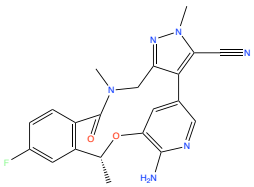
Simeprevir
MW: 749
De novo
Oral



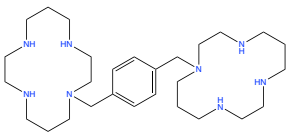
Voxilaprevir
MW: 868
De novo
Oral



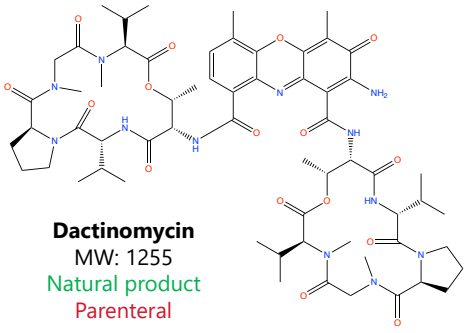
Verteporfin
MW: 718
Natural product derivative
Parenteral



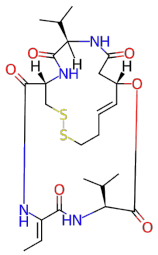
Lorlatinib
MW: 406
De novo
Oral



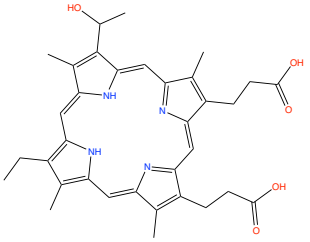
Plerixafor
MW: 502
De novo
Parenteral



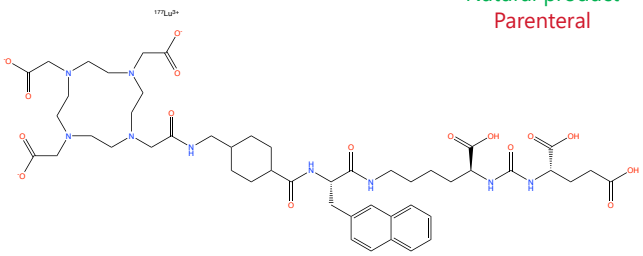
Dactinomycin
MW: 1255
Natural product
Parenteral



Romidepsin
MW: 540
Natural product
Parenteral



Porfimer sodium
MW: 605
Natural product derivative
Parenteral



Lutetium Lu-177 Vipivotide Tetraxetan
MW: 1216
Natural product derivative
Parenteral

Figure A.1: (continuation)

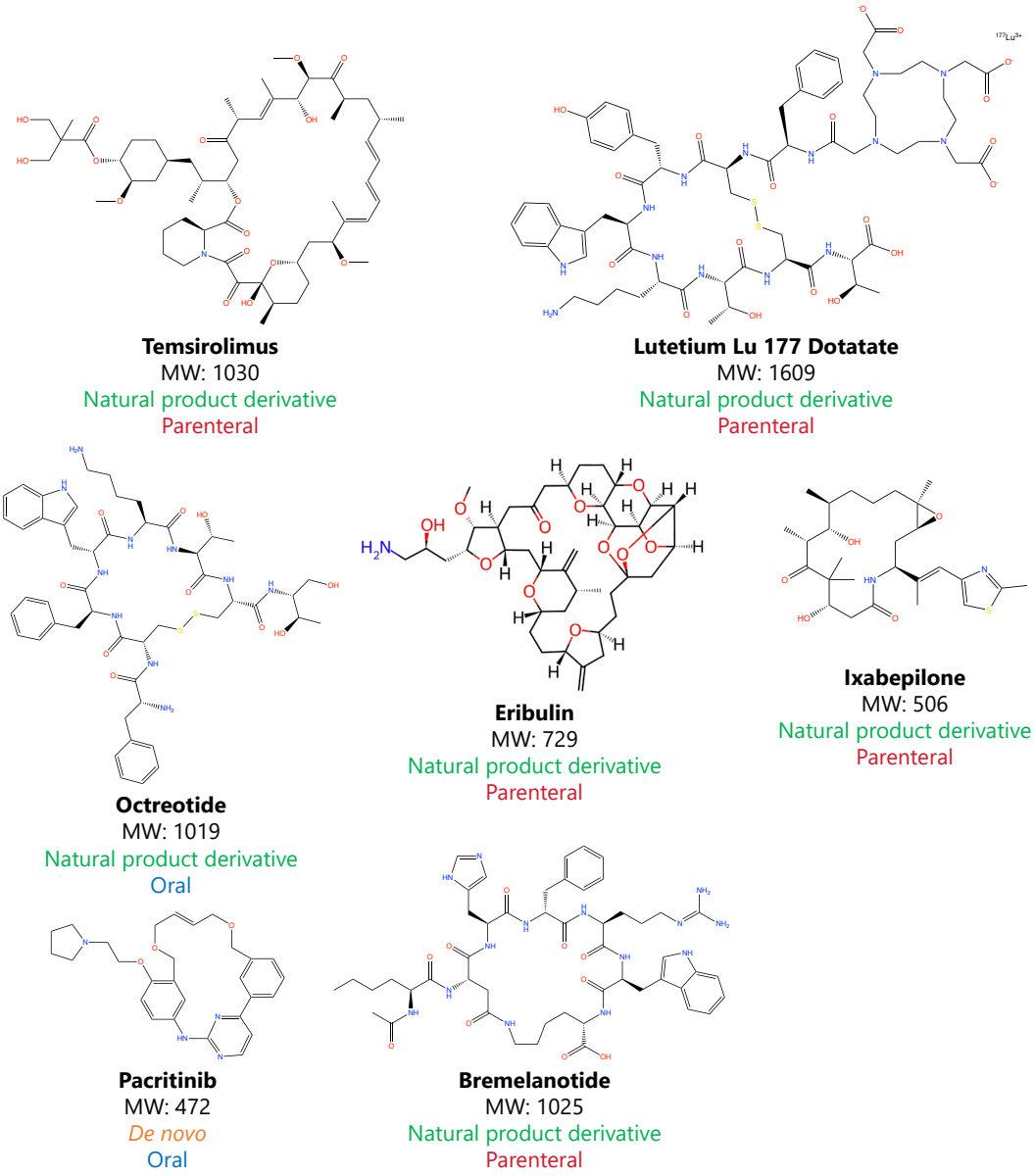
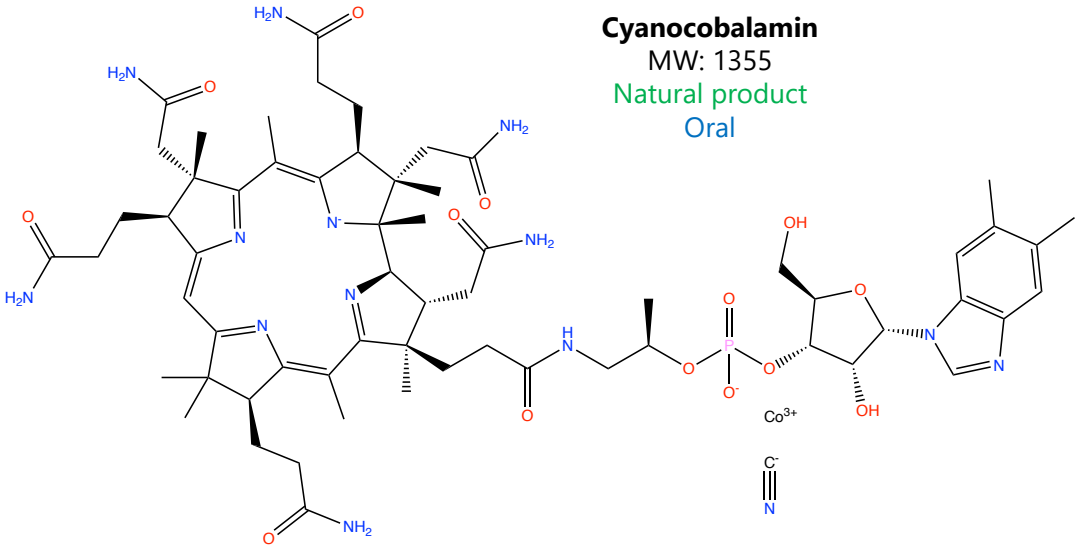


Figure A.1: (continuation)



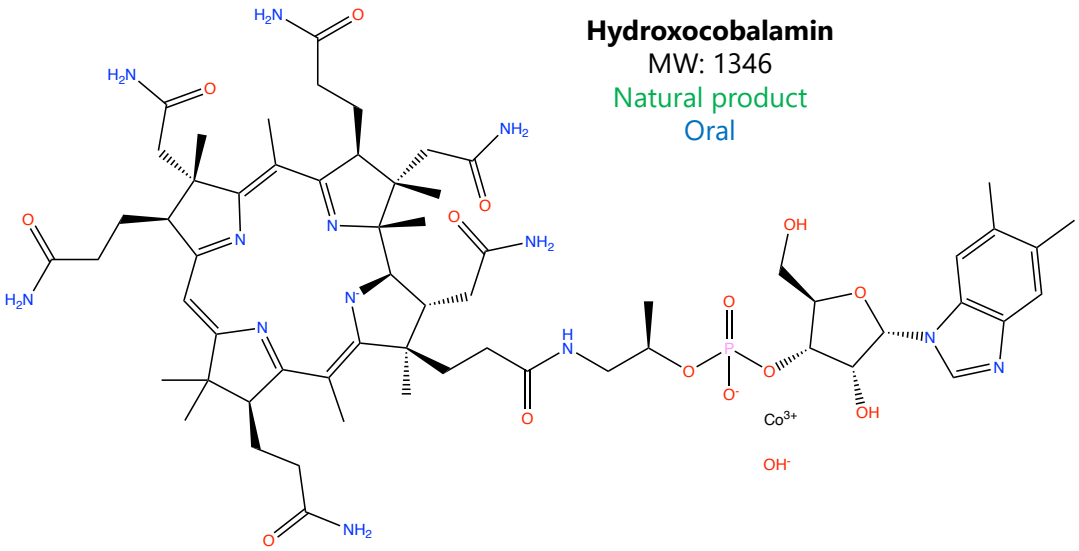


Figure A.1: (continuation)

Appendix B

Full List of Calculated $1,857$ Descriptors

The full list and descriptions of the Mordred descriptors can be found in `https://mordred-descriptor.github.io/documentation/master/descriptors.html`. The full list of the MOE descriptors is shown in Table B.1 and Table B.2. The full list of the RDkit descriptors is shown in Table B.3.

Table B.1: Full list of calculated 206 MOE 2D descriptors.

Type	Name	Type	Name	Type	Name
Physical Properties	apol	Atom Counts and Bond Counts	SMR_VSA0	Kier & Hall Connectivity and Kappa Shape Indices	lip_violation
	bpol		SMR_VSA1		nmol
	density		SMR_VSA2		opr_brigid
	FCharge		SMR_VSA3		opr_leadlike
	mr		SMR_VSA4		opr_nring
	SMR		SMR_VSA5		opr_nrot
	Weight		SMR_VSA6		opr_violation
	logP(o/w)		SMR_VSA7		ast_violation
	SlogP		a_aro		ast_violation_ext
	logS		a_count		ast_fraglike
	mutagenic		a_heavy		ast_fraglike_ext
	reactive		a_ICM		rings
	rsynth		a_IC		VAdjMa
	TPSA		a_nH		VAdjEq
	vdw_vol		a_nB		chi0
	vdw_area		a_nC		chi0_C
Hueckel Theory Descriptors	h_ema		a_nN		chi1
	h_emd		a_nO		chi1_C
	h_emd_C		a_nF		chi0v
	h_log_pbo		a_nP		chi0v_C
	h_log_dbo		a_nS		chi1v
	h_mr		a_nCl		chi1v_C
	h_logP		a_nBr		Kier1
	h_logS		a_nI		Kier2
	h_pavgQ		b_1rotN		Kier3
	h_pstates		b_1rotR		KierA1
	h_pstrain		b_ar		KierA2
	h_pKa		b_count		KierA3
	h_pKb		b_double		KierFlex
	h_logD		b_heavy		zagreb
Subdivided Surface Areas	SlogP_VSA0		b_rotN	Adjacency and Distance Matrix Descriptors	balabanJ
	SlogP_VSA1		b_rotR		BCUT_PEOE_0
	SlogP_VSA2		b_single		BCUT_PEOE_1
	SlogP_VSA3		b_max1len		BCUT_PEOE_2
	SlogP_VSA4		b_triple		BCUT_PEOE_3
	SlogP_VSA5		chiral		BCUT_SLOGP_0
	SlogP_VSA6		chiral_u		BCUT_SLOGP_1
	SlogP_VSA7		lip_acc		BCUT_SLOGP_2
	SlogP_VSA8		lip_don		BCUT_SLOGP_3
	SlogP_VSA9		lip_druglike		BCUT_SMR_0

Table B.1: (continuation)

Type	Name
Adjacency and Distance Matrix Descriptors	BCUT_SMR_1
	BCUT_SMR_2
	BCUT_SMR_3
	diameter
	petitjean
	GCUT_PEOE_0
	GCUT_PEOE_1
	GCUT_PEOE_2
	GCUT_PEOE_3
	GCUT_SLOGP_0
	GCUT_SLOGP_1
	GCUT_SLOGP_2
	GCUT_SLOGP_3
	GCUT_SMR_0
	GCUT_SMR_1
	GCUT_SMR_2
	GCUT_SMR_3
	petitjeanSC
	radius
	VDistEq
	VDistMa
	weinerPath
	weinerPol
Pharmacophore Feature Descriptors	a_acc
	a_acid
	a_base
	a_don
	a_donacc
	a_hyd
	vsa_acc
	vsa_acid
	vsa_base
	vsa_don
	vsa_hyd
	vsa_other
	vsa_pol
Partial Charge Descriptors	PC+
	Q_PC+
	PEOE_PC+
	PC-
Partial Charge Descriptors	Q_PC-
	PEOE_PC-
	RPC+
	Q_RPC+
	PEOE_RPC+
	RPC-
	Q_RPC-
	PEOE_RPC-
	Q_VSA_POS
	PEOE_VSA_POS
	Q_VSA_NEG
	PEOE_VSA_NEG
	Q_VSA_PPOS
	PEOE_VSA_PPOS
	Q_VSA_PNEG
	PEOE_VSA_PNEG
	Q_VSA_HYD
	PEOE_VSA_HYD
	Q_VSA_POL
	PEOE_VSA_POL
	Q_VSA_FPOS
	PEOE_VSA_FPOS
	Q_VSA_FNEG
	PEOE_VSA_FNEG
	Q_VSA_FPPOS
	PEOE_VSA_FPPOS
	Q_VSA_FPNEG
	PEOE_VSA_FPNEG
	Q_VSA_FHYD
	PEOE_VSA_FHYD
	Q_VSA_FPOL
	PEOE_VSA_FPOL
	PEOE_VSA+6
	PEOE_VSA+5
	PEOE_VSA+4
	PEOE_VSA+3
	PEOE_VSA+2
	PEOE_VSA+1
	PEOE_VSA+0
	PEOE_VSA-0
Partial Charge Descriptors	PEOE_VSA-1
	PEOE_VSA-2
	PEOE_VSA-3
	PEOE_VSA-4
	PEOE_VSA-5
	PEOE_VSA-6

Table B.2: Full list of calculated 117 MOE 3D descriptors.

Type	Name	Type	Name	Type	Name
Potential Energy Descriptors	E	Surface Area, Volume and Shape Descriptors	vsurf_IW3	Conformation Dependent Charge Descriptors	vsurf_DD13
	E_ang		vsurf_IW4		vsurf_DD23
	E_ele		vsurf_IW5		vsurf_HL1
	E_nb		vsurf_IW6		vsurf_HL2
	E_oop		vsurf_IW7		vsurf_A
	E_sol		vsurf_IW8		vsurf_CP
	E_stb		vsurf_CW1		vsurf_Wp1
	E_str		vsurf_CW2		vsurf_Wp2
	E_strain		vsurf_CW3		vsurf_Wp3
	E_tor		vsurf_CW4		vsurf_Wp4
	E_vdw		vsurf_CW5		vsurf_Wp5
Surface Area, Volume and Shape Descriptors	ASA		vsurf_CW6		vsurf_Wp6
	dens		vsurf_CW7		vsurf_Wp7
	glob		vsurf_CW8		vsurf_Wp8
	pmi		vsurf_EWmin1		vsurf_HB1
	pmi1		vsurf_EWmin2		vsurf_HB2
	pmi2		vsurf_EWmin3		vsurf_HB3
	pmi3		vsurf_DW12		vsurf_HB4
	npr1		vsurf_DW13		vsurf_HB5
	npr2		vsurf_DW23		vsurf_HB6
	rgyr		vsurf_D1		vsurf_HB7
	std_dim1		vsurf_D2		vsurf_HB8
	std_dim2		vsurf_D3		ASA+
	std_dim3		vsurf_D4		ASA-
	vol		vsurf_D5		ASA_H
	VSA		vsurf_D6		ASA_P
	vsurf_V		vsurf_D7		DASA
	vsurf_S		vsurf_D8		CASA+
	vsurf_R		vsurf_ID1		CASA-
	vsurf_G		vsurf_ID2		DCASA
	vsurf_W1		vsurf_ID3		dipole
	vsurf_W2		vsurf_ID4		FASA+
	vsurf_W3		vsurf_ID5		FASA-
	vsurf_W4		vsurf_ID6		FCASA+
	vsurf_W5		vsurf_ID7		FCASA-
	vsurf_W6		vsurf_ID8		FASA_H
	vsurf_W7		vsurf_EDmin1		FASA_P
	vsurf_W8		vsurf_EDmin2		
	vsurf_IW1		vsurf_EDmin3		
	vsurf_IW2		vsurf_DD12		

Table B.3: Full list of calculated 208 RDKit 2D descriptors.

MaxEStateInde	PEOE_VSA3	HeavyAtomCount	fr_barbitur
MinEStateIndex	PEOE_VSA4	NHOHCount	fr_benzene
MaxAbsEStateIndex	PEOE_VSA5	NOCount	fr_benzodiazepine
MinAbsEStateIndex	PEOE_VSA6	NumAliphaticCarbocycles	fr_bicyclic
qed	PEOE_VSA7	NumAliphaticHeterocycles	fr_diazo
MolWt	PEOE_VSA8	NumAliphaticRings	fr_dihydropyridine
HeavyAtomMolWt	PEOE_VSA9	NumAromaticCarbocycles	fr_epoxide
ExactMolWt	SMR_VSA1	NumAromaticHeterocycles	fr_ester
NumValenceElectrons	SMR_VSA10	NumAromaticRings	fr_ether
NumRadicalElectrons	SMR_VSA2	NumHAcceptors	fr_furan
MaxPartialCharge	SMR_VSA3	NumHDonors	fr_guanido
MinPartialCharge	SMR_VSA4	NumHeteroatoms	fr_halogen
MaxAbsPartialCharge	SMR_VSA5	NumRotatableBonds	fr_hdrzine
MinAbsPartialCharge	SMR_VSA6	NumSaturatedCarbocycles	fr_hdrzone
FpDensityMorgan1	SMR_VSA7	NumSaturatedHeterocycles	fr_imidazole
FpDensityMorgan2	SMR_VSA8	NumSaturatedRings	fr_imide
FpDensityMorgan3	SMR_VSA9	RingCount	fr_isocyan
BCUT2D_MWHI	SlogP_VSA1	MolLogP	fr_isothiocyan
BCUT2D_MWLOW	SlogP_VSA10	MolMR	fr_ketone
BCUT2D_CHGHI	SlogP_VSA11	fr_Al_COO	fr_ketone_Topliss
BCUT2D_CHGLO	SlogP_VSA12	fr_Al_OH	fr_lactam
BCUT2D_LOGPHI	SlogP_VSA2	fr_Al_OH_noTert	fr_lactone
BCUT2D_LOGPLOW	SlogP_VSA3	fr_ArN	fr_methoxy
BCUT2D_MRHI	SlogP_VSA4	fr_Ar_COO	fr_morpholine
BCUT2D_MRLOW	SlogP_VSA5	fr_Ar_N	fr_nitrile
BalabanJ	SlogP_VSA6	fr_Ar_NH	fr_nitro
BertzCT	SlogP_VSA7	fr_Ar_OH	fr_nitro_arom
Chi0	SlogP_VSA8	fr_COO	fr_nitro_arom_nonortho
Chi0n	SlogP_VSA9	fr_COO2	fr_nitroso
Chi0v	TPSA	fr_C_O	fr_oxazole
Chi1	EState_VSA1	fr_C_O_noCOO	fr_oxime
Chi1n	EState_VSA10	fr_C_S	fr_para_hydroxylation
Chi1v	EState_VSA11	fr_HOCCN	fr_phenol
Chi2n	EState_VSA2	fr_Imine	fr_phenol_noOrthoHbond
Chi2v	EState_VSA3	fr_NH0	fr_phos_acid
Chi3n	EState_VSA4	fr_NH1	fr_phos_ester
Chi3v	EState_VSA5	fr_NH2	fr_piperdine
Chi4n	EState_VSA6	fr_N_O	fr_piperzine
Chi4v	EState_VSA7	fr_Ndealkylation1	fr_priamide
HallKierAlpha	EState_VSA8	fr_Ndealkylation2	fr_prisulfonamd
Ipc	EState_VSA9	fr_Nhpyrrole	fr_pyridine
Kappa1	VSA_EState1	fr_SH	fr_quatN
Kappa2	VSA_EState10	fr_aldehyde	fr_sulfide
Kappa3	VSA_EState2	fr_alkyl_carbamate	fr_sulfonamd
LabuteASA	VSA_EState3	fr_alkyl_halide	fr_sulfone
PEOE_VSA1	VSA_EState4	fr_allylic_oxid	fr_term_acetylene
PEOE_VSA10	VSA_EState5	fr_amide	fr_tetrazole
PEOE_VSA11	VSA_EState6	fr_amidine	fr_thiazole
PEOE_VSA12	VSA_EState7	fr_aniline	fr_thiocyan
PEOE_VSA13	VSA_EState8	fr_aryl_methyl	fr_thiophene
PEOE_VSA14	VSA_EState9	fr_azide	fr_unbrch_alkane
PEOE_VSA2	FractionCSP3	fr_azo	fr_urea

Appendix C

Development of a Comprehensive Database of Membrane Permeability of Cyclic Peptides (CycPeptMPDB)

C.1 Overview of CycPeptMPDB Framework

As shown in Fig. C.1, CycPeptMPDB is a comprehensive database recording the membrane permeability of cyclic peptides based on data obtained from published papers and pharmaceutical patents. It mainly contains two types of data for cyclic peptides: (1) property information, i.e., experimental values of membrane permeability and physical quantities such as LogP (an index of lipophilicity) estimated from chemical structure, and (2) chemical structure information, i.e., sequence information described by HELM and monomers as partial structures constituting the cyclic peptides. CycPeptMPDB provides several functions, such as data storage, statistics and visualization, and searching and analysis.

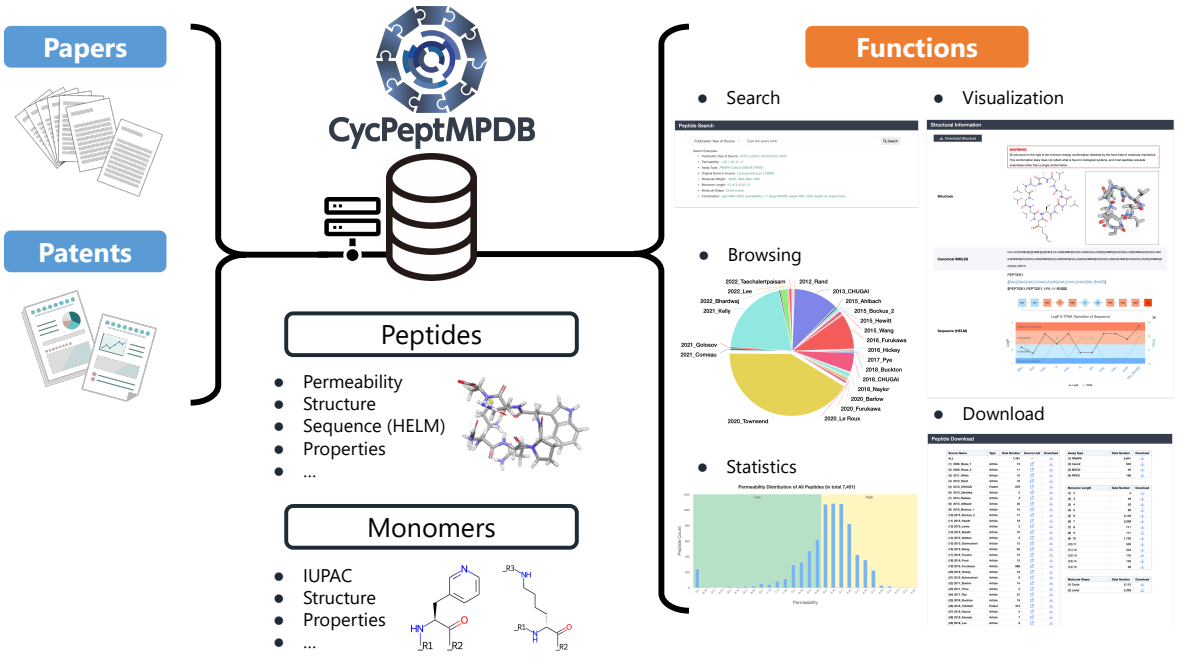


Figure C.1: Basic framework of CycPeptMPDB. CycPeptMPDB data were collected from published papers and patents of pharmaceutical companies and then manually inspected. Information in various formats was deposited into a PostgreSQL-based database for various web-based functions.

C.2 3D Structure Generation of Cyclic Peptides

Chemical structural information of the collected cyclic peptides was recorded in SMILES notation. Additionally, as conformation generation for cyclic peptides is computationally expensive, we generated the 3D structure of each cyclic peptide using RDKit software (version 2020.09.1), allowing users to quickly start relevant research. We generated 5,000 conformations per peptide (with imposing macrocycle torsion angle preferences, `useMacrocycleTorsions=True`) and removed redundant conformations with RMSD less than 1.0 Å. Next, the structure optimization of each conformation was performed using the UFF force field, and the top structure with the lowest potential energy was selected. This approach provided a computationally efficient way to obtain the 3D structure of cyclic peptides. However, it should be noted that the minimum energy conformation obtained by molecular mechanics force fields may not necessarily reflect the true conformations of the peptides in biological systems. Furthermore, most peptides are likely to populate conformational ensembles rather than a single conformation. The 3D structure of the cyclic peptide can be viewed online and downloaded in SDF format.

C.3 Introduction of Web Page Functions

C.3.1 Peptide browsing function

As mentioned in Section 4.2, CycPeptMPDB includes 7,334 structurally diverse cyclic peptides (the number including duplicated structures from all publications was 7,451) from 45 papers and 2 pharmaceutical company patents. As shown in Table 4.1, only 6 publications reported more than 100 peptides, 2020_Townsend[163] with 3,086 peptides accounting for more than 40% of the total. In addition, when browsing peptides, we prepared three classification methods in addition to browsing by Data Source: Assay Type, Monomer Length, and Molecule Shape (Fig. C.2 (A)). When classified by monomer length (peptide sequence length), according to the monomer splitting method used in this study (cleavage of peptide and ester bond), the monomer length ranged from 2 to 15. Furthermore, the behavior of side chains of cyclic peptides and the formation of hydrogen bonds between side chains and the main chain can have a significant impact on membrane permeability; therefore, it may be necessary to separate the treatment of Circle and Lariat peptides. Circle peptides accounted for under 70% (5,115) of the total and Lariat peptides for about 30% (2,336). Moreover, detailed information for the source can also be accessed if browsed by Data Source, as shown in Fig. C.2 (B). After navigating to the corresponding subset list page, the brief table of peptides displays basic information of peptides, including: CycPeptMPDB ID, 2D structure image, HELM, permeability, molecular weight, monomer length, and LogP (Fig. C.2 (B)). If users want to refine the list of accessed peptides further, the search function on the upper right of the table can be used. This search function differs from the search function described in Section C.3.2 in that it can filter peptides that partially match the contents of the table (data source name, publication year of source, original name in source, and molecule shape are provided in addition to the table contents).

C.3.2 Peptide search function for quick data retrieval

In addition to peptide browsing, users can use the peptide search function to quickly find their target peptides. The search module supports conditional searches for peptides by seven options and their logical combinations. These options include publication year of source, permeability, assay type, original compound name in source, molecular weight, monomer length, and molecule shape. Numeric options such as permeability can also be searched by range; see CycPeptMPDB usage for specific search examples and detailed instructions. In addition to the homepage, users can also use the peptide

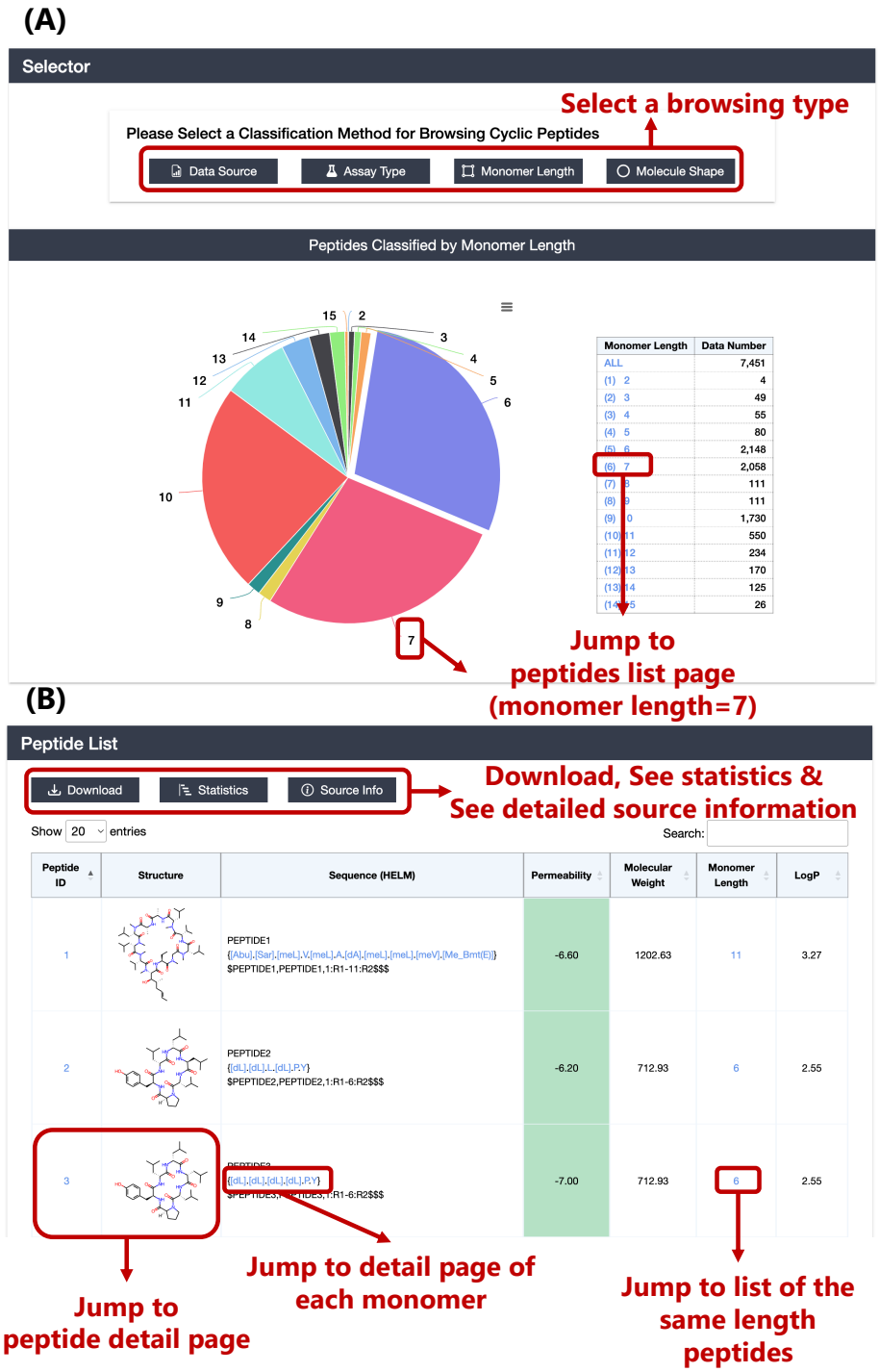


Figure C.2: (A) Classification method selection for browsing peptides and browsing page. The case when Monomer Length is selected is shown as an example. (B) Peptide's list page. The background color of the permeability cell is yellow when the permeability is High ($\mathrm{LogP}_{\mathrm{exp}} \geq -6.00$) and green when it is Low ($\mathrm{LogP}_{\mathrm{exp}} < -6.00$).

search function from the search box in the upper right corner of each page.

C.3.3 Visualization functions on peptide detail page

We incorporated several useful functions in the peptide detail page for peptides and monomers. First, for peptides of the same structure reported in multiple sources, we listed all published membrane permeability measurements in the peptide information section (Fig. C.3 (A)). In addition to measurements from different assays of the same source, the measured membrane permeabilities between each source are also different. This function allows users to quickly select the measured membrane permeability values obtained under different measurement environments. Because the number of 3D structures that cyclic peptides can take is enormous, generating 3D structures requires a large amount of computational resources. Therefore, to facilitate the use of CycPeptMPDB, we generated 5,000 conformations per peptide with RDKit software as described in Section C.2. The most stable single conformation was selected and stored (Fig. C.3 (B)). Finally, to increase the readability of the HELM representation and support sequence-based analysis, we also created HELM image and LogP and TPSA transition diagrams for the sequence (Fig. C.3 (B)). Using these functions, users can quickly capture the change in peptide sequence and partial characterization.

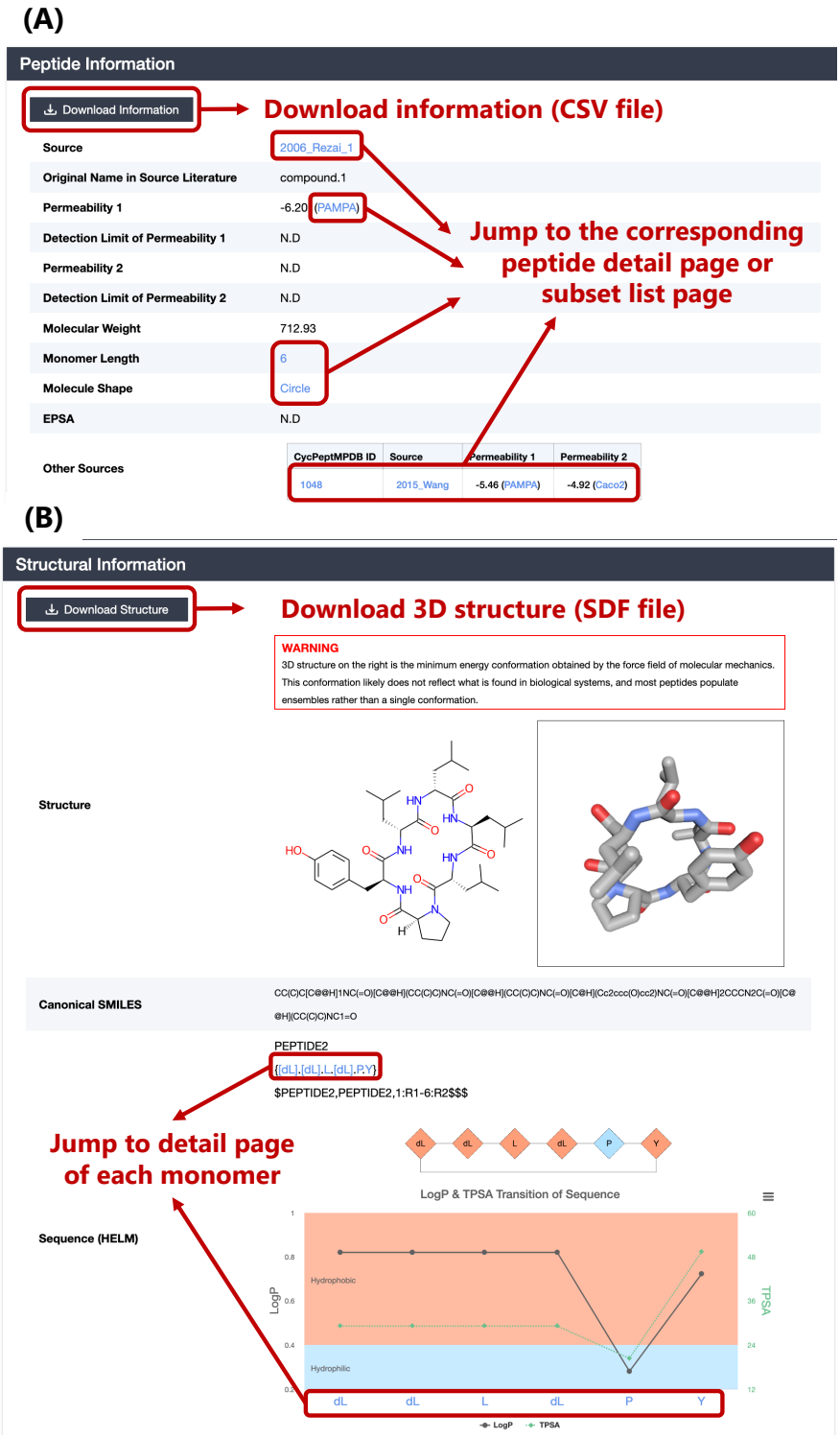


Figure C.3: (A) Peptide information section of the peptide detail page. (B) The structural information section of the peptide detail page. HELM images and LogP transition diagrams are colored by the LogP value of each monomer.

C.3.4 Browsing and visualization functions of monomers

A total of 312 monomers were defined as substructures that comprise the peptides, and they were classified into 21 categories by their natural analog (20 natural amino acids and unknown (X)). Natural analogs were established by referring to the description of each monomer in PubChem and the monomer library of ChEMBL. Among these 21 categories, categories F (38) and G (38) included the most monomers, and there were two other categories with more than 20 monomers: A (25) and S (26). We provided a browsing function for monomers by natural analog (Fig. C.4 (A)). After navigating to the corresponding subset list page, the brief table displays the basic information of monomers, including symbol, 2D structure image, monomer type (Backbone or Terminal), natural analog, attachment points (R1–R3), molecular weight, and LogP (Fig. C.4 (B)). Next, as shown in Fig. C.5 (A), we included the PubChem CID of the monomer and created a link to PubChem in the monomer detail page. Users can obtain more diverse information on the monomer from PubChem. Moreover, the monomer detail page lists the distribution of the number of peptides containing each monomer and the membrane permeability distribution of these peptides (Fig. C.5 (B)). This function will assist users in performing monomer-level analysis.

C.4 Data and Software Availability

All information recorded in CycPeptMPDB can be downloaded from `http://cycpeptmpdb.com/download/`. The structure and membrane permeability of all cyclic peptides recorded in CycPeptMPDB was collected from published papers and patents. The list of source publications is shown in Table 4.1 or `http://cycpeptmpdb.com/resources/statistics/`. The implementations of CycPeptMPDB used Docker (`https://www.docker.com/`). All data were stored in a PostgreSQL-based database and managed by pgAdmin4 (version 6.14, `https://www.pgadmin.org/`). The website was implemented by Django (version 3.2, `https://www.djangoproject.com/`), a high-level web framework with Python (version 3.8.3). The web page was constructed using HTML, CSS, and JavaScript; dynamic chart visualization was performed using Highcharts (`https://www.highcharts.com/`), and 3D structures were presented using ChemDoodle Web Components (`https://web.chemdoodle.com/`). In addition, RDKit software (version 2020.09.1, `https://www.rdkit.org/`) was used for 3D structure generation of cyclic peptides, descriptor calculation of cyclic peptides and monomers, and 2D structures image gen-

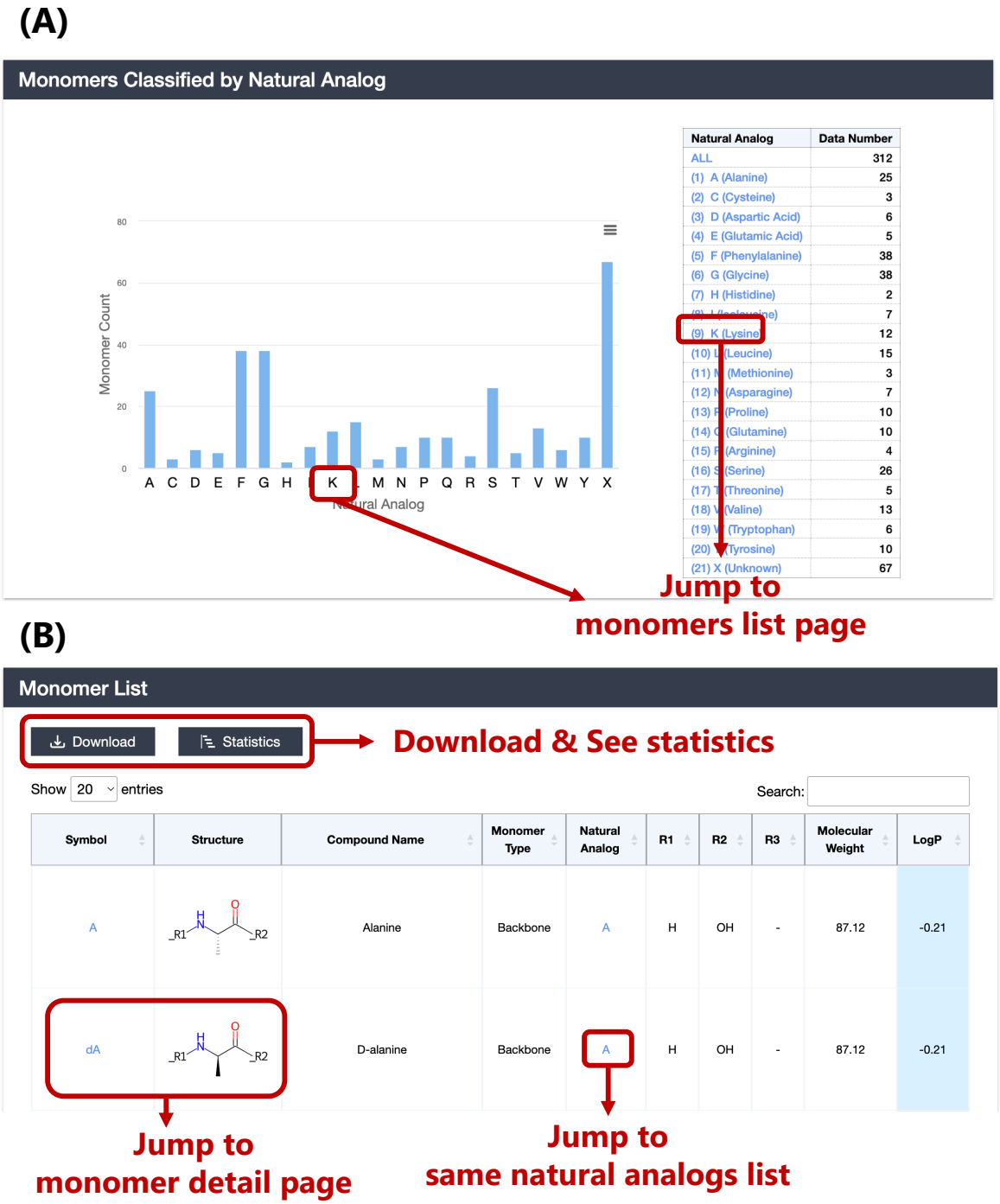


Figure C.4: Monomer (A) browsing and (B) list pages. LogP cell background color is blue when LogP is Extremely Hydrophilic (LogP < -0.60), light blue when Hydrophilic ($-0.60 \leq$ LogP < 0.40), orange when Hydrophobic ($0.40 \leq$ LogP < 1.40), and red when Extremely Hydrophobic ($1.40 \leq$ LogP).

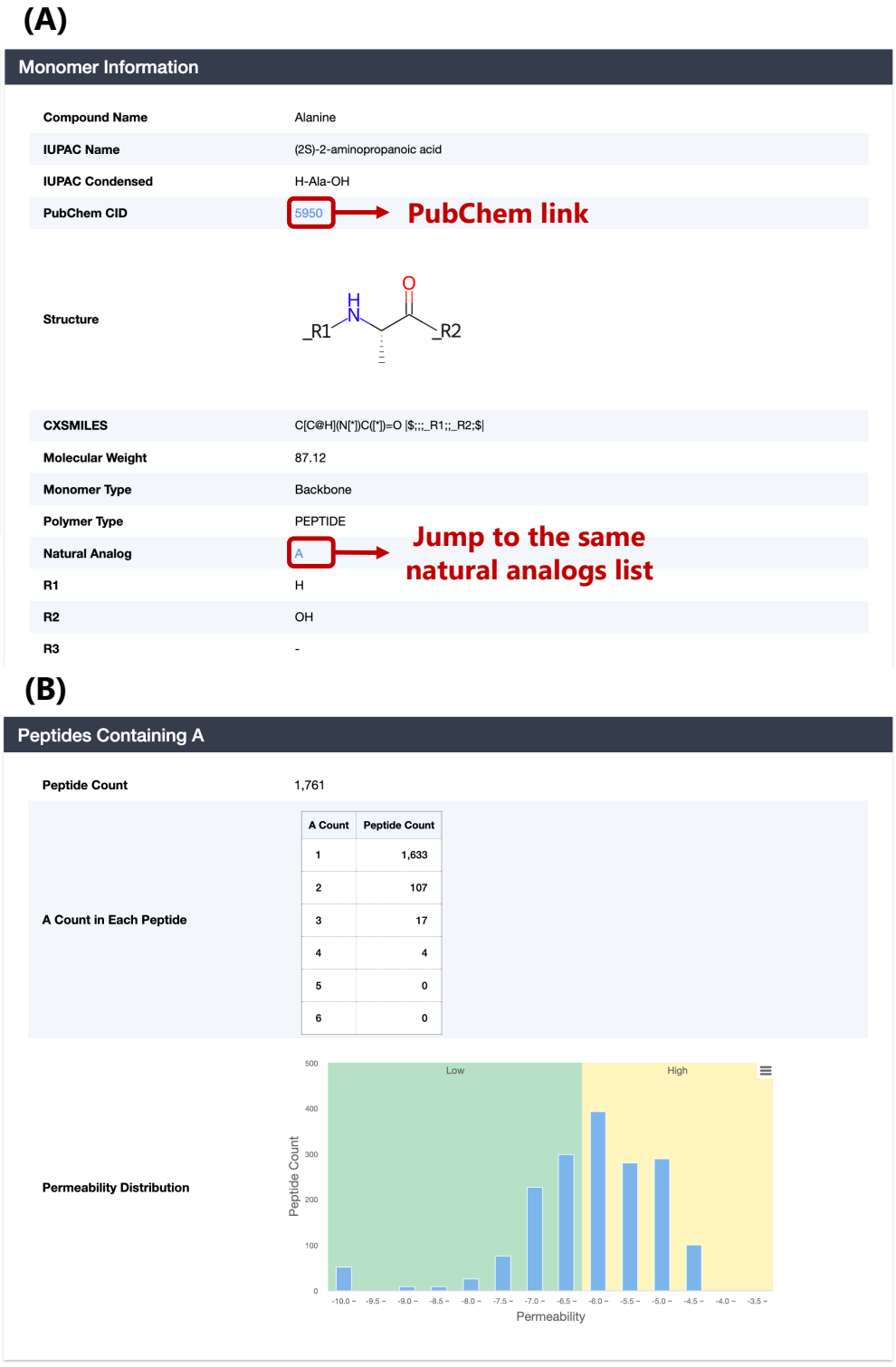


Figure C.5: (A) Monomer information section of the monomer detail page. (B) Statistics section of peptides containing current monomer.

eration of cyclic peptides and monomers. Furthermore, as mentioned in the Methods section, the IUPAC names of monomers referred to the PubChem database (`https://pubchem.ncbi.nlm.nih.gov/`) were included, and some of them were generated by the STOUT software (version 2.0, `https://github.com/Kohulan/Smiles-TO-iUpac-Translator`) [171].

References

[1] Steven M Paul, Daniel S Mytelka, Christopher T Dunwiddie, Charles C Persinger, Bernard H Munos, Stacy R Lindborg, and Aaron L Schacht. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discov.*, 9(3):203–214, 2010.

[2] Duxin Sun, Wei Gao, Hongxiang Hu, and Simon Zhou. Why 90% of clinical drug development fails and how to improve it? *Acta Pharm. Sin. B.*, 12(7):3049–3062, 2022.

[3] Teresa A F Cardote and Alessio Ciulli. Cyclic and macrocyclic peptides as chemical tools to recognise protein surfaces and probe protein–protein interactions. *ChemMedChem*, 11(8):787–794, 2016.

[4] Alexander A Vinogradov, Yizhen Yin, and Hiroaki Suga. Macrocyclic Peptides as Drug Candidates: Recent Progress and Remaining Challenges. *J. Am. Chem. Soc.*, 141(10):4167–4181, 2019.

[5] Diego Garcia Jimenez, Vasanthanathan Poongavanam, and Jan Kihlberg. Macrocycles in drug discovery—learning from the past for the future. *J. Med. Chem.*, 66(8):5377–5396, 2023.

[6] Patrick G Dougherty, Ashweta Sahni, and Dehua Pei. Understanding cell penetration of cyclic peptides. *Chem. Rev.*, 119(17):10241–10287, 2019.

[7] Tina R White, Chad M Renzelman, Arthur C Rand, Taha Rezai, Cayla M McEwen, Vladimir M Gelev, Rushia A Turner, Roger G Linington, Siegfried S F Leung, Amit S Kalgutkar, Jonathan N Bauman, Yizhong Zhang, Spiros Liras, David A Price, Alan M Mathiowetz, Matthew P Jacobson, and R Scott Lokey. On-resin N-methylation of cyclic peptides for discovery of orally bioavailable scaffolds. *Nat. Chem. Biol.*, 7(11):810–817, 2011.

[8] John R Frost, Conor C G Scully, and Andrei K Yudin. Oxadiazole grafts in peptide macrocycles. *Nat. Chem.*, 8(12):1105–1111, 2016.

[9] Yuki Hosono, Satoshi Uchida, Moe Shinkai, Chad E Townsend, Colin N Kelly, Matthew R Naylor, Hsiau-wei Lee, Kayoko Kanamitsu, Mayumi Ishii, Ryosuke Ueki, Takumi Ueda, Koh Takeuchi, Masatake Sugita, Yutaka Akiyama, R Scott Lokey,

Jumpei Morimoto, and Shinsuke Sando. Amide-to-ester substitution as a stable alternative to N-methylation for increasing membrane permeability in cyclic peptides. *Nat. Commun.*, 14(1):1416, 2023.

[10] Pritha Ghosh, Nishant Raj, Hitesh Verma, Monika Patel, Sohini Chakraborti, Bhavesh Khatri, Chandrashekar M Doreswamy, S R Anandakumar, Srinivas Seekallu, M B Dinesh, Gajanan Jadhav, Prem Narayan Yadav, and Jayanta Chatterjee. An amide to thioamide substitution improves the permeability and bioavailability of macrocyclic peptides. *Nat. Commun.*, 14(1):6050, 2023.

[11] Jaru Taechalertpaisarn, Satoshi Ono, Okimasa Okada, Timothy C Johnstone, and R Scott Lokey. A new amino acid for improving permeability and solubility in macrocyclic peptides through side chain-to-backbone hydrogen bonding. *J. Med. Chem.*, 65(6):5072–5084, 2022.

[12] Frank Hutter, Holger Hoos, and Kevin Leyton-Brown. An efficient approach for assessing hyperparameter importance. In *Proceedings of the ICML*, pages 754–762. PMLR, 2014.

[13] Elena K Schneider, Johnny X Huang, Vincenzo Carbone, Meiling Han, Yan Zhu, Sue Nang, Keith K Khoo, Johnson Mak, Matthew A Cooper, Jian Li, and Tony Velkov. Plasma protein binding structure–activity relationships related to the n-terminus of daptomycin. *ACS Infect. Dis.*, 3(3):249–258, 2017.

[14] Chugai Pharma. Co., Ltd. Peptide compound cyclization method, 2013. Patent WO2013100132A1.

[15] Akihiro Furukawa, Chad E Townsend, Joshua Schwochert, Cameron R Pye, Maria A Bednarek, and R Scott Lokey. Passive Membrane Permeability in Cyclic Peptomer Scaffolds Is Robust to Extensive Variation in Side Chain Functionality and Backbone Geometry. *J. Med. Chem.*, 59(20):9503–9512, 2016.

[16] Douglas R Cary, Masaki Ohuchi, Patrick C Reid, and Keiichi Masuya. Constrained Peptides in Drug Discovery and Development. *J. Syn. Org. Chem. Jpn.*, 75(11):1171–1178, 2017.

[17] Masatake Sugita, Satoshi Sugiyama, Takuya Fujie, Yasushi Yoshikawa, Keisuke Yanagisawa, Masahito Ohue, and Yutaka Akiyama. Large-scale membrane permeability prediction of cyclic peptides crossing a lipid bilayer based on enhanced sampling molecular dynamics simulations. *J. Chem. Inf. Model.*, 61(7):3681–3695, 2021.

[18] Markus Muttenthaler, Glenn F King, David J Adams, and Paul F Alewood. Trends in peptide drug discovery. *Nat. Rev. Drug Discov.*, 20(4):309–325, 2021.

[19] Christopher A Lipinski, Franco Lombardo, Beryl W Dominy, and Paul J Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.*, 23(1–3):3–25, 1997.

[20] Daria de Raffele and Ioana M Ilie. Unlocking novel therapies: Cyclic peptide design for amyloidogenic targets through synergies of experiments, simulations, and machine learning. *Chem. Comm.*, 60(6):632–645, 2024.

[21] Prakash Kulkarni, Supriyo Bhattacharya, Srisairam Achuthan, Amita Behal, Mohit Kumar Jolly, Sourabh Kotnala, Atish Mohanty, Govindan Rangarajan, Ravi Salgia, and Vladimir Uversky. Intrinsically disordered proteins: critical components of the wetware. *Chem. Rev.*, 122(6):6614–6633, 2022.

[22] Michelle R Arkin, Yinyan Tang, and James A Wells. Small-molecule inhibitors of protein-protein interactions: progressing toward the reality. *Chem. Biol.*, 21(9):1102–1114, 2014.

[23] Shuzhe Wang, Gerhard Konig, Hans-Jorg Roth, Marianne Fouché, Stephane Rodde, and Sereina Riniker. Effect of flexibility, lipophilicity, and the location of polar residues on the passive membrane permeability of a series of cyclic decapeptides. *J. Med. Chem.*, 64(17):12761–12773, 2021.

[24] Howook Hwang, Thom Vreven, Joël Janin, and Zhiping Weng. Protein–protein docking benchmark version 4.0. *Proteins: Struct., Funct., Bioinf.*, 78(15):3111–3114, 2010.

[25] Hiroyuki Miyachi, Kayoko Kanamitsu, Mayumi Ishii, Eri Watanabe, Akira Katsuyama, Satoko Otsuguro, Fumika Yakushiji, Mizuki Watanabe, Kouhei Matsui, Yukina Sato, Satoshi Shuto, Takashi Tadokoro, Shunsuke Kita, Takanori Matsumaru, Akira Matsuda, Tomoyasu Hirose, Masato Iwatsuki, Yasuteru Shigeta, Tetsuo Nagano, Hirotatsu Kojima, Satoshi Ichikawa, Toshiaki Sunazuka, and Katsumi Maenaka. Structure, solubility, and permeability relationships in a diverse middle molecule library. *Bioorg. Med. Chem. Lett.*, 37:127847, 2021.

[26] Jonathan C Fuller, Nicholas J Burgoyne, and Richard M Jackson. Predicting druggable binding sites at the protein–protein interface. *Drug Discov. Today*, 14(3–4):155–161, 2009.

[27] David J Craik, David P Fairlie, Spiros Liras, and David Price. The future of peptide-based drugs. *Chem. Biol. Drug Des.*, 81(1):136–147, 2013.

[28] Giuseppe Ermondi, Maura Vallaro, Gilles Goetz, Marina Shalaeva, and Giulia Caron. Updating the portfolio of physicochemical descriptors related to permeability in the beyond the rule of 5 chemical space. *Eur. J. Pharm. Sci.*, 146:105274, 2020.

[29] Susan V Onrust, Harriet M Lamb, and Julia A Barman Balfour. Rituximab. *Drugs*, 58:79–88, 1999.

[30] David Schrama, Ralph A Reisfeld, and Jürgen C Becker. Antibody targeted drugs as cancer therapeutics. *Nat. Rev. Drug Discov.*, 5(2):147–159, 2006.

[31] William R Strohl. Structure and function of therapeutic antibodies approved by the US FDA in 2023. *Antib. Ther.*, 7(2):132–156, 2024.

[32] Kohzoh Imai and Akinori Takaoka. Comparing antibody and small-molecule therapies for cancer. *Nat. Rev. Cancer.*, 6(9):714–727, 2006.

[33] Gregory L Verdine and Loren D Walensky. The challenge of drugging undruggable targets in cancer: lessons learned from targeting BCL-2 family members. *Clin. Cancer Res.*, 13(24):7264–7270, 2007.

[34] Joanna B Opalinska and Alan M Gewirtz. Nucleic-acid therapeutics: basic principles and recent applications. *Nat. Rev. Drug Discov.*, 1(7):503–514, 2002.

[35] Watshara Shoombuatong, Nalini Schaduangrat, Reny Pratiwi, and Chanin Nantasenamat. THPep: A machine learning-based approach for predicting tumor homing peptides. *Comput. Biol. Chem.*, 80:441–451, 2019.

[36] Gene Hopping, Jackson Kellock, Ravi Pratap Barnwal, Peter Law, James Bryers, Gabriele Varani, Byron Caughey, and Valerie Daggett. Designed α-sheet peptides inhibit amyloid formation by targeting toxic oligomers. *eLife*, 3:e01681, 2014.

[37] Tanishq Chamoli, Alisha Khera, Akanksha Sharma, Anshul Gupta, Sonam Garg, Kanishk Mamgain, Aayushi Bansal, Shriya Verma, Ankit Gupta, Hema K Alajangi, Gurpal Singh, and Ravi P Barnwal. Peptide utility (PU) search server: a new tool for peptide sequence search from multiple databases. *Heliyon*, 8(12):e12283, 2022.

[38] Dong In Kim, So Hee Han, Hahnbeom Park, Sehwan Choi, Mandeep Kaur, Euimin Hwang, Seong Jae Han, Jung Yeon Ryu, Hae Kap Cheong, Ravi Pratap Barnwal, and Yong Beom Lim. Pseudo-isolated α-helix platform for the recognition of deep and narrow targets. *J. Am. Chem. Soc.*, 144(34):15519–15528, 2022.

[39] Akshita Thakur, Akanksha Sharma, Hema K Alajangi, Pradeep Kumar Jaiswal, Yongbeom Lim, Gurpal Singh, and Ravi Pratap Barnwal. In pursuit of next-generation therapeutics: Antimicrobial peptides against superbugs, their sources, mechanism of action, nanotechnology-based delivery, and clinical applications. *Int. J. Biol. Macromol.*, 218:135–156, 2022.

[40] Andrei K Yudin. Macrocycles: lessons from the distant past, recent developments, and future directions. *Chem. Sci.*, 6(1):30–49, 2015.

[41] David A Price, Heather Eng, Kathleen A Farley, Gilles H Goetz, Yong Huang, Zhaodong Jiao, Amit S Kalgutkar, Natasha M Kablaoui, Bhagyashree Khunte, Spiros Liras, Chris Limberakis, Alan M Mathiowetz, Roger B Ruggeri, Jun-Min Quan, and Zhen Yang. Comparative pharmacokinetic profile of cyclosporine (CsA) with a decapeptide and a linear analogue. *Org. Biomol. Chem.*, 15(12):2501–2506, 2017.

[42] Youla S Tsantrizos. The design of a potent inhibitor of the hepatitis C virus NS3 protease: BILN 2061—from the NMR tube to the clinic. *Pept. Sci.*, 76(4):309–323, 2004.

[43] Fabio Begnini, Vasanthanathan Poongavanam, Björn Over, Marie Castaldo, Stefan Geschwindner, Patrik Johansson, Mohit Tyagi, Christian Tyrchan, Lisa Wissler, Peter Sjö, Stefan Schiesser, and Jan Kihlberg. Mining natural products for macrocycles to drug difficult targets. *J. Med. Chem.*, 64(2):1054–1072, 2020.

[44] Jamie Mallinson and Ian Collins. Macrocycles in new drug discovery. *Future Med. Chem.*, 4(11):1409–1438, 2012.

[45] Gaurav Bhardwaj, Jacob O'Connor, Stephen Rettie, Yen-Hua Huang, Theresa A Ramelot, Vikram Khipple Mulligan, Gizem Gokce Alpkilic, Jonathan Palmer, Asim K Bera, Matthew J Bick, Maddalena Di Piazza, Xinting Li, Parisa Hosseinzadeh, Timothy W Craven, Roberto Tejero, Anna Lauko, Ryan Choi, Calina Glynn, Linlin Dong, Robert Griffin, Wesley C van Voorhis, Jose Rodriguez, Lance Stewart, Gaetano T Montelione, David Craik, and David Baker. Accurate de novo design of membrane-traversing macrocycles. *Cell*, 185(19):3520–3532, 2022.

[46] Bradley C. Doak, Jie Zheng, Doreen Dobritzsch, and Jan Kihlberg. How Beyond Rule of 5 Drugs and Clinical Candidates Bind to Their Targets. *J. Med. Chem.*, 59(6):2312–2327, 2016.

[47] Huiya Zhang and Shiyu Chen. Cyclic peptide drugs approved in the last two decades (2001–2021). *RSC Chem. Biol.*, 3(1):18–31, 2022.

[48] Takatsugu Kosugi and Masahito Ohue. Design of cyclic peptides targeting protein–protein interactions Using alphaFold. *Int. J. Mol. Sci.*, 24(17):13257, 2023.

[49] Catrin Sohrabi, Andrew Foster, and Ali Tavassoli. Methods for generating and screening libraries of genetically encoded cyclic peptides in drug discovery. *Nat. Rev. Chem.*, 4(2):90–101, 2020.

[50] Yusuke Yamagishi, Ikuo Shoji, Shoji Miyagawa, Takashi Kawakami, Takayuki Katoh, Yuki Goto, and Hiroaki Suga. Natural product-like macrocyclic N-methyl-peptide inhibitors against a ubiquitin ligase uncovered from a ribosome-expressed de novo library. *Chem. Biol.*, 18(12):1562–1570, 2011.

[51] Nasir K Bashiruddin, Mikihito Hayashi, Masanobu Nagano, Yan Wu, Yukiko Matsunaga, Junichi Takagi, Tomoki Nakashima, and Hiroaki Suga. Development of cyclic peptides with potent in vivo osteogenic activity through RaPID-based affinity maturation. *Proc. Natl. Acad. Sci. USA*, 117(49):31070–31077, 2020.

[52] Ziyang Zhang, Rong Gao, Qi Hu, Hayden Peacock, D Matthew Peacock, Shizhong Dai, Kevan M Shokat, and Hiroaki Suga. GTP-state-selective cyclic peptide ligands of K-Ras (G12D) block its interaction with Raf. *ACS Cent. Sci.*, 6(10):1753–1761, 2020.

[53] Li Di and Edward H Kerns. Application of pharmaceutical profiling assays for optimization of drug-like properties. *Curr. Opin. Drug Discov. Dev.*, 8(4):495–504, 2005.

[54] Gian P Camenisch. Drug disposition classification systems in discovery and development: a comparative review of the BDDCS, ECCS and ECCCS concepts. *Pharmaceut. Res.*, 33:2583–2593, 2016.

[55] Alessandro Zorzi, Kaycie Deyle, and Christian Heinis. Cyclic peptide therapeutics: past, present and future. *Curr. Opin. Chem. Biol.*, 38:24–29, 2017.

[56] Adrian Whitty, Mengqi Zhong, Lauren Viarengo, Dmitri Beglov, David R Hall, and Sandor Vajda. Quantifying the chameleonic properties of macrocycles and other high-molecular-weight drugs. *Drug Discov. Today*, 21(5):712–717, 2016.

[57] Emma Danelius, Vasanthanathan Poongavanam, Stefan Peintner, Lianne HE Wieske, Máté Erdélyi, and Jan Kihlberg. Solution conformations explain the chameleonic behaviour of macrocyclic drugs. *Chem. Eur. J.*, 26(23):5231–5244, 2020.

[58] Dongjae Lee, Sungjin Lee, Jieun Choi, Yoo-Kyung Song, Min Ju Kim, Dae-Seop Shin, Myung Ae Bae, Yong-Chul Kim, Chin-Ju Park, Kyeong-Ryoon Lee, Jun-Ho Choi, and Jiwon Seo. Interplay among conformation, intramolecular hydrogen bonds, and chameleonicity in the membrane permeability and cyclophilin A binding of macrocyclic peptide cyclosporin O derivatives. *J. Med. Chem.*, 64(12):8272–8286, 2021.

[59] Eric Biron, Jayanta Chatterjee, Oded Ovadia, Daniel Langenegger, Joseph Brueggen, Daniel Hoyer, Herbert A Schmid, Raz Jelinek, Chaim Gilon, Amnon Hoffman, and Horst Kessler. Improving oral bioavailability of peptides by multiple N-methylation: somatostatin analogues. *Angew. Chem. Int. Ed.*, 47(14):2595–2599, 2008.

[60] Manfred Kansy, Frank Senner, and Klaus Gubernator. Physicochemical high throughput screening: parallel artificial membrane permeation assay in the description of passive absorption processes. *J. Med. Chem.*, 41(7):1007–1010, 1998.

[61] Ian Lewis, Michael Schaefer, Trixie Wagner, Lukas Oberer, Emine Sager, Peter Wipfli, and Thomas Vorherr. A detailed investigation on conformation, permeability and PK properties of two related cyclohexapeptides. *Int. J. Pept. Res. Ther.*, 21:205–221, 2015.

[62] Allen R Hilgers, Robert A Conradi, and Philip S Burton. Caco-2 cell monolayers as a model for drug transport across the intestinal mucosa. *Pharm. Res.*, 7:902–910, 1990.

[63] Jennifer D Irvine, Lori Takahashi, Karen Lockhart, Jonathan Cheong, John W Tolan, HE Selick, and J Russell Grove. MDCK (Madin–Darby canine kidney) cells: a tool for membrane permeability screening. *J. Pharm. Sci.*, 88(1):28–33, 1999.

[64] Li Di, Carrie Whitney-Pickett, John P Umland, Hui Zhang, Xun Zhang, David F Gebhard, Yurong Lai, James J Federico III, Ralph E Davidson, Russ Smith, Eric L Reyner, Caroline Lee, Bo Feng, Charles Rotter, Manthena V Varma, Sarah Kempshall, Katherine Fenner, Ayman F El-kattan, Theodore E Liston, and Matthew D Troutman. Development of a new permeability assay using low-efflux MDCKII cells. *J. Pharm. Sci.*, 100(11):4974–4985, 2011.

[65] Rei Miyamoto, Takashi Nozawa, Mayuko Kimura, Koichi Shiozuka, and Kenji Tabata. Development and validation of semiautomated 96-well transport assay using LLC-PK1 cells transfected with human P-glycoprotein for high-throughput screening. *Assay Drug Dev. Tech.*, 13(2):79–87, 2015.

[66] Conan K Wang, Susan E Northfield, Joakim E Swedberg, Barbara Colless, Stephanie Chaousis, David A Price, Spiros Liras, and David J Craik. Exploring experimental and computational markers of cyclic peptides: Charting islands of permeability. *Eur. J. Med. Chem.*, 97:202–213, 2015.

[67] George Lambrinidis, Theodosia Vallianatou, and Anna Tsantili-Kakoulidou. In vitro, in silico and integrated strategies for the estimation of plasma protein binding. A review. *Adv. Drug Deliv. Rev.*, 86:27–45, 2015.

[68] Yawen Yuan, Shuo Chang, Zheng Zhang, Zhigang Li, Size Li, Peng Xie, Wai-Ping Yau, Haishu Lin, Weimin Cai, Yanchun Zhang, and Xiaoqiang Xiang. A novel strategy for prediction of human plasma protein binding using machine learning techniques. *Chemometr. Intell. Lab. Syst.*, 199:103962, 2020.

[69] Sho Ito, Akinobu Senoo, Satoru Nagatoishi, Masahito Ohue, Masaki Yamamoto, Kouhei Tsumoto, and Naoki Wakui. Structural basis for the binding mechanism of human serum albumin complexed with cyclic peptide dalbavancin. *J. Med. Chem.*, 63(22):14045–14053, 2020.

[70] Jiunn H Lin, David M Cocchetto, and Daniel E Duggan. Protein binding as a primary determinant of the clinical pharmacokinetic properties of non-steroidal anti-inflammatory drugs. *Clin. Pharmacokinet.*, 12:402–432, 1987.

[71] Junichi Enokizono. Assessment of protein binding. *Folia Pharmacol. Jpn.*, 134(2):78–81, 2009.

[72] M Volpp and U Holzgrabe. Determination of plasma protein binding for sympathomimetic drugs by means of ultrafiltration. *Eur. J. Pharm. Sci.*, 127:175–184, 2019.

[73] Masatake Sugita, Takuya Fujie, Keisuke Yanagisawa, Masahito Ohue, and Yutaka Akiyama. Lipid composition is critical for accurate membrane permeability prediction of cyclic peptides by molecular dynamics simulations. *J. Chem. Inf. Model.*, 62(18):4549–4560, 2022.

[74] Andreas Eitel, Jost Tobias Springenberg, Luciano Spinello, Martin Riedmiller, and Wolfram Burgard. Multimodal deep learning for robust RGB-D object recognition. In *Proceedings of the 2015 IEEE/RSJ International Conference on IROS*, pages 681–687. IEEE, 2015.

[75] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, and Brian Kingsbury. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Process. Mag.*, 29(6):82–97, 2012.

[76] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andrew J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

[77] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[78] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, and Tsuhan Chen. Recent advances in convolutional neural networks. *Pattern Recognit.*, 77:354–377, 2018.

[79] Hojjat Salehinejad, Sharan Sankar, Joseph Barfett, Errol Colak, and Shahrokh Valaee. Recent advances in recurrent neural networks. *arXiv*, pages 1–21, 2017. DOI: 10.48550/arXiv.1801.01078.

[80] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. A comprehensive survey on graph neural networks. *IEEE Trans. Neural Netw. Learn. Syst.*, 32(1):4–24, 2020.

[81] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Adv. Neural. Inf. Process. Syst.*, 30, 2017.

[82] Jianan Li, Keisuke Yanagisawa, Masatake Sugita, Takuya Fujie, Masahito Ohue, and Yutaka Akiyama. CycPeptMPDB: A comprehensive database of membrane permeability of cyclic peptides. *J. Chem. Inf. Model.*, 63(7):2240–2250, 2023.

[83] Jianan Li, Keisuke Yanagisawa, and Yutaka Akiyama. CycPeptMP: enhancing membrane permeability prediction of cyclic peptides with multi-level molecular features and data augmentation. *Brief. Bioinform.*, 25(5):bbae417, 2024.

[84] Jianan Li, Keisuke Yanagisawa, Yasushi Yoshikawa, Masahito Ohue, and Yutaka Akiyama. Plasma protein binding prediction focusing on residue-level features and circularity of cyclic peptides by deep learning. *Bioinformatics*, 38(4):1110–1117, 2022.

[85] Satoshi Ono, Matthew R Naylor, Chad E Townsend, Chieko Okumura, Okimasa Okada, and R Scott Lokey. Conformation and permeability: cyclic hexapeptide diastereomers. *J. Chem. Inf. Model.*, 59(6):2952–2963, 2019.

[86] Jagna Witek, Shuzhe Wang, Benjamin Schroeder, Robin Lingwood, Andreas Dounas, Hans Jörg Roth, Marianne Fouché, Markus Blatter, Oliver Lemke, Bettina Keller, and

Sereina Riniker. Rationalization of the membrane permeability differences in a series of analogue cyclic decapeptides. *J. Chem. Inf. Model.*, 59(1):294–308, 2019.

[87] Flaviu Cipcigan, Paul Smith, Jason Crain, Anders Hogner, Leonardo De Maria, Antonio Llinas, and Ekaterina Ratkova. Membrane permeability in cyclic peptides is modulated by core conformations. *J. Chem. Inf. Model.*, 61(1):263–269, 2020.

[88] Yuji Sugita and Yuko Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.*, 314(1–2):141–151, 1999.

[89] Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *Proc. Natl. Acad. Sci. USA.*, 99(20):12562–12566, 2002.

[90] Donald Hamelberg, John Mongan, and J Andrew McCammon. Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. *J. Chem. Phys.*, 120(24):11919–11929, 2004.

[91] Nobuyuki Nakajima, Haruki Nakamura, and Akinori Kidera. Multicanonical ensemble generated by molecular dynamics simulation for enhanced conformational sampling of peptides. *J. Phys. Chem. B*, 101(5):817–824, 1997.

[92] Taha Rezai, Jonathan E Bock, Mai V Zhou, Chakrapani Kalyanaraman, R Scott Lokey, and Matthew P Jacobson. Conformational flexibility, internal hydrogen bonding, and passive membrane permeability: successful in silico prediction of the relative permeabilities of cyclic peptides. *J. Am. Chem. Soc.*, 128(43):14073–14080, 2006.

[93] Siegfried S F Leung, Daniel Sindhikara, and Matthew P Jacobson. Simple predictive models of passive membrane permeability incorporating size-dependent membrane-water partition. *J. Chem. Inf. Model.*, 56(5):924–929, 2016.

[94] Björn Over, Pär Matsson, Christian Tyrchan, Per Artursson, Bradley C Doak, Michael A Foley, Constanze Hilgendorf, Stephen E Johnston, Maurice D Lee, Richard J Lewis, Patrick McCarren, Giovanni Muncipinto, Ulf Norinder, Matthew WD Perry, Jeremy R Duvall, and Jan Kihlberg. Structural and conformational determinants of macrocycle cell permeability. *Nat. Chem. Biol.*, 12(12):1065–1074, 2016.

[95] Matteo Rossi Sebastiano, Bradley C Doak, Maria Backlund, Vasanthanathan Poongavanam, Björn Over, Giuseppe Ermondi, Giulia Caron, Pär Matsson, and Jan Kihlberg. Impact of dynamically exposed polarity on permeability and solubility of chameleonic drugs beyond the rule of 5. *J. Med. Chem.*, 61(9):4189–4202, 2018.

[96] Vito Digiesi, Víctor de la Oliva Roque, Maura Vallaro, Giulia Caron, and Giuseppe Ermondi. Permeability prediction in the beyond-Rule-of 5 chemical space: Focus on cyclic hexapeptides. *Eur. J. Pharm. Biopharm.*, 165:259–270, 2021.

[97] Vasanthanathan Poongavanam, Yoseph Atilaw, Sofie Ye, Lianne HE Wieske, Mate Erdelyi, Giuseppe Ermondi, Giulia Caron, and Jan Kihlberg. Predicting the permeability of macrocycles from conformational sampling–limitations of molecular flexibility. *J. Pharm. Sci.*, 110(1):301–313, 2021.

[98] Billy J Williams-Noonan, Melissa N Speer, Tu C Le, Maiada M Sadek, Philip E Thompson, Raymond S Norton, Elizabeth Yuriev, Nicholas Barlow, David K Chalmers, and Irene Yarovsky. Membrane Permeating Macrocycles: Design Guidelines from Machine Learning. *J. Chem. Inf. Model.*, 62(19):4605–4619, 2022.

[99] Yoshifumi Fukunishi, Tadaaki Mashimo, Takashi Kurosawa, Yoshinori Wakabayashi, Hironori K Nakamura, and Koh Takeuchi. Prediction of Passive Membrane Permeability by Semi-Empirical Method Considering Viscous and Inertial Resistances and Different Rates of Conformational Change and Diffusion. *Mol. Inform.*, 39(1–2):1900071, 2020.

[100] William M Hewitt, Siegfried S F Leung, Cameron R Pye, Alexandra R Ponkey, Maria Bednarek, Matthew P Jacobson, and R Scott Lokey. Cell-permeable cyclic peptides from synthetic libraries inspired by natural products. *J. Am. Chem. Soc.*, 137(2):715–721, 2015.

[101] Xiang-Wei Zhu, Alexander Sedykh, Hao Zhu, Shu-Shen Liu, and Alexander Tropsha. The use of pseudo-equilibrium constant affords improved QSAR models of human plasma protein binding. *Pharm. Res.*, 30:1790–1798, 2013.

[102] Brandall L Ingle, Brandon C Veber, John W Nichols, and Rogelio Tornero-Velez. Informing the human plasma protein binding of environmental chemicals by machine learning in the pharmaceutical space: applicability domain and limits of predictability. *J. Chem. Inf. Model.*, 56(11):2243–2252, 2016.

[103] Lixia Sun, Hongbin Yang, Jie Li, Tianduanyi Wang, Weihua Li, Guixia Liu, and Yun Tang. In silico prediction of compounds binding to human plasma proteins by QSAR models. *ChemMedChem*, 13(6):572–581, 2018.

[104] Reiko Watanabe, Tsuyoshi Esaki, Hitoshi Kawashima, Yayoi Natsume-Kitatani, Chioko Nagao, Rikiya Ohashi, and Kenji Mizuguchi. Predicting fraction unbound in human plasma from chemical structure: improved accuracy in the low value ranges. *Mol. Pharm.*, 15(11):5302–5311, 2018.

[105] Cosimo Toma, Domenico Gadaleta, Alessandra Roncaglioni, Andrey Toropov, Alla Toropova, Marco Marzo, and Emilio Benfenati. QSAR development for plasma protein binding: influence of the ionization state. *Pharm. Res.*, 36:1–9, 2019.

[106] Njabulo J Gumede, Parvesh Singh, Myalowenkosy I Sabela, Krishna Bisetty, Laura Escuder-Gilabert, María-José Medina-Hernández, and Salvador Sagrado. Experimental-like affinity constants and enantioselectivity estimates from flexible docking. *J. Chem. Inf. Model.*, 52(10):2754–2759, 2012.

[107] Katrina W Lexa, Elena Dolghih, and Matthew P Jacobson. A structure-based model for predicting serum albumin binding. *PLoS ONE*, 9(4):e93323, 2014.

[108] Ferenc Zsila, Zsolt Bikadi, David Malik, Peter Hari, Imre Pechan, Attila Berces, and Eszter Hazai. Evaluation of drug–human serum albumin binding interactions with

support vector machine aided online automated docking. *Bioinformatics*, 27(13):1806–1813, 2011.

[109] Haiyan Li, Zhuxi Chen, Xuejun Xu, Xiaofan Sui, Tao Guo, Wei Liu, and Jiwen Zhang. Predicting human plasma protein binding of drugs using plasma protein interaction QSAR analysis (PPI-QSAR). *Biopharm. Drug Dispos.*, 32(6):333–342, 2011.

[110] Lijuan Chen and Xin Chen. Results of molecular docking as descriptors to predict human serum albumin binding affinity. *J. Mol. Graph. Model.*, 33:35–43, 2012.

[111] Takashi Tajimi, Naoki Wakui, Keisuke Yanagisawa, Yasushi Yoshikawa, Masahito Ohue, and Yutaka Akiyama. Computational prediction of plasma protein binding of cyclic peptides from small molecule experimental data using sparse modeling techniques. *BMC Bioinform.*, 19:157–170, 2018.

[112] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. *Adv. Neural. Inf. Process. Syst.*, 28, 2015.

[113] Connor W Coley, Regina Barzilay, William H Green, Tommi S Jaakkola, and Klavs F Jensen. Convolutional embedding of attributed molecular graphs for physical property prediction. *J. Chem. Inf. Model.*, 57(8):1757–1772, 2017.

[114] Ziqiao Zhang, Jihong Guan, and Shuigeng Zhou. FraGAT: a fragment-oriented multi-scale graph attention model for molecular property prediction. *Bioinformatics*, 37(18):2981–2987, 2021.

[115] Hou Yee Choo, JunJie Wee, Cong Shen, and Kelin Xia. Fingerprint-enhanced graph attention network (FinGAT) model for antibiotic discovery. *J. Chem. Inf. Model.*, 63(10):2928–2935, 2023.

[116] Michael Withnall, Edvard Lindelöf, Ola Engkvist, and Hongming Chen. Building attention and edge message passing neural networks for bioactivity and physical–chemical property prediction. *J. Cheminform.*, 12(1):1–18, 2020.

[117] Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. ChemBERTa: large-scale self-supervised pretraining for molecular property prediction. *arXiv*, pages 1–7, 2020. DOI: 10.48550/arXiv.2010.09885.

[118] Xiao-Chen Zhang, Cheng-Kun Wu, Zhi-Jiang Yang, Zhen-Xing Wu, Jia-Cai Yi, Chang-Yu Hsieh, Ting-Jun Hou, and Dong-Sheng Cao. MG-BERT: leveraging unsupervised atomic representation learning for molecular property prediction. *Brief. Bioinform.*, 22(6):bbab152, 2021.

[119] Ross Irwin, Spyridon Dimitriadis, Jiazhen He, and Esben Jannik Bjerrum. Chemformer: a pre-trained transformer for computational chemistry. *Mach. Learn.: Sci. Technol.*, 3(1):015022, 2022.

[120] Łukasz Maziarka, Tomasz Danel, Sławomir Mucha, Krzysztof Rataj, Jacek Tabor, and Stanisław Jastrzebski. Molecule attention transformer. *arXiv*, pages 1–11, 2020. DOI: 10.48550/arXiv.2002.08264.

[121] Dexiong Chen, Leslie O'Bray, and Karsten Borgwardt. Structure-aware transformer for graph representation learning. In *Proceedings of the ICML*, pages 3469–3489. PMLR, 2022.

[122] Jian Gao, Zheyuan Shen, Yufeng Xie, Jialiang Lu, Yang Lu, Sikang Chen, Qingyu Bian, Yue Guo, Liteng Shen, Jian Wu, Binbin Zhou, Tingjun Hou, Qiaojun He, Jinxin Che, and Xiaowu Dong. TransFoxMol: predicting molecular property with focused attention. *Brief. Bioinform.*, 24(5):bbad306, 2023.

[123] Yinghui Jiang, Shuting Jin, Xurui Jin, Xianglu Xiao, Wenfan Wu, Xiangrong Liu, Qiang Zhang, Xiangxiang Zeng, Guang Yang, and Zhangming Niu. Pharmacophoric-constrained heterogeneous graph transformer model for molecular property prediction. *Commun. Chem.*, 6(1):60, 2023.

[124] Apakorn Kengkanna and Masahito Ohue. Enhancing property and activity prediction and interpretation using multiple molecular graph representations with MMGX. *Commun. Chem.*, 7(1):74, 2024.

[125] Chemical Computing Group ULC. Molecular Operating Environment (MOE), version 2019.01, 2019. Montreal, QC, Canada.

[126] Gregory A Landrum. RDKit: Open-source cheminformatics, version 2022.09.5, 2022. https://www.rdkit.org.

[127] Sereina Riniker and Gregory A Landrum. Better informed distance geometry: using what we know to improve conformation generation. *J. Chem. Inf. Model.*, 55(12):2562–2574, 2015.

[128] Shuzhe Wang, Jagna Witek, Gregory A Landrum, and Sereina Riniker. Improving conformer generation for small rings and macrocycles based on distance geometry and experimental torsional-angle preferences. *J. Chem. Inf. Model.*, 60(4):2044–2058, 2020.

[129] Hirotomo Moriwaki, Yu-Shi Tian, Norihito Kawashita, and Tatsuya Takagi. Mordred: a molecular descriptor calculator. *J. Cheminform.*, 10(1):1–14, 2018.

[130] Alan Jović, Karla Brkić, and Nikola Bogunović. A review of feature selection methods with applications. In *Proceedings of the MIPRO*, pages 1200–1205. IEEE, 2015.

[131] Łukasz Maziarka, Dawid Majchrowski, Tomasz Danel, Piotr Gaiński, Jacek Tabor, Igor Podolak, Paweł Morkisz, and Stanisław Jastrzebski. Relative molecule self-attention transformer. *J. Cheminform.*, 16(1):3, 2024.

[132] Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A Hunter, Costas Bekas, and Alpha A Lee. Molecular transformer: a model for

uncertainty-calibrated chemical reaction prediction. *ACS Cent. Sci.*, 5(9):1572–1583, 2019.

[133] Guorui Zhou, Na Mou, Ying Fan, Qi Pi, Weijie Bian, Chang Zhou, Xiaoqiang Zhu, and Kun Gai. Deep interest evolution network for click-through rate prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 5941–5948. AAAI, 2019.

[134] Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. Molecular contrastive learning of representations via graph neural networks. *Nat. Mach. Intell.*, 4(3):279–287, 2022.

[135] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *J. Big Data*, 6(1):1–48, 2019.

[136] Esben Jannik Bjerrum. SMILES enumeration as data augmentation for neural network modeling of molecules. *arXiv*, pages 1–7, 2017. DOI: 10.48550/arXiv.1703.07076.

[137] Taha Rezai, Bin Yu, Glenn L Millhauser, Matthew P Jacobson, and R Scott Lokey. Testing the conformational hypothesis of passive membrane permeability using synthetic cyclic peptide diastereomers. *J. Am. Chem. Soc.*, 128(8):2510–2511, 2006.

[138] Arthur C Rand, Siegfried S F Leung, Heather Eng, Charles J Rotter, Raman Sharma, Amit S Kalgutkar, Yizhong Zhang, Manthena V Varma, Kathleen A Farley, Bhagyashree Khunte, Chris Limberakis, David A Price, Spiros Liras, Alan M Mathiowetz, Matthew P Jacobson, and R Scott Lokey. Optimizing PK properties of cyclic peptides: the effect of side chain substitutions on permeability and clearance. *MedChemComm*, 3(10):1282–1289, 2012.

[139] Serge Zaretsky, Conor C G Scully, Alan J Lough, and Andrei K Yudin. Exocyclic control of turn induction in macrocyclic peptide scaffolds. *Chem. Eur. J.*, 19(52):17668–17672, 2013.

[140] Daniel S Nielsen, Huy N Hoang, Rink Jan Lohman, Timothy A Hill, Andrew J Lucke, David J Craik, David J Edmonds, David A Griffith, Charles J Rotter, Roger B Ruggeri, David A Price, Spiros Liras, and David P Fairlie. Improving on Nature: Making a Cyclic Heptapeptide Orally Bioavailable. *Angew. Chem. Int. Ed.*, 53(45):12059–12063, 2014.

[141] Christopher L Ahlbach, Katrina W Lexa, Andrew T Bockus, Valerie Chen, Phillip Crews, Matthew P Jacobson, and R Scott Lokey. Beyond cyclosporine A: conformation-dependent passive membrane permeabilities of cyclic peptide natural products. *Future Med. Chem.*, 7(16):2121–2130, 2015.

[142] Andrew T Bockus, Katrina W Lexa, Cameron R Pye, Amit S Kalgutkar, Jarret W Gardner, Kathryn C R Hund, William M Hewitt, Joshua A Schwochert, Emerson Glassey, David A Price, Alan M Mathiowetz, Spiros Liras, Matthew P Jacobson, and R Scott Lokey. Probing the Physicochemical Boundaries of Cell Permeability and Oral

Bioavailability in Lipophilic Macrocycles Inspired by Natural Products. *J. Med. Chem.*, 58(11):4581–4589, 2015.

[143] Andrew T Bockus, Joshua A Schwochert, Cameron R Pye, Chad E Townsend, Vong Sok, Maria A Bednarek, and R Scott Lokey. Going Out on a Limb: Delineating the Effects of β-Branching, N-Methylation, and Side Chain Size on the Passive Permeability, Solubility, and Flexibility of Sanguinamide A Analogues. *J. Med. Chem.*, 58(18):7409–7418, 2015.

[144] Udaya Kiran Marelli, Jacqueline Bezençon, Eduard Puig, Beat Ernst, and Horst Kessler. Enantiomeric cyclic peptides with different caco-2 permeability suggest carrier-mediated transport. *Chem. Eur. J.*, 21(22):8023–8027, 2015.

[145] Daniel S Nielsen, Rink Jan Lohman, Huy N Hoang, Timothy A Hill, Alun Jones, Andrew J Lucke, and David P Fairlie. Flexibility versus Rigidity for Orally Bioavailable Cyclic Hexapeptides. *ChemBioChem*, 16(16):2289–2293, 2015.

[146] Joshua Schwochert, Rushia Turner, Melissa Thang, Ray F Berkeley, Alexandra R Ponkey, Kelsie M Rodriguez, Siegfried S F Leung, Bhagyashree Khunte, Gilles Goetz, Chris Limberakis, Amit S Kalgutkar, Heather Eng, Michael J Shapiro, Alan M Mathiowetz, David A Price, Spiros Liras, Matthew P Jacobson, and R Scott Lokey. Peptide to Peptoid Substitutions Increase Cell Permeability in Cyclic Hexapeptides. *Org. Lett.*, 17(12):2928–2931, 2015.

[147] Marianne Fouché, Michael Schäfer, Jörg Berghausen, Sandrine Desrayaud, Markus Blatter, Philippe Piéchon, Ina Dix, Aimar Martingarcia, and Hans Jörg Roth. Design and Development of a Cyclic Decapeptide Scaffold with Suitable Properties for Bioavailability and Oral Exposure. *ChemMedChem*, 11(10):1048–1059, 2016.

[148] Jennifer L Hickey, Serge Zaretsky, Megan A St. Denis, Sai Kumar Chakka, M Monzur Morshed, Conor CG Scully, Andrew L Roughton, and Andrei K Yudin. Passive membrane permeability of macrocycles can be controlled by exocyclic amide bonds. *J. Med. Chem.*, 59(11):5368–5376, 2016.

[149] Joshua Schwochert, Yongtong Lao, Cameron R Pye, Matthew R Naylor, Prashant V Desai, Isabel C Gonzalez Valcarcel, Jaclyn A Barrett, Geri Sawada, Maria-Jesus Blanco, and R Scott Lokey. Stereochemistry balances cell permeability and solubility in the naturally derived phepropeptin cyclic peptides. *ACS Med. Chem. Lett.*, 7(8):757–761, 2016.

[150] Markus Boehm, Kevin Beaumont, Rhys Jones, Amit S Kalgutkar, Liying Zhang, Karen Atkinson, Guoyun Bai, Janice A Brown, Heather Eng, Gilles H Goetz, Brian R Holder, Bhagyashree Khunte, Sarah Lazzaro, Chris Limberakis, Sangwoo Ryu, Michael J Shapiro, Laurie Tylaska, Jiangli Yan, Rushia Turner, Siegfried S F Leung, Mahesh Ramaseshan, David A Price, Spiros Liras, Matthew P Jacobson, David J Earp, R Scott Lokey, Alan M Mathiowetz, and Elnaz Menhaji-Klotz. Discovery of potent and orally

bioavailable macrocyclic peptide–peptoid hybrid CXCR7 modulators. *J. Med. Chem.*, 60(23):9653–9663, 2017.

[151] Cameron R Pye, William M Hewitt, Joshua Schwochert, Terra D Haddad, Chad E Townsend, Lyns Etienne, Yongtong Lao, Chris Limberakis, Akihiro Furukawa, Alan M Mathiowetz, David A Price, Spiros Liras, and R Scott Lokey. Nonclassical Size Dependence of Permeation Defines Bounds for Passive Adsorption of Large Drug Molecules. *J. Med. Chem.*, 60(5):1665–1672, 2017.

[152] Laura K Buckton and Shelli R McAlpine. Improving the Cell Permeability of Polar Cyclic Peptides by Replacing Residues with Alkylated Amino Acids, Asparagines, and d-Amino Acids. *Org. Lett.*, 20(3):506–509, 2018.

[153] Chugai Pharma. Co., Ltd. Cyclic peptide compound having high membrane permeability, and library containing same, 2018. Patent WO2018225864A1.

[154] Júlia García-Pindado, Tom Willemse, Rebecca Goss, Bert UW Maes, Ernest Giralt, Steven Ballet, and Meritxell Teixidó. Bromotryptophans and their incorporation in cyclic and bicyclic privileged peptides. *Biopolymers*, 109(10):e23112, 2018.

[155] Masato Kaneda, Shinsaku Kawaguchi, Nobutaka Fujii, Hiroaki Ohno, and Shinya Oishi. Structure-Activity Relationship Study on Odoamide: Insights into the Bioactivities of Aurilide-Family Hybrid Peptide-Polyketides. *ACS Med. Chem. Lett.*, 9(4):365–369, 2018.

[156] Leo LH Lee, Laura K Buckton, and Shelli R McAlpine. Converting polar cyclic peptides into membrane permeable molecules using N-methylation. *Pept. Sci.*, 110(3):e24063, 2018.

[157] Matthew R Naylor, Andrew M Ly, Mason J Handford, Daniel P Ramos, Cameron R Pye, Akihiro Furukawa, Victoria G Klein, Ryan P Noland, Quinn Edmondson, Alexandra C Turmon, William M Hewitt, Joshua Schwochert, Chad E Townsend, Colin N Kelly, Maria Jesus Blanco, and R Scott Lokey. Lipophilic Permeability Efficiency Reconciles the Opposing Roles of Lipophilicity in Membrane Permeability and Aqueous Solubility. *J. Med. Chem.*, 61(24):11169–11182, 2018.

[158] Suelem D Ramalho, Conan K Wang, Gordon J King, Karl A Byriel, Yen Hua Huang, Vanderlan S Bolzani, and David J Craik. Synthesis, Racemic X-ray Crystallographic, and Permeability Studies of Bioactive Orbitides from Jatropha Species. *J. Nat. Prod.*, 81(11):2436–2445, 2018.

[159] Nicholas Barlow, David K Chalmers, Billy J Williams-Noonan, Philip E Thompson, Raymond S Norton, and Philip E Thompson. Improving Membrane Permeation in the beyond Rule-of-Five Space by Using Prodrugs to Mask Hydrogen Bond Donors. *ACS Chem. Biol.*, 15(8):2070–2078, 2020.

[160] Akihiro Furukawa, Joshua Schwochert, Cameron R Pye, Daigo Asano, Quinn D Edmondson, Alexandra C Turmon, Victoria G Klein, Satoshi Ono, Okimasa Okada, and

R Scott Lokey. Drug-like properties in macrocycles above MW 1000: backbone rigidity versus side-chain lipophilicity. *Angew. Chem. Int. Ed.*, 59(48):21571–21577, 2020.

[161] Yuki Hosono, Jumpei Morimoto, Chad Townsend, Colin N Kelly, Matthew R Naylor, Hsiau-Wei Lee, R Scott Lokey, and Shinsuke Sando. Amide-to-Ester Substitution Improves Membrane Permeability of a Cyclic Peptide Without Altering Its Three-Dimensional Structure. *ChemRxiv*, pages 1–17, 2020. DOI: 10.26434/chemrxiv.12272861.v1.

[162] Antoine Le Roux, Emilie Blaise, Pierre-Luc Boudreault, Christian Comeau, Annie Doucet, Marilena Giarrusso, Marie-Pierre Collin, Thomas Neubauer, Florian Kölling, Andreas H Göller, Lea Seep, Dieudonné T Tshitenge, Matthias Wittwer, Maximilian Kullmann, Alexander Hillisch, Joachim Mittendorf, and Eric Marsault. Structure–permeability relationship of semipeptidic macrocycles—understanding and optimizing passive permeability and efflux ratio. *J. Med. Chem.*, 63(13):6774–6783, 2020.

[163] Chad Townsend, Eva Jason, Matthew R Naylor, Cameron R Pye, Joshua A Schwochert, Quinn Edmondson, and R Scott Lokey. The passive permeability landscape around geometrically diverse hexa-and heptapeptide macrocycles. *ChemRxiv*, pages 1–21, 2020. DOI: 10.26434/chemrxiv.13335941.v1.

[164] Christian Comeau, Benjamin Ries, Thomas Stadelmann, Jacob Tremblay, Sylvain Poulet, Ulrike Fröhlich, Jérôme Côté, Pierre-Luc Boudreault, Rabeb Mouna Derbali, Philippe Sarret, Michel Grandbois, Grégoire Leclair, Sereina Riniker, and Éric Marsault. Modulation of the passive permeability of semipeptidic macrocycles: N- and C-methylations fine-tune conformation and properties. *J. Med. Chem.*, 64(9):5365–5383, 2021.

[165] Andrei A Golosov, Alec N Flyer, Jakal Amin, Charles Babu, Christian Gampe, Jingzhou Li, Eugene Liu, Katsumasa Nakajima, David Nettleton, Tajesh J Patel, Patrick C Reid, Lihua Yang, and Lauren G Monovich. Design of thioether cyclic peptide scaffolds with passive permeability and oral exposure. *J. Med. Chem.*, 64(5):2622–2633, 2021.

[166] Colin N Kelly, Chad E Townsend, Ajay N Jain, Matthew R Naylor, Cameron R Pye, Joshua Schwochert, and R Scott Lokey. Geometrically diverse lariat peptide scaffolds reveal an untapped chemical space of high membrane permeability. *J. Am. Chem. Soc.*, 143(2):705–714, 2020.

[167] Dongjae Lee, Jung-Ah Kang, Chanseok Lim, Sunjae Bae, Jieun Choi, Minji Park, Yong-Chul Kim, Yuri Cho, Sung-Gyoo Park, and Jiwon Seo. Entry inhibition of hepatitis B virus using cyclosporin O derivatives with peptoid side chain incorporation. *Bioorg. Med. Chem.*, 68:116862, 2022.

[168] George J Saunders and Andrei K Yudin. Property-Driven Development of Passively Permeable Macrocyclic Scaffolds Using Heterocycles. *Angew. Chem. Int. Ed.*, 61(33):e202206866, 2022.

[169] Takashi Tamura, Masaaki Inoue, Yuji Yoshimitsu, Ichihiko Hashimoto, Noriyuki Ohashi, Kyosuke Tsumura, Koo Suzuki, Takayoshi Watanabe, and Takahiro Hohsaka. Chemical synthesis and cell-free expression of thiazoline ring-bridged cyclic Peptides and their properties on biomembrane permeability. *Bull. Chem. Soc. Jpn.*, 95(2):359–366, 2022.

[170] Tianhong Zhang, Hongli Li, Hualin Xi, Robert V. Stanton, and Sergio H. Rotstein. HELM: A Hierarchical Notation Language for Complex Biomolecule Structure Representation. *J. Chem. Inf. Model.*, 52(10):2796–2806, 2012.

[171] Kohulan Rajan, Achim Zielesny, and Christoph Steinbeck. STOUT: SMILES to IUPAC names using neural machine translation. *J. Cheminformatics*, 13(1):1–14, 2021.

[172] Pritha Ghosh, Nishant Raj, Hitesh Verma, Monika Patel, Sohini Chakraborti, Bhavesh Khatri, Chandrashekar M Doreswamy, S R Anandakumar, Srinivas Seekallu, M B Dinesh, Gajanan Jadhav, Prem Narayan Yadav, and Jayanta Chatterjee. An amide to thioamide substitution improves the permeability and bioavailability of macrocyclic peptides. *Nat. Commun.*, 14(1):6050, 2023.

[173] Atsushi Ohta, Mikimasa Tanada, Shojiro Shinohara, Yuya Morita, Kazuhiko Nakano, Yusuke Yamagishi, Ryusuke Takano, Shiori Kariyuki, Takeo Iida, Atsushi Matsuo, Kazuhisa Ozeki, Takashi Emura, Yuuji Sakurai, Koji Takano, Atsuko Higashida, Miki Kojima, Terushige Muraoka, Ryuuichi Takeyama, Tatsuya Kato, Kaori Kimura, Kotaro Ogawa, Kazuhiro Ohara, Shota Tanaka, Yasufumi Kikuchi, Nozomi Hisada, Ryuji Hayashi, Yoshikazu Nishimura, Kenichi Nomura, Tatsuhiko Tachibana, Machiko Irie, Hatsuo Kawada, Takuya Torizawa, Naoaki Murao, Tomoya Kotake, Masahiko Tanaka, Shiho Ishikawa, Taiji Miyake, Minoru Tamiya, Masako Arai, Aya Chiyoda, Sho Akai, Hitoshi Sase, Shino Kuramoto, Toshiya Ito, Takuya Shiraishi, Tetsuo Kojima, and Hitoshi Iikura. Validation of a new methodology to create oral drugs beyond the rule of 5 for intracellular tough targets. *J. Am. Chem. Soc.*, 145(44):24035–24051, 2023.

[174] Mikimasa Tanada, Minoru Tamiya, Atsushi Matsuo, Aya Chiyoda, Koji Takano, Toshiya Ito, Machiko Irie, Tomoya Kotake, Ryuuichi Takeyama, Hatsuo Kawada, Ryuji Hayashi, Shiho Ishikawa, Kenichi Nomura, Noriyuki Furuichi, Yuya Morita, Mirai Kage, Satoshi Hashimoto, Keiji Nii, Hitoshi Sase, Kazuhiro Ohara, Atsushi Ohta, Shino Kuramoto, Yoshikazu Nishimura, Hitoshi Iikura, and Takuya Shiraishi. Development of orally bioavailable peptides targeting an intracellular protein: from a hit to a clinical KRAS inhibitor. *J. Am. Chem. Soc.*, 145(30):16610–16620, 2023.

[175] Catherine Bergeron, Christopher Bérubé, Henry Lamb, Yasuko Koda, David J Craik, Sónia Troeira Henriques, Normand Voyer, and Nicole Lawrence. Analogs of Cyclic Peptide Mortiamide-D From Marine Fungi Have Improved Membrane Permeability and Kill Drug-Resistant Melanoma Cells. *Pept. Sci.*, page e24380, 2024.

[176] Justin H Faris, Emel Adaligil, Nataliya Popovych, Satoshi Ono, Mifune Takahashi, Huy Nguyen, Emile Plise, Jaru Taechalertpaisarn, Hsiau-Wei Lee, Michael F T Koehler,

Christian N Cunningham, and R Scott Lokey. Membrane Permeability in a Large Macrocyclic Peptide Driven by a Saddle-Shaped Conformation. *J. Am. Chem. Soc.*, 146(7):4582–4591, 2024.

[177] Mirai Kage, Ryuji Hayashi, Atsushi Matsuo, Minoru Tamiya, Shino Kuramoto, Kazuhiro Ohara, Machiko Irie, Aya Chiyoda, Koji Takano, Toshiya Ito, Tomoya Kotake, Ryuuichi Takeyama, Shiho Ishikawa, Kenichi Nomura, Noriyuki Furuichi, Yuya Morita, Satoshi Hashimoto, Hatsuo Kawada, Yoshikazu Nishimura, Keiji Nii, Hitoshi Sase, Atsushi Ohta, Tetsuo Kojima, Hitoshi Iikura, Mikimasa Tanada, and Takuya Shiraishi. Structure-activity relationships of middle-size cyclic peptides, KRAS inhibitors derived from an mRNA display. *Bioorg. Med. Chem.*, 110:117830, 2024.

[178] Huy M Ly, Michael Desgagné, Duc Tai Nguyen, Christian Comeau, Ulrike Froehlich, Éric Marsault, and Pierre-Luc Boudreault. Insights on Structure–Passive Permeability Relationship in Pyrrole and Furan-Containing Macrocycles. *J. Med. Chem.*, 67(5):3711–3726, 2024.

[179] Yuko Otani, Asami Ichinose, Xihong Wang, Zhihan Huang, Akitomo Kasahara, Mayumi Ishii, Eri Watanabe, Kayoko Kanamitsu, Kempei Tai, Hiroyuki Kusuhara, Takumi Ueda, Koh Takeuchi, and Tomohiko Ohwada. An N-ortho-nitrobenzylated benzanilide amino acid enables control of the conformation and membrane permeability of cyclic peptides. *Chem. Commun.*, 60(69):9242–9245, 2024.

[180] R K H Galvao, M C U Araujo, G E José, M J C Pontes, E C Silva, and T C B Saldanha. A method for calibration and validation subset partitioning. *Talanta*, 67(4):736–740, 2005.

[181] Tomohiro Ban, Masahito Ohue, and Yutaka Akiyama. Efficient hyperparameter optimization by using Bayesian optimization for drug-target interaction prediction. In *Proceedings of the 2017 IEEE 7th ICCABS*, pages 1–6. IEEE, 2017.

[182] James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. *Adv. Neural. Inf. Process. Syst.*, 24, 2011.

[183] Yoshihiko Ozaki, Yuki Tanigaki, Shuhei Watanabe, and Masaki Onishi. Multiobjective tree-structured parzen estimator for computationally expensive optimization problems. In *Proceedings of the 2020 GECCO*, pages 533–541. ACM, 2020.

[184] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631. ACM, 2019.

[185] Ruochi Zhang, Haoran Wu, Yuting Xiu, Kewei Li, Ningning Chen, Yu Wang, Yan Wang, Xin Gao, and Fengfeng Zhou. PepLand: a large-scale pre-trained peptide representation model for a comprehensive landscape of both canonical and non-canonical amino acids. *arXiv*, pages 1–29, 2023. DOI: 10.48550/arXiv.2311.04419.

[186] Lujing Cao, Zhenyu Xu, Tianfeng Shang, Chengyun Zhang, Xinyi Wu, Yejian Wu, Silong Zhai, Zhajun Zhan, and Hongliang Duan. Multi_CycGT: A Deep Learning-Based Multimodal Model for Predicting the Membrane Permeability of Cyclic Peptides. *J. Med. Chem.*, 67(3):1888–1899, 2024.

[187] Zixu Wang, Yangyang Chen, Xiucai Ye, and Tetsuya Sakurai. CyclePermea: Membrane Permeability Prediction of Cyclic Peptides with a Multi-Loss Fusion Network. In *Proceedings of the IJCNN*, pages 1–8. IEEE, 2024.

[188] Xiaorong Tan, Qianhui Liu, Yanpeng Fang, Yingli Zhu, Fei Chen, Wenbin Zeng, Defang Ouyang, and Jie Dong. Predicting Peptide Permeability Across Diverse Barriers: A Systematic Investigation. *Mol. Pharm.*, 21(8):4116–4127, 2024.

[189] Yunxiang Yu, Mengyun Gu, Hai Guo, Yabo Deng, Danna Chen, Jianwei Wang, Caixia Wang, Xia Liu, Wenjin Yan, and Jinqi Huang. MuCoCP: a priori chemical knowledge-based multimodal contrastive learning pre-trained neural network for the prediction of cyclic peptide membrane penetration ability. *Bioinformatics*, 40(8), 2024.

[190] Aaron Lee Feller and Claus O Wilke. Peptide-specific chemical language model successfully predicts membrane diffusion of cyclic peptides. *bioRxiv*, pages 1–21, 2024. DOI: 10.1101/2024.08.09.607221.

[191] Craig Knox, Mike Wilson, Christen M Klinger, Mark Franklin, Eponine Oler, Alex Wilson, Allison Pon, Jordan Cox, Na Eun Chin, Seth A Strawbridge, Marysol Garcia-Patino, Ray Kruger, Aadhavya Sivakumaran, Selena Sanford, Rahil Doshi, Nitya Khetarpal, Omolola Fatokun, Daphnee Doucet, Ashley Zubkowski, Dorsa Yahya Rayat, Hayley Jackson, Karxena Harford, Afia Anjum, Mahi Zakir, Fei Wang, Siyang Tian, Brian Lee, Jaanus Liigand, Harrison Peters, Ruo Qi (Rachel) Wang, Tue Nguyen, Denise So, Matthew Sharp, Rodolfo da Silva, Cyrella Gabriel, Joshua Scantlebury, Marissa Jasinski, David Ackerman, Timothy Jewison, Tanvir Sajed, Vasuk Gautam, and David S Wishart. DrugBank 6.0: the DrugBank knowledgebase for 2024. *Nucleic Acids Res.*, 52(D1):D1265–D1275, 2024.

[192] W Couet, N Gregoire, S Marchand, and O Mimoz. Colistin pharmacokinetics: the fog is lifting. *Clin. Microbiol. Infect.*, 18(1):30–39, 2012.

[193] David T Bearden. Clinical pharmacokinetics of quinupristin/dalfopristin. *Clin. Pharmacokinet*, 43:239–252, 2004.

[194] Simulations Plus Inc. ADMET Predictor 10.0, 2020. California, Lancaster, USA.

[195] Antón Garcia-Diaz, Xosé R Fdez-Vidal, Xosé M Pardo, and Raquel Dosil. Saliency from hierarchical adaptation through decorrelation and variance normalization. *Image Vis. Comput.*, 30(1):51–64, 2012.

[196] Matthias Kümmerer, Lucas Theis, and Matthias Bethge. Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. *arXiv*, pages 1–12, 2014. DOI: 10.48550/arXiv.1411.1045.

[197] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. How to explain individual classification decisions. *J. Mach. Learn. Res.*, 11:1803–1831, 2010.

[198] Schrödinger LLC. Schrödinger Release 2019-1: Glide, 2019. New York, NY, USA.

[199] Schrödinger LLC. Schrödinger Release 2019-1: LigPrep, 2019. New York, NY, USA.

[200] Schrödinger LLC. Schrödinger Release 2019-1: MacroModel, 2019. New York, NY, USA.

[201] Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Methodol.*, 58(1):267–288, 1996.

[202] Satoshi Hara and Takanori Maehara. Enumerate lasso solutions for feature selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, pages 1985–1991. AAAI, 2017.

[203] Yang Zhou, Rong Jin, and Steven Chu-Hong Hoi. Exclusive lasso for multi-task feature selection. In *Proceedings of the AISTATS*, pages 988–995. JMLR, 2010.

[204] Francis R Bach. Bolasso: model consistent lasso estimation through the bootstrap. In *Proceedings of the ICML*, pages 33–40. ACM, 2008.

[205] D A Case, K Belfon, I Y Ben-Shalom, S R Brozell, D S Cerutti, T E III Cheatham, V W D Cruzeiro, T A Darden, R E Duke, G Giambasu, M K Gilson, H Gohlke, A W Goetz, R Harris, S Izadi, S A Izmailov, K Kasavajhala, A Kovalenko, R Krasny, T Kurtzman, T S Lee, S LeGrand, P Li, C Lin, J Liu, T Luchko, R Luo, V Man, K M Merz, Y Miao, O Mikhailovskii, G Monard, H Nguyen, A Onufriev, F Pan, S Pantano, R Qi, D R Roe, A Roitberg, C Sagui, S Schott-Verdugo, J Shen, C L Simmerling, N R Skrynnikov, J Smith, J Swails, R C Walker, J Wang, L Wilson, R M Wolf, X Wu, Y Xiong, Y Xue, D M York, and P A Kollman. AMBER 2020, 2020. University of California, San Francisco.

[206] Zhi-Li Zuo, Ling Guo, and Ricardo L Mancera. Free energy of binding of coiled-coil complexes with different electrostatic environments: the influence of force field polarisation and capping. *Nat. Prod. Bioprospect.*, 4:285–295, 2014.

[207] Junichi Higo, Nobutoshi Ito, Masataka Kuroda, Satoshi Ono, Nobuyuki Nakajima, and Haruki Nakamura. Energy landscape of a peptide consisting of α-helix, 310-helix, β-turn, β-hairpin, and other disordered conformations. *Prot. Sci.*, 10(6):1160–1171, 2001.

[208] Miho Isegawa, Bo Wang, and Donald G Truhlar. Electrostatically embedded molecular tailoring approach and validation for peptides. *J. Chem. Theory Comput.*, 9(3):1381–1393, 2013.

[209] Lei Diao and Bernd Meibohm. Pharmacokinetics and pharmacokinetic–pharmacodynamic correlations of therapeutic peptides. *Clin. Pharmacokinet.*, 52:855–868, 2013.

[210] Huifeng Zhao, Dejun Jiang, Chao Shen, Jintu Zhang, Xujun Zhang, Xiaorui Wang, Dou Nie, Tingjun Hou, and Yu Kang. Comprehensive Evaluation of 10 Docking Programs on a Diverse Set of Protein–Cyclic Peptide Complexes. *J. Chem. Inf. Model.*, 64(6):2112–2124, 2024.

[211] Yuqi Zhang and Michel F Sanner. Docking flexible cyclic peptides with AutoDock CrankPep. *J. Chem. Theory Comput.*, 15(10):5161–5168, 2019.

[212] Vicky Charitou, Siri C Van Keulen, and Alexandre MJJ Bonvin. Cyclization and docking protocol for cyclic peptide–protein modeling using HADDOCK2.4. *J. Chem. Theory Comput.*, 18(6):4027–4040, 2022.

[213] Shahrzad Ahangarzadeh, Mohammad M Kanafi, Simzar Hosseinzadeh, Ahad Mokhtarzadeh, Mahmood Barati, Javad Ranjbari, and Lobat Tayebi. Bicyclic peptides: types, synthesis and applications. *Drug Discov. Today*, 24(6):1311–1319, 2019.

List of Publications

Journal Papers

1. **Jianan Li**, Keisuke Yanagisawa, Yasushi Yoshikawa, Masahito Ohue, Yutaka Akiyama. "Plasma protein binding prediction focusing on residue-level features and circularity of cyclic peptides by deep learning", *Bioinformatics*, **38**(4): 1110–1117, 2022. DOI: 10.1093/bioinformatics/btab726.

2. **Jianan Li**, Keisuke Yanagisawa, Masatake Sugita, Takuya Fujie, Masahito Ohue, Yutaka Akiyama. "CycPeptMPDB: A Comprehensive Database of Membrane Permeability of Cyclic Peptides", *Journal of Chemical Information and Modeling*, **63**(7): 2240–2250, 2023. DOI: 10.1021/acs.jcim.2c01573.

3. **Jianan Li**, Keisuke Yanagisawa, Yutaka Akiyama. "CycPeptMP: enhancing membrane permeability prediction of cyclic peptides with multi-level molecular features and data augmentation", *Briefings in Bioinformatics*, **25**(5): bbae417, 2024. DOI: 10.1093/bib/bbae417.

4. Masatake Sugita, Yudai Noso, **Jianan Li**, Takuya Fujie, Keisuke Yanagisawa, Yutaka Akiyama. "Protocol for Membrane Permeability Prediction of Cyclic Peptides using Descriptors from Extended Ensemble Molecular Dynamics Simulations and 2D Descriptors", *Journal of Chemical Information and Modeling*, 2024. (submitted)